Petra Perner (Ed.)

Advances in Data Mining

Applications and Theoretical Aspects

LNAI 7987

13th Industrial Conference, ICDM 2013
New York, NY, USA, July 2013
Proceedings

Lecture Notes in Artificial Intelligence 7987

Subseries of Lecture Notes in Computer Science

Petra Perner (Ed.)

Advances in Data Mining

Applications and Theoretical Aspects

13th Industrial Conference, ICDM 2013
New York, NY, USA, July 16-21, 2013
Proceedings

Volume Editor

Petra Perner
Institute of Computer Vision
and Applied Computer Sciences, IBaI
Kohlenstraße 2
04107 Leipzig, Germany
E-mail: pperner@ibai-institut.de

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-39735-6 e-ISBN 978-3-642-39736-3
DOI 10.1007/978-3-642-39736-3
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2013943124

CR Subject Classification (1998): I.2.6, I.2, H.2.8, J.3, H.3, I.4-5, J.1

LNCS Sublibrary: SL 7 – Artificial Intelligence

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The 13 event of the Industrial Conference on Data Mining ICDM was held in New York (www.data-mining-forum.de) running under the umbrella of the World Congress "The Frontiers in Intelligent Data and Signal Analysis, DSA 2013."

For this edition, the Program Committee received 112 submissions. After the peer-review process, we accepted 33 high-quality papers for oral presentation, of which 22 included in this proceeding book. The topics range from theoretical aspects of data mining to applications of data mining, such as in multimedia data, in marketing, finance and telecommunication, in medicine and agriculture, and in process control, industry and society. Extended versions of selected papers will appear in the*International Journal Transactions on Machine Learning and Data Mining* (www.ibai-publishing.org/journal/mldm).

In all, 30 papers were selected for poster presentations and six for industry paper presentations that are published in the *ICDM Poster and Industry Proceeding* by *ibai-publishing* (www.ibai-publishing.org).

In conjunction with ICDM, four workshops were run focusing on special hot application-oriented topics in data mining: the Workshop on Case-Based Reasoning (CBR-MD), Data Mining in Marketing (DMM), and the Workshop on Data Mining in Agriculture (DMA). All workshop papers are published in the *workshop proceedings* by ibai-publishing (www.ibai-publishing.org).

A tutorial on Data Mining, a tutorial on Case-Based Reasoning, a tutorial on Intelligent Image Interpretation and Computer Vision in Medicine, Biotechnology, Chemistry & Food Industry, a tutorial on Big Data and Text Analysis and a tutorial on Standardization in Immunofluorescence were held before the conference.

We were pleased to give out the best paper award for ICDM for the seventh time this year. There are four announcement mentioned at www.data-mining-forum.de. The final decision was made by the Best Paper Award Committee based on the presentation by the authors and the discussion with the auditorium. The ceremony took place at the end of the conference. This prize is sponsored by ibai solutions (www.ibai-solutions.de), one of the leading companies in data mining for marketing, Web mining and e-commerce.

The conference was rounded up by an outlook session on new challenging topics in data mining before the Best Paper Award Ceremony.

We would like to thank all reviewers for their highly professional work and their effort in reviewing the papers.

We also thank the members of the Institute of Applied Computer Sciences, Leipzig, Germany (www.ibai-institut.de), who handled the conference as secretariat. We appreciate the help and understanding of the editorial staff at Springer Verlag, and in particular Alfred Hofmann, who supported the publication of these proceedings in the LNAI series.

Last, but not least, we wish to thank all the speakers and participants who contributed to the success of the conference. We hope to see you in 2014 in Sankt Petersburg at the next World Congress "The Frontiers in Intelligent Data and Signal Analysis, DSA 2014" (`www.worldcongressdsa.com`) that combines under its roof the following three events: International Conferences Machine Learning and Data Mining MLDM, the Industrial Conference on Data Mining ICDM, and the International Conference on Mass Data Analysis of Signals and Images in Medicine, Biotechnology, Chemistry and Food Industry MDA.

July 2013 Petra Perner

Organization

Chair

Petra Perner	IBaI, Leipzig, Germany

Program Committee

Ajith Abraham	Machine Intelligence Research Labs, USA
Andrea Ahlemeyer-Stubbe	ENBIS, Amsterdam, The Netherlands
Eva Armengol	IIA CSIC, Spain
Brigitte Bartsch-Spörl	BSR Consulting GmbH, Germany
Orlando Belo	University of Minho, Portugal
Isabelle Bichindaritz	State University of New York, USA
Leon Bobrowski	Bialystok Technical University, Poland
Marc Boullé	France Télécom, France
Shirley Coleman	University of Newcastle, UK
Juan M. Corchado	Universidad de Salamanca, Spain
Antonio Dourado	University of Coimbra, Portugal
Jeroen de Bruin	Medical University of Vienna, Austria
Peter Funk	Mälardalen University, Sweden
Geert Gins	KU Leuven, Belgium
Warwick Graco	ATO, Australia
Osman Hegazy	Cairo University, Egypt
Gary F. Holness	Delaware State University, USA
Pedro Isaias	Universidade Aberta, Portugal
Piotr Jedrzejowicz Gdynia	Maritime University, Poland
Martti Juhola	University of Tampere, Finland
Janusz Kacprzyk	Polish Academy of Sciences, Poland
Mineichi Kudo	Hokkaido University, Japan
Mehmed Kantardzic	University of Louisville, USA
David Manzano Macho	Ericsson Research Spain, Spain
Dunja Mladenic	Jozef Stefan Institute, Slovenia
Eduardo F. Morales	INAOE, Ciencias Computacionales, Mexico
Stefania Montani	Università del Piemonte Orientale, Italy
Jerry Oglesby	SAS Institute Inc., USA
Wieslaw Paja	University of Information Technology and Management in Rzeszow, Poland
Eric Pauwels	CWI Amsterdam, The Netherlands
Mykola Pechenizkiy	Eindhoven University of Technology, The Netherlands
Jonas Poelmans	KU Leuven, Belgium

Georg Ruß	Otto-von-Guericke-Universität Magdeburg, Germany
Rainer Schmidt	University of Rostock, Germany
Kaoru Shimada	Fukuoka Dental College, Japan
Yanbo J. Wang	China Minsheng Banking Corporation Ltd., China
Claus Weihs	University of Dortmund, Germany
Yong Zheng	DePaul University, USA

Table of Contents

Mining and Information Integration Practice for Chinese Bibliographic Database of Life Sciences

Heng Chen[1], Yi Jin[2], Yan Zhao[3], Yongjuan Zhang[1], Chengcai Chen[1], Jilin Sun[1], and Shen Zhang[1]

[1] Shanghai Information Center for Life Sciences/ Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China
{hengchen,yjzhang,ccchen,jlsun,zhangshen}@sibs.ac.com
[2] Shanghai Jiao Tong University Library, Shanghai, China
y_king@hotmail.com
[3]College of International Business, Shanghai International Studies University, Shanghai , China
zhaoyan2000@shisu.edu.cn

Abstract. With fast development of life science research, countless achievements have been generated and scattered in various literatures. Information providers are facing the challenge of satisfying users' needs for more efficient and intelligent retrieval. Information integration and mining are promising ways that become more and more important. This paper describes how protein related information is mined from the Chinese Biological Abstract (CBA) database, and integrated with corresponding information in the Universal Protein Resource (Uniprot database). With the collaboration of European Bioinformatics Institute (EBI), integration with corresponding protein information in the Uniprot database is achieved. This paper describes the integration and mapping between Chinese bibliographic databases and authoritative factual databases through relevant text mining works. It would be helpful for extension, utilization and mining of Chinese bibliographic resources, as well as cross lingual information retrieval, integration, and mining.

Keywords: bibliographic database, factual database, information integration, text mining.

1 Introduction

As one of the most active research fields, life science generates countless achievements that scatter in various literatures every year. Although most of them are accessible through databases and web sites, it is still a problem for users to identify what they really need from enormous search results. So information integration and mining are essential to meet users' needs for more efficient and intelligent retrieval. Some successful works have been carried out, such as GOPubMed, which can automatically recognize concepts from user's search query to PubMed and display papers containing relevant terms[1], and Entrez, an integrated search system that enables access to multiple National Center for Biotechnology Information (NCBI)

P. Perner (Ed.): ICDM 2013, LNAI 7987, pp. 1–10, 2013.

databases[2]. Similar works are also reported by Pasquier[3], McGarry[4], Alexopoulou[5] and Sahoo[6], etc. However, all these works were based on western language literatures. Compared to studies of U.S. and European countries, information integration and mining related to Chinese life science literatures are still less extensive and thorough in China. Conducted works are basically for the purpose of research or demonstration, and can't actually provide services based on information integration and mining. This paper describes our efforts to mine protein related information from the Chinese Biological Abstract (CBA) database, and integrate with relevant information in the Universal Protein Resource (Uniprot database).

CBA is a comprehensive bibliographic database that collects Chinese research achievements in life sciences. CBA is run by Shanghai Information Center for Life Sciences (SICLS) of Chinese Academy of Sciences since 1985 and covers over 800 Chinese periodicals related to biology, medicine, basic medicine, basic agriculture, and biological interdisciplinary sciences. CBA reflects the latest development trends of life sciences and bio-medical researches in China. UniProt database contains comprehensive factual databases for protein sequence and annotation data, and is a collaboration between the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR)[7]. Because of the rapid accumulation of genome sequences for many organisms, currently research attention is turning to the identification and functions of proteins encoded by these genomes. With the increasing volume and variety of protein sequences and functional information, UniProt database can serve as a central resource of protein sequence, structure and function, providing rich, consistent and nonredundant protein information[8]. EBI and SICLS have established collaborative relationship and EBI literature database also contains some selected CBA data (around 130 thousand records). As part of the collaboration, we hope to create links between Chinese bibliographic database (CBA) and factual databases (UniProt database). There are no similar works being reported in China so far. This work would be very helpful for Chinese users to obtain relevant protein information easily and rapidly from search results. Since it is quite convenient to correctly translate the protein name from Chinese into English, subsequently, it is easily to use the English name to access UniProt database website to retrieve relevant information.

2 System Structure of CBA

Figure 1 depicts the system structure of CBA. Selected data records that qualify the standards of CBA are processed first before being loaded into the CBA database. During this procedure, proteins contained in records are identified and their features are annotated according to the controlled-vocabulary dictionary. Links to UniProt database are also generated based on accession numbers or names of identified proteins. Some optimization rules are applied in order to improve search efficiency. Protein names in search results can be highlighted by users and if they want to know more about these proteins, just click on them and follow links to access UniProt

databases. Then detailed protein information and annotation can easily be obtained from UniProt database, such as protein sequences, sequence annotations, functional annotations, sequence blast, alignment and homologous analysis, etc.

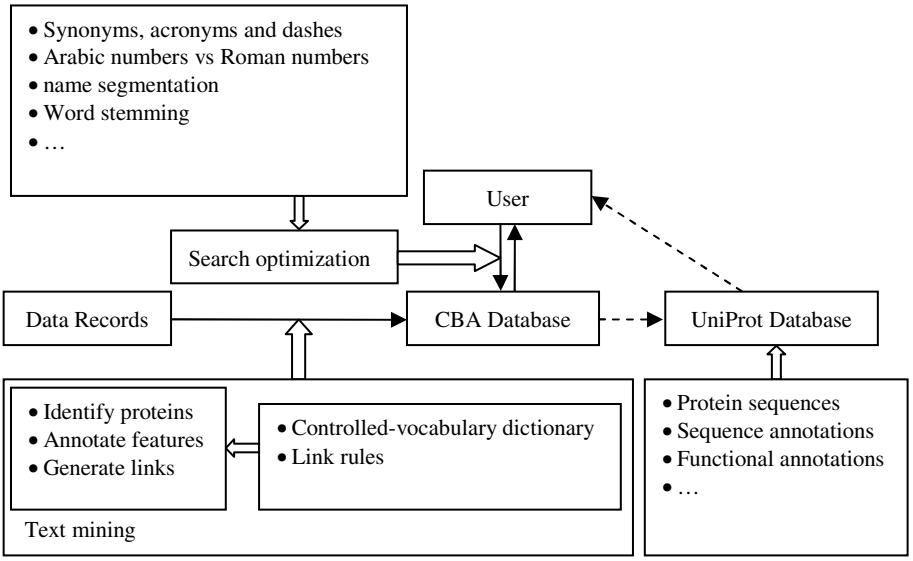


Fig. 1. System structure of CBA

3 English-Chinese Controlled-Vocabulary Dictionary for Proteins

An English-Chinese Controlled-vocabulary dictionary is essential for mining protein information efficiently and accurately. The dictionary is based on standardized nomenclature and controlled vocabularies of UniProt database, since UniProt database makes use of the official nomenclature defined by international committees while still providing the published synonyms[8]. Besides protein names, the dictionary also includes relevant information needed for text mining. The most important work here is to establish mapping relationships between English protein names and corresponding Chinese protein names. For each protein name, we must take into consideration of its synonyms, homonyms, acronyms, and different writing habits, etc. SICLS also invited Professor Weimin Zhu of EBI to Shanghai and his comprehensive introduction to UniProt database and detailed technical information required for text mining are very helpful for our collaboration. The work starts with downloading relevant data from UniProt database. Then we extract needed items from these data, including accession numbers, protein names and their synonyms, and create an English protein name list. According to Chinese classified Thesaurus, Chinese Medical Subject Headings (CMeSH), and keywords and their synonyms in CBA database, we add corresponding Chinese protein names and two protein features to the list and then build the English-Chinese Controlled-vocabulary dictionary. Since a protein has many features, two features (functional/structural, soluble/insoluble)

annotated here are mainly for the purpose of test. More features can be added later based on actual needs. The mapping between English protein names and Chinese protein names enables text mining of proteins for CBA records.

Figure 2 is a screenshot of part of the English-Chinese Controlled-vocabulary dictionary, in which field A records UniProt database accession numbers of proteins and field B is the English protein name relevant to accession numbers in field A. Field C, D, and E are corresponding Chinese protein names. It can be noted that many proteins have several Chinese names. Mapping these relationships is therefore the most important and time-consuming part of work. Currently the dictionary contains more than 15,000 proteins and we will continue this work until all downloaded data are processed.

	A	B	C	D	E
1	Accession	name	cname	cname2	cname3
2	P09196,P18144	Atrial natriuretic factor	心钠素	心房钠尿因子	心房利钠因子
3	O46540,P01160,P01161,P05125,P07499,P07500,F	Atrial natriuretic peptide	心钠素	心房钠尿肽	心钠肽
4	P16066,P18293,P18910	Atrial natriuretic peptide A-type receptor	心钠素A型受体	心房钠尿肽A型受体	心钠肽A型受体
5	P16067,P20594,P46197,P55202,Q6VVW5	Atrial natriuretic peptide B-type receptor	心钠素B型受体	心房钠尿肽B型受体	心钠肽B型受体
6	P10730,P17342,P41740,P70180	Atrial natriuretic peptide C-type receptor	心钠素C型受体	心房钠尿肽C型受体	心钠肽C型受体
7	Q9Y5Q5,Q9Z319	Atrial natriuteric peptide-converting enzyme	心钠素转化酶	心房钠尿肽转化酶	心钠肽转化酶
8	P61436,P61437	Heat shock 10 kDa protein	热激蛋白10	热休克蛋白10	
9	P08110	Heat shock 108 kDa protein	热激蛋白108	热休克蛋白108	
10	Q60446,Q61699,Q92598	Heat shock 110 kDa protein	热激蛋白110	热休克蛋白110	
11	P26210	Heat shock 12 kDa protein	热激蛋白12	热休克蛋白12	
12	Q12988	Heat shock 17 kDa protein	热激蛋白17	热休克蛋白17	
13	Q12315,Q3ZBK7,Q4KLN4,Q53GS7,Q5RAS2,Q6DF	Nucleoporin GLE1	核孔蛋白GLE1		
14	P08523,P47874,Q64288	Olfactory marker protein	嗅觉标记蛋白		
15	O14463,O17486,O22022,O30974,O51088,O51890,	Thioredoxin	硫氧还蛋白		
16	P0A4L1,P0A4L2,P0AA25,P0AA26,P0AA27,P0AA2	Thioredoxin 1	硫氧还蛋白1		
17	P00275	Thioredoxin C-1	硫氧还蛋白C-1		
18	O64394,O64432,O65049,P68176,P80028,P84564,(	Thioredoxin H-type	硫氧还蛋白H型		
19	P22217	Thioredoxin I	硫氧还蛋白I		
20	P00734,P00735,P18292,P19221,P84122,Q5R537	Thrombin light chain	凝血酶轻链		
21	P25116,P26824,P30558,P47749,P56488,Q00991	Thrombin receptor	凝血酶受体		
22	P55085,P55086,Q63645	Thrombin receptor-like 1	凝血酶受体样蛋白1		

Fig. 2. Screenshot of Part of English-Chinese Controlled-vocabulary dictionary for proteins

4 Text Mining

Based on English-Chinese Controlled-vocabulary dictionary for proteins, text mining then can be performed on selected data records that qualify the standards of CBA. During the course of text mining, following works have been conducted:

- Text protein mining principle and algorithm design are as follows,
  Text protein mining principle schematic diagram is shown as Fig. 3

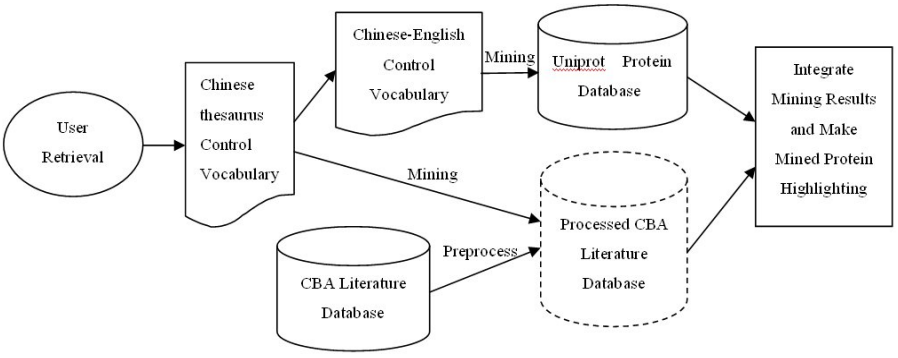


Fig. 3. Text protein mining principle schematic diagram

Before text protein mining was performed, the CBA literature database was preprocessed. Afterwards,text mining was realized through the functional expression mine_with_one() using C++ language with the following basic algorithm:

```
size_t mine_with_one(const std::string& text, const
std::string& key)

{
    std::string copy = text;
    size_t offset = 0;
    size_t pos = copy.find(key);
    if(pos == string::npos)
    {
        std::string key2 = key;
        key2[0] = toupper(key2[0]);
        pos = copy.find(key2);
    }
    while(pos != string::npos)
    {
        bool ok1 = true;
        bool ok2 = true;
        size_t p = pos;
        if(p + offset >= 1) {
            ok1 = isDelimiter(text[offset + p - 1]);
        }
        p = pos + key.length();
        ok2 = (p == copy.size() - 1) || (offset + p ==
text.size()) ||
              isDelimiter(text[offset + p]) ||
isEnding(text[offset + p]);

        if(ok1 && ok2)
        {
            return pos + offset;
        }
        if(pos  + key.length() >= copy.length())
        {
            return string::npos;
        }
        offset += pos + 1;
        copy = copy.substr(pos+1);
        pos = copy.find(key);
    }
    return std::string::npos;
}
```

Furthermore, link and integration between CBA literature database and Uniprot protein database was realized through the functional expression makeup_text() using C++ language with the following basic algorithm:

```
  std::string makeup_text(const std::string& text, const
StrStrMap& word)
  {
      typedef std::map<size_t, CtrlRecord> ReplaceMap;
      typedef std::pair<size_t, CtrlRecord> ReplacePair;
      ReplaceMap result;
      SubStrList strList;
      strList.push_back(SubStrElement(text, 0));
      StrStrMap::const_iterator cit;
      SubStrList::iterator lit;
      for(cit = word.begin(); cit != word.end(); cit++)
      {
          std::string key = cit->first;
          for(lit     =     strList.begin();     lit     !=
strList.end(); ++lit)
          {
              size_t offset = lit->offset;
              size_t pos = mine_with_one(lit->str, key);
              if(pos == string::npos)
              {
                  continue;
              }
              result.insert(ReplacePair(pos    +    offset,
CtrlRecord(key, cit->second)));
              SubStrList::iterator newlit = lit;
              if(pos < lit->str.length() - key.length())
                  {
                  newlit       =       strList.insert(lit,
SubStrElement(lit->str.substr(pos + key.length()),
                    offset + pos + key.length()));
              }
              newlit       =       strList.insert(newlit,
SubStrElement(lit->str.substr(0, pos), offset));
              strList.erase(lit);
              lit = newlit;
              offset += pos + key.length();
          }
      }
      QString
urlTemp("http://www.uniprot.org/uniprot/?query=name:%tname");
```

```
        QString  hrefTemp("<a  class='external  textmined'
target='_blank' href='%url'>%name</a>");
        std::string makeup;
        size_t start = 0;
        ReplaceMap::iterator i;
        for(i = result.begin(); i != result.end(); i++)
        {
            QString url = urlTemp;
            QString href = hrefTemp;
            url.replace("%tname",       QString::fromUtf8(i-
>second.tname.c_str()));
            href.replace("%name",       QString::fromUtf8(i-
>second.name.c_str()));
            href.replace("%url", url);
            makeup.append(text.substr(start,    i->first    -
start));
            makeup.append(std::string(href.toUtf8()));
            start = i->first + i->second.name.length();
        }
        makeup.append(text.substr(start));
        return makeup;
    }
```

- Protein identification: If data records contain any proteins that are included in the dictionary, these proteins are identified and tagged. In search results, they can be highlighted by simply clicking a button (See figure 4).
- Feature annotation: two features, namely functional/structural feature and soluble/insoluble feature are currently annotated. More features can also be annotated if necessary.

Fig. 4. A screenshot of CBA search interface

- Link generation: In order to integrate identified proteins with UniProt database, links that follow linking rules of UniProt database are generated based on their accession numbers or names. Links are attached with Chinese protein names in search results and users can click these links to access relevant protein information in UniProt database. Thus we establish links between a Chinese bibliographic database with the factual protein science databases.

5 Protein Name-Based Bilingual Retrieval

Both English protein names and Chinese protein names can be used to perform retrieval in CBA. Based on the English-Chinese Controlled-vocabulary dictionary for proteins, CBA can return search results that contain only Chinese protein names (See figure 4).

6 Search Optimization

When the CBA system was up and running, at first we found that both the recall rate and precision rate were unsatisfactory after trial use. After careful analysis and study, causes are identified and search optimizations are conducted as follows:

- Since protein names may have synonyms and acronyms, and people may write them in different ways, such as the use of dashes (e.g." insulin-like growth factor I " vs " insulin like growth factor I "), or Arabic numbers vs Roman numbers (e.g. "protein I" vs "protein 1"), it is necessary and possible to optimize searches so as to ensure the recall rate while not compromising the precision rate. Otherwise even trivial differences in protein names may often cause failure of queries. So before search criteria are submitted for retrieval, optimization should be performed to take into account of above cases. In these cases to treat them as equivalent can efficiently improve the recall rate.
- When part of a protein name is also a protein name, improper segmentation of names may cause the precision rate to fall rapidly because such case is very common, many protein names form this way. For example, “thioredoxin H-type” and “thioredoxin C-1” are protein names, as part of them, “thioredoxin” is also a valid protein name. If “thioredoxin H-type” or “thioredoxin C-1” appears in records and is improperly segmented as “thioredoxin”, when querying for “thioredoxin”, search results will include records containing “thioredoxin H-type” or “thioredoxin C-1”, which certainly are inaccurate results and decrease the precision rate. Furthermore, links generated subsequently are of course incorrect and will connect to information related to “thioredoxin” rather than “thioredoxin H-type” or “thioredoxin C-1”. To avoid this, name segmentation should be performed in such a way that longer names would have higher priority for segmentation. A long name is always segmented first and short names contained in this long name will be ignored and should not be segmented.

- Previously, CBA will stem words during the course of retrieval. For example, search words such as “convert”, “converts”, and “converting” are treated as equivalent. But as far as protein names are concerned, such stemming is relatively improper. In most cases, stemming an entered protein name may cause many unrelated records appearing in search results. So it is reasonable not to perform stemming if an entered word can be determined as a protein name. This rule is also suitable for other names, such as gene or organism names.

7 Feedback and Future Work

Since the CBA system was officially up and running in March 2011, we have got many feedbacks from users. Most of them love the convenience of easily searching protein names, locating highlighted protein names in search results, and accessing UniProt database for detailed protein information through links. They appreciate the idea of integrating Chinese bibliographic resources with authoritative factual protein science databases. But they also raised some questions and proposed many advices. Overall, however, the feedback has been very positive so far. According to users’ suggestions and problems we have discovered, following issues are currently being considered and actually some of them are being undertaken in order to further enhance the system and make it more efficient and convenient:

- To add more protein names and their features into the English-Chinese Controlled-vocabulary dictionary. This work is continuously being conducted and actually we also plan to add relationships of proteins and other relevant information so as to finally construct a Chinese protein ontology. Then it would be possible to realize semantic-based text mining and provide users with knowledge-based information service.
- To integrate more factual scientific databases, especially factual gene databases. Some users are also interested in other special fields, such as evidence-based medicine, AIDS, etc. If search results of a special topic from a bibliographic database can be integrated with relevant factual scientific databases, it is certainly very helpful and convenient for users. This is an interesting direction for information integration and knowledge mining.
- Based on the English-Chinese Controlled-vocabulary dictionary and CMeSH, further text mining can be performed to recognize concepts from user’s search query. Thus besides normal search results, records containing terms relevant to these concepts can also be displayed. So this will provide users an easy way to access concept relevant information.

8 Conclusion

With the fast development of life science research, to satisfy user’s information needs is becoming an inevitable challenge. Information integration and mining are playing a more and more important role. But almost all successful works are based on western

language literatures so far. This paper describes our efforts to mine protein related information and knowledge from the CBA, and integrate with relevant information in the Uniprot database. This is the first time that a Chinese bibliographic database establishes the link with western language factual scientific databases. Users love the convenience of easily searching protein names, locating highlighted protein names in search results, and accessing UniProt database for detailed protein information and knowledge through links. The work would be helpful for utilization and mining of Chinese bibliographic resources, as well as cross lingual information retrieval, integration, and mining. We hope with further enhancements, CBA can provide better services for its users and acts as a hub for Chinese life science information service.

Acknowledgements. This work is supported by Science and Technology Commission of Shanghai Municipality and Shanghai Municipal Bureau of Press and Publication (Grant No.08dz1501000), Shanghai Pujiang Program, The Choosing Excellent Program for the Outstanding Talent Introduced in Chinese Academy of Sciences in The Fields of Bibliographical Information and Periodical Publication and 2011 Innovation and Teaching Scientific Research Team Plan of Shanghai International Studies University. We wish to thank all participants of this project at Shanghai Information Center for Life Sciences of Chinese Academy of Sciences and Shanghai International Studies University. We are also grateful to professor Weimin Zhu for his technical direction and help in the database construction and information integration.

References

1. Doms, A., Schroeder, M.: GoPubMed: exploring PubMed with the gene ontology. Nucleic Acids Research 33, 783–786 (2005)
2. Maglott, D., Ostell, J., Pruitt, K., Tatusova, T.: Entrez Gene: gene-centered information at NCBI. Nucleic Acids Research 39, 52–57 (2011)
3. Pasquier, C.: Biological data integration using Semantic Web technologies. Biochimie 90, 584–594 (2008)
4. McGarry, K., Garfield, S., Morris, N.: Recent trends in knowledge and data integration for the life sciences. Expert Systems 23(5), 330–341 (2006)
5. Alexopoulou, D., Wächter, T., Pickersgill, L., Eyre, C., Schroeder, M.: Terminologies for text-mining; an experiment in the lipoprotein metabolism domain. BMC Bioinformatics 9(suppl. 4), S2 (2008)
6. Sahoo, S., Bodenreider, O., Zeng, K., Sheth, A.: An experiment in integrating large biomedical knowledge resources with RDF: Application to associating genotype and phenotype information. In: 16th International World Wide Web Conference (WWW 2007) on Health Care and Life Sciences Data Integration for the Semantic Web, Banff, Canada, pp. 8–12 (2007)
7. Jain, E., Bairoch, A., Duvaud, S., et al.: Infrastructure for the life sciences: design and implementation of the UniProt website. BMC Bioinformatics 10, 136–136 (2009)
8. Bairoch, A., Apweiler, R., Wu, C., et al.: The Universal Protein Resource (UniProt). Nucleic Acids Research 33, 154–159 (2005)

An Automated Search Space Reduction Methodology for Large Databases

Angel Kuri-Morales

Departamento de Computación, Instituto Tecnológico Autónomo de México, Mexico
akuri@itam.mx

Abstract. Given the present need for Customer Relationship and the increased growth of the size of databases, many new approaches to large database clustering and processing have been attempted. In this work we propose a methodology based on the idea that statistically proven search space reduction is possible in practice. Following a previous methodology two clustering models are generated: one corresponding to the full data set and another pertaining to the sampled data set. The resulting empirical distributions were mathematically tested by applying an algorithmic verification.

Keywords: Large databases, Sampling, Space reduction, Preprocessing, Clustering.

1 Introduction

Commercial enterprises are importantly oriented to continuously improving customer-business relationship. With the increasing influence of CRM[1] Systems, such companies dedicate more time and effort to maintain better customer-business relationships. The effort implied in getting to better know the customer involves the accumulation of very large data bases where the largest possible quantity of data regarding the customer is stored.

Data warehouses offer a way to access detailed information about the customer's history, business facts and other aspects of the customer's behavior. The databases constitute the information backbone for any well established company. However, from each step and every new attempted link of the company to its customers the need to store increasing volumes of data arises. Hence databases and data warehouses are always growing up in terms of number of registers and tables which will allow the company to improve the general vision of the customer.

Data warehouses are difficult to characterize when trying to analyze the customers from company's standpoint. This problem is generally approached through the use of data mining techniques [1]. However, to attempt direct clustering over a data base of several terabytes with millions of registers results in a costly and not always fruitful effort. There have been many attempts to solve this problem. For instance, with the

[1] Customer Relationship Management.

P. Perner (Ed.): ICDM 2013, LNAI 7987, pp. 11–24, 2013.

use of parallel computation, the optimization of clustering algorithms, via alternative distributed and grid computing and so on. But still the more efficient methods are unwieldy when attacking the clustering problem for databases as considered above.

In this article we present a methodology derived from the practical solution of an automated clustering process over large database from a real large sized (over 25 million customers) company. We emphasize the way we used statistical methods to reduce the search space of the problem as well as the treatment given to the customer's information stored in multiple tables of multiple databases. This paper is an improved extension of [21] where the behavior of the sampled variables was modeled empirically. Here we advance a method which allows us to avoid the limited search of hand-picked models in order to verify the equivalence between the universe and the sample.

For confidentiality the name of the company and the actual final results of the customer characterization are withheld.

Paper Outline

The outline of the paper is as follows. First, we give an overview of the analysis of large databases in section 2; next we give a clustering, sampling, and feature selection overview. In section 3 we briefly discuss the case study treated with the proposed methodology. Explanation of the methodology follows in Section 4. Finally, we conclude in Section 5.

2 Analysis of Large Databases

To extract the best information of a database it is convenient to use a set of strategies or techniques which will allow us to analyze large volumes of data. These tools are generically known as data mining (DM) which targets on new, valuable, and nontrivial information in large volumes of data. It includes techniques such as clustering (which corresponds to non-supervised learning) and statistical analysis (which includes, for instance, sampling and multivariate analysis).

2.1 Clustering in Large Databases

Clustering is a popular data mining task which consist of processing a large volume of data to obtain groups where the elements of each group exhibit quantifiably (under some measure) small differences between them and, contrariwise, large dissimilarities between elements of different groups. Given its importance as a very important data mining task, clustering has been the subject of multiple research efforts and has proven to be useful for many purposes [3].

Many techniques and algorithms for clustering have been developed, improved and applied [4], [5], [6]. Some of them try to ease the process on a large database as in [7], [8] and [9]. On the other hand, the so-called "Divide and Merge" [10] or "Snakes and Sandwiches" [11] methods refer to clustering attending to the physical storage of the records comprising data warehouses. Another strategy to work with a large database is based upon the idea of working with statistical sampling optimization [12].

2.2 Sampling and Feature Selection

Sampling is a statistical method to select a certain number of elements from a population to be included in a sample. There exist two sampling types: probabilistic and nonprobabilistic. For each of these categories there exists a variety of sub methods. The probabilistic better known ones include: a) Random sampling, b) Systematic sampling and c) Stratified sampling. On the other hand the nonprobabilistic ones include methods such as convenience sampling, judgment sampling and quota sampling. There are many ways to select the elements from a data set and some of them are discussed in [13]. This field of research, however, continues to be an open one [14], [15].

The use of sampling for data mining has received some criticism since there is always a possibility that such sampling may hamper a clustering algorithm's capability to find small clusters appearing in the original data [12]. However, small clusters are not always significant; such is the case of costumer clusters. Since the main objective of the company is to find significant and, therefore, large customer clusters, a small cluster that may not be included in a sample is not significant for CRM.

Apart from the sampling theory needed to properly reduce the search space, we need to perform feature selection to achieve desirable smaller dimensionality. In this regard we point out that feature selection has been the main object of many researches [16], [17], and these have resulted in a large number of methods and algorithms [18]. One such method is "multivariate analysis". This is a scheme (as treated here) which allows us to synthesize a functional relation between a dependent and two or more independent variables. There are many techniques to perform a multivariate analysis. For instance, multivariate regression analysis, principal component analysis, variance and covariance analysis, canonical correlation analysis, etc., [19]. Here we focus on the explicit determination of a functional which maximizes the resulting correlation coefficient while minimizing its standard error. In the present approach this is solved by automatically a) Determining the degree of a polynomial expansion and b) Calculating the coefficients of such expansion, as will be discussed in the sequel.

3 Case Study

A data mining project was conducted for a very large multi-national Banking company (one of the largest in the world) hereinafter referred to as the "Company". The Company has several databases with information about its different customers, including data about services contracted, services' billing (registered over a period of several years) and other pertinent characterization data. The Company offers a large variety of services to millions of users in several countries. Its databases are stored on IBM Universal Database version 7.0. In our study we applied a specific data mining tool (which we will refer to as "the miner") which works directly on the database. We also developed a set of auxiliary programs intended to help in data pre-processing.

The actual customer information that was necessary for the clustering process was extracted from multiple databases in the Company. Prior to the data mining process,

the Company's experts conducted an analysis of the different existent databases and selected the more important variables and associated data related to the project's purpose: to determine the behavior of strategic variables in order to expedite decision making. Due to the variety of platforms and databases, such process of selection and collection of relevant information took several months and several hundred man-hours.

The resulting database displayed a table structure that contains information about the characteristics of the customers, products or services contracted for the customer and monthly billing data over a one year period.

To test the working methodology the project team worked with a set of 29,112,157 registers, from a selected subset of 52 strategic variables divided in 3 consolidated data tables.

Main Objective

As stated above, the main objective of the data mining project was to characterize the customers of the Company allowing in the near future - in accordance to customer characteristics - to offer a faster and simpler analysis of the behavior of the strategic variables such that decision making is more agile while preserving its accuracy.

4 Methodology

In order to apply a methodology whereupon the search space is efficiently and effectively reduced it is necessary to comply with several steps leading to the adequate representation and/or behavior of the data regardless of its primary origin. These steps are discussed in what follows.

- Data preprocessing
- Search space reduction
- Clustering

4.1 Data Preprocessing

This step included data cleaning by exhaustively searching for incomplete, inconsistent or missing data [20]. Resulting from this process unrecoverable registers were deleted. The number of such deleted records, however, was not significant.

From the original multiple-tables structure we defined a single-table view structure for which a process of denormalization was performed. This followed from an analysis of the key-structure. In this view tables with the same key were merged and tables with different keys were included in the referenced tables as additional columns. The transformation resulted in a view with a structure with 52 attributes.

4.2 Search Space Reduction

To reduce the search space we work with the original data to obtain a sample which is not only a subspace but, rather, one that properly represents the original (full) set of data. We reduce the set both horizontally (reducing the number of tuples) and vertically (reducing the number of attributes) to obtain the "minable view". Simultaneous reduction - horizontal and vertical - yields the smallest representation of the original data set. Vertical reduction is possible from traditional statistical methods, while horizontal reduction, basically, consists of finding the best possible sample. The following subsections discuss how we performed both reductions.

Vertical Reduction

To perform vertical reduction, multivariate analysis is required. There exist many methods to reduce the original number of variables. Here we simply used Pearson's correlation coefficients. An exploration for correlated variables was performed over the original data. We calculated a correlation matrix for 87 variables. We considered (after consulting with the experts) that those variables exhibiting a correlation factor equal or larger than 0.75 were redundant. Hence, from the original 87 variables only 52 remained as informationally interesting. In principle, out of a set of correlated variables only one is needed for clustering purposes. Which of these is to be retained is irrelevant; in fact, we wrote a program which simply performed a sequential binary search to select the (uncorrelated) variables to be retained.

Horizontal Reduction

This step is based on the hypothesis that a sample will adequately represent the full set of data if the information present in the sample is approximately equal to the one of the universe. To this effect we appealed to simple concepts of Shannon's information theory. The size of the sample was determined by incrementally calculating the entropy H(S) of the data base from

$$H(S) = \sum_{i=1}^{N} P(S_i) I(S_i)$$

Where S_i is a symbol of the data set, $P(S_i)$ is the probability that this symbol is randomly selected and $I(S_i)$ is the information in the symbol; $I(S_i) = -log_2[P(S_i)]$. To begin with the symbols are unknown. The data base is sampled and every new (unseen) symbol is added to a list. If a symbol is already on the list a counter is incremented and the $P(S_i)$ is approximated from its proportion relative to the symbols that have been read. In every step $H(S)$ is calculated and compared to the value of the previous iteration. When the difference between the two last values of $H(S)$ falls below a predefined threshold it is considered that sampling new elements from the data base will not supply significant new information and the process stops. This process is schematized in figure 1.

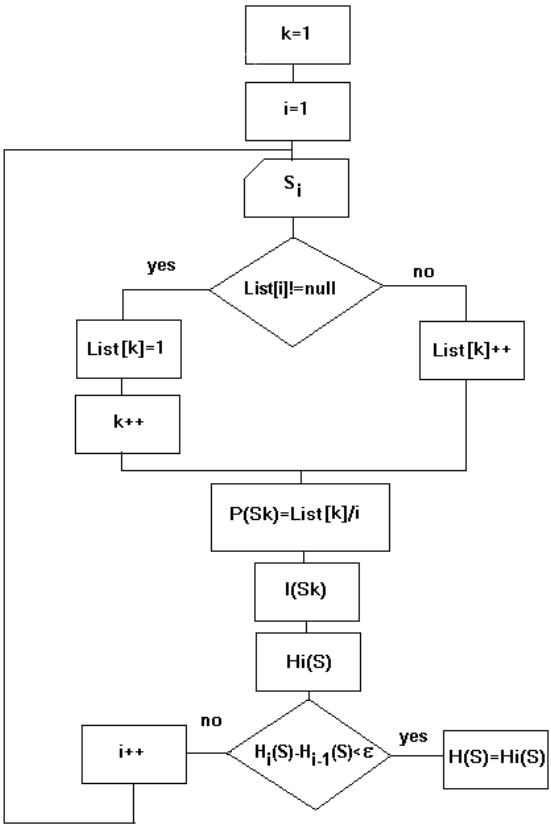


Fig. 1. Incremental entropy calculation

The process is iterated for all the (say *n*) variables. Every iteration yields a maximum *i* which we denote as C_i and the size of the sample corresponds to the largest one of the *n* (call it C_X). From the sample we validated the representative adequacy of this subset. A central issue to our work was the way the sample is validated. The process consists of the following steps:

1. Select two random samples of size C_X.
2. Select n_θ (see below) couples of variables (To prove that, within each sample, the behavior of the selected variables is statistically equivalent).
3. Find the best regressive function of the selected couples in both samples. If the functions exhibit different behaviors (see below) then make C_X ← C_X+K (where K is an adequate increment size selected a priori) and go to step 1.
4. Perform steps 2 and 3 as long as there are more variables to evaluate.

In step 3 we programmatically analyzed, in every case, polynomial models of degrees from 2 to 8. It is well known that any analytic function may be closely approximated by a polynomial expansion, as illustrated in Figure 2.

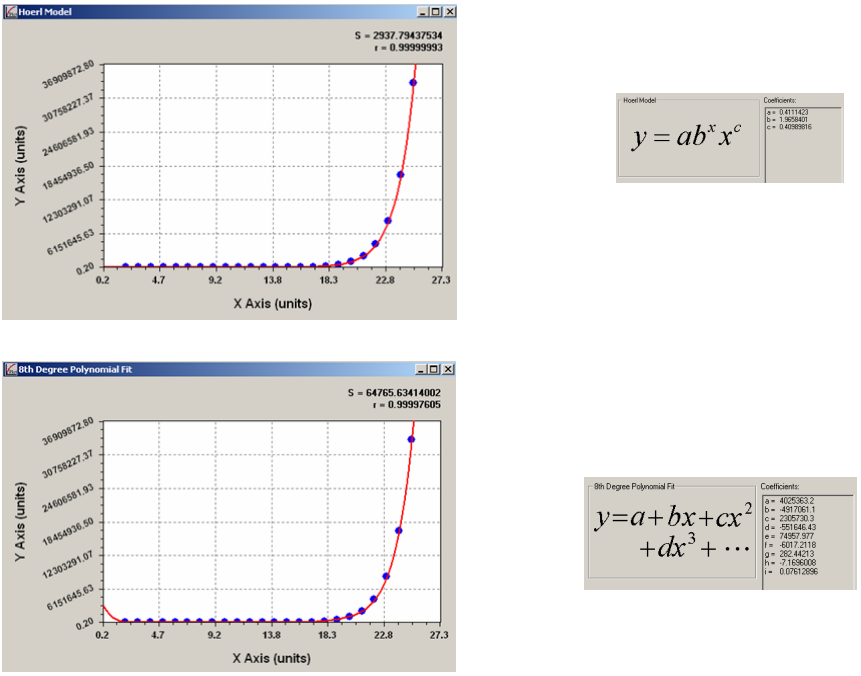


Fig. 2. Approximation of Hoerl's Model with a Polynomial of degree 8

Pairs of variables (v_1, v_2) were randomly selected and polynomials $P_k(v_2)$ of degree k were found (one for each of the two samples) such that $v_1 = P_k(v_2)$ for $k=2,\ldots,8$. Then the best fit for each of the two samples was compared, as shown in Figure 3. The headings "VAR1" and "VAR2" indicate which variables were selected; "BESTD1" and "BESTD2" indicate the degrees of the best approximation; "ABSDIF" indicates tha absolute approximation difference and "PERCDIF" the percentual difference. If the percentual difference between the polynomials of sample 1 and sample 2 was larger than 5% the samples were rejected.

In principle we should compare all variable's couples. The number of possible pairings is given by $\sum_{i=1}^{n-1} i = n(n-1)/2$. However, this (in general) implies a very large number of comparisons, with the need of the corresponding calculation of a very large number of polynomials. Therefore, we randomly selected n_θ couples of variables, where we set the proportion of variables that we wish to satisfy the minimum error criterion (which we denote with π) and the reliability of the sample (which we denote with γ). Let $N_\theta \subset N$ be the number of couples which behave as expected and n_θ its cardinality. Then $n_\theta = \pi N$ .The probability P_S that in a simple of size S

all the elements are in N_θ is $P_S = \frac{(\pi N)!(N-S)!}{N!(\pi N - S)!}$ We, therefore, specify that $\frac{n_\theta!(N-S)!}{N!(n_\theta - S)!} \leq 1-\gamma$. Then, given π and γ we may solve for the value of S and estimate the number of couples to test. For example, for *N=500*, π = *0.95*, γ=*.95* (N_θ=475) we have that *S=55*. That is, where an exhaustive analysis would imply solving 500 functions, sampling the indicated couples we just have to analyze 55 to ensure, with a 95% reliability that 475 of those couples would satisfy our requirement: that the two samples be similar.

VAR1	VAR2	BESTD1	BESTD2	ABSDIF	PERCDIF
10	3	7	8	0.0000445365	0.01%
16	10	3	7	0.0000196449	0.01%
24	16	4	6	0.0002663321	0.03%
2	1	8	3	0.0170915966	0.44%
20	16	4	8	0.0002044512	1.83%
7	20	7	4	0.0003220067	1.93%
19	1	8	6	0.0007318391	2.03%
10	15	8	8	0.0001581633	2.10%
5	1	8	8	0.0005979911	2.15%
13	3	5	2	0.0000306959	2.19%
4	10	8	2	0.0008559179	2.28%
24	11	8	7	0.0321750508	2.30%
19	5	8	8	0.0000132690	2.43%
5	27	5	8	0.0000291109	2.60%
10	8	7	5	0.0000142587	2.66%
1	5	6	6	0.0000155038	2.67%
27	7	8	4	0.0498684561	2.76%
15	10	6	6	0.0008779180	2.77%
21	7	8	8	0.0005055243	2.78%
23	3	8	6	0.0000228125	2.80%
19	13	8	6	0.0008211535	2.88%
25	21	8	5	0.2377070904	3.00%
20	17	5	8	0.0002921991	3.03%
4	26	8	7	0.0277383297	3.15%

Fig. 3. Comparison of best polynomial approximations in both samples

The probability of displaying the results shown in Figure 3. by chance alone is less than 10^{-12}. We must stress the fact that this analysis is only possible because we were able to numerically characterize each of the subsets. Furthermore, not only characterization was proven; we also showed that, in every case, the said characterization was similar when required and dissimilar in other cases.

4.3 Clustering Phase

Once the search space is reduced the clustering phase is reached. Before attempting the clustering proper, we impose certain a priori assumptions, as follows.

- The number of clusters is to be determined automatically (without applying any aprioristic rules).
- The "best" number (N) of clusters is derived from information theoretical arguments.

- The theoretical N is to be validated empirically from the expert analysis of the characteristics of such clusters.

In order to comply with our assumptions we follow the next steps:

1. Consecutively obtaining the clusters (via a Fuzzy C Means algorithm) assuming n clusters for n=2, 3, …., k; where "k" represents the largest acceptable number of clusters.
2. Determine the "optimal" number of clusters according to "elbow" criterion [10].
3. Clustering with a self organizing map algorithm to find the optimal segmentation.

The minable view with the 129 variables was processed. The Fuzzy C Means (FCM) algorithm was used on the uncorrelated data and the elbow criterion was applied [21]. It is important to stress the fact that the use of fuzzy logic allows us to determine the content of information (the entropy) in every one of the N clusters into which the data set is divided. Other clustering algorithms based on crisp logic do not provide such alternative. Since the elements of a fuzzy cluster belong to all clusters it is possible to establish an analogy between the membership degree of an element in the set and the probability of its appearance. In this sense, the "entropy" is calculated as the expected value of the membership for a given cluster. Therefore we are able to calculate the partition's entropy PE (see below). Intuitively, as the number of clusters is increased the value of PE increases since the structure within a cluster is disrupted. In the limit, where there is a cluster for every member in the set, PE is maximal. On the other hand, we are always able to calculate the partition coefficient: a measure of how compact a set is. In this case, such measure of compactness decreases with N. The elbow criterion stipulates that the "best" N corresponds to the point where the corresponding *tendencies* of PE to increase and PC to decrease *simultaneously* change. That is, when the curvature of the graph of tendencies changes we are faced with an optimal number of clusters. Table 1 displays part of the numeric data values of PC and PE. These coefficients were calculated with formulas 1 and 2.

$$PC = \sum_{k=1}^{K} \sum_{i=1}^{c} \frac{(\mu_{ik})^2}{K} \tag{1}$$

$$PE = -\frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{c} \mu_{ik} \ln(\mu_{ik}) \tag{2}$$

Table 1. Numeric data for the elbow criterion

Clusters	2	3	4	5	6	7	8	9	10	11	12
PC	0.879	0.770	0.642	0.560	0.498	0.489	0.413	0.414	0.400	0.359	0.349
PE	0.204	0.436	0.639	0.812	0.982	1.036	1.220	1.224	1.272	1.403	1.433

Figure 4 shows the graph for the numeric data of table 3. In the graph the "elbow" point is located between the cluster 6 and 7, indicating that there is a high probability that the optimal number of clusters is in that point, i.e. N=6.

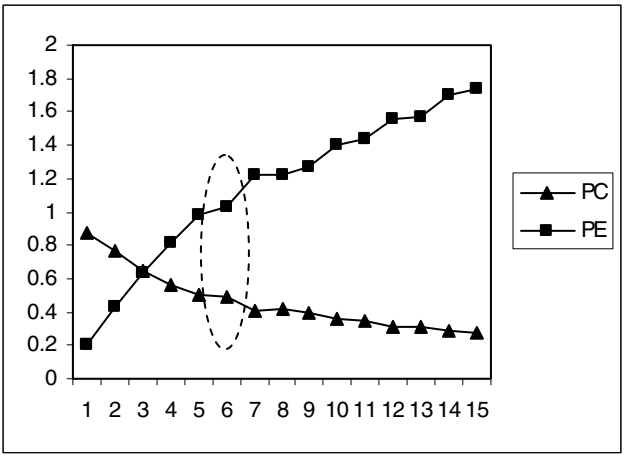


Fig. 4. Graph for the elbow criterion

The last phase of our analysis implied the use of the miner and the theoretically determined best number of clusters, as shown in figure 5.

The graph in figure 5 shows at the left side the percentage of elements grouped in each cluster. On the right the neuron number which represents the cluster. Each cluster shows the more important variables for the results, ordered by Chi-squared characterization of the variable's behavior in the cluster and in the whole sample. The cluster information for the Company can be extracted from the graph and reports supplied by the tool. We should now prove that clustering resulting from the reduced search space reflects a correct clustering view of the population.

4.4 Validation of the Reduced Search Space

To ease the understanding of the process in what follows we will call the clustering model from the sample "Model1"; likewise, we will call the clustering model derived from the complete data "Model2".

We followed the next steps:

1. Reduce the original data set only vertically.
2. Execute a clustering process over the full set of data to obtain Model2.
3. Label all the original data set and the sample data set with Model1 and Model2.
4. Compare the resulting distribution of elements labeled with both models.

The results are discussed in what as follows.

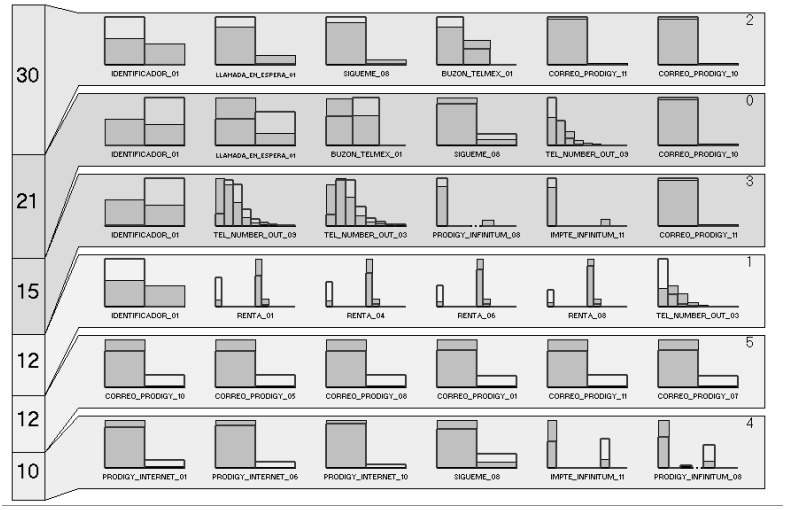


Fig. 5. Graph view of the clustering result supplied by the Miner

Comparison of Model1 and Model2

Table 2 shows the percentages for the two models. The names of the clusters were replaced by letters to avoid possible confusions with the neuron numbers shown in the Miner's results. As table 4 shows, the result clusters are very similar.

Table 2. Clusters' comparison for Model1 and Model2

Clusters	Model1 (%)	Model2 (%)	Difference (%)
A	30	27	3
B	21	20	1
C	15	18	3
D	12	15	3
E	12	12	0
F	10	8	2

Clustering from Sampling (Model1)

Having the clustering Model1, we labeled the sample data and the full data sets. The resulting distribution of elements into the different 6 clusters was expressed in percentages for comparison effects. As the table 5 shows the resulting distribution for the sample and for the full data are almost equal. This proves that the sample represents the full data set adequately.

Table 3. Labeling from the Model1 applied to the sampled and full data sets

Cluster	Sample (%)	Full Data Set (%)	Difference (%)
A	30.06	30.24	0.18
B	21.01	20.91	0.10
C	15.45	15.37	0.08
D	12.27	12.25	0.02
E	11.54	11.55	0.01
F	9.67	9.68	0.01

Cross Validation

Finally, we labeled the sample and the entire data with both algorithms for a set of selected variables. The results are shown in table 4.

Table 4. Comparison of Full and Sampled Data Clusters

Segment	Full Data Base	Sample 1	Sample 2
V1	0.0018	0.0018	0.0018
V2	0.1731	0.1731	0.17
V3	0.0011	0.0011	0.0013
V4	0.0192	0.019	0.0198
V5	0.0001	0.0001	0.0001
V6	0.3047	0.305	0.3023
V7	0.0911	0.0914	0.0891
V8	0.0001	0.0001	0.0001
V9	0.0002	0.0001	0.0002
V10	0.0093	0.0093	0.0094
V11	0.0026	0.0027	0.0026
V12	0.0012	0.0013	0.0012
V13	0.0254	0.0252	0.0254
TOTAL REGISTERS	**29,112,157**	**680,710**	**250,000**

As table 4 shows the differences between the distributions of elements into the clusters are similar between the two clustering models. Analog clusters share the same cardinality with a difference of less than 3%.

5 Conclusions

As we pointed out in the introduction, data mining may be an important strategic tool for commercial enterprises. But the management of large volumes of data (both physically and logically) may become a practical problem of large proportions and difficult to solve. Applying the methodology advanced herein it is possible to drastically reduce the size of the data base to be processed. In this case we were able to reduce the size in close to 0.60%. Originally we had to deal with 2,175 million elements (i.e 25,000,000 registers with 87 attributes each); instead we used a sample with only 13 million such elements (250,000 records with 52 attributes). The reduced sample, however, performed in a way that made it statistically indistinguishable from the original data. Apart from the benefit resulting from having quicker access to strategic information the use of this methodology yields economic benefits derived from the ability to process a smaller sample (increased speed and capacity for data processing; decreased amount of primary and secondary storage, costs of software and hardware, among others). Considering that the company had important improvements with the application of the results of this investigation, we consider that continuing research is needed and justified, since much work remains to be done if we wish to set a bound on the characteristics of the data which will allow us to generalize the results reported here.

Acknowledgments. Although the determination of the experimental probability distributions was achieved from the application of software especially designed by the author I wish to acknowledge that clustering was performed with SAS® Business Analytics Software.

References

1. Palpanas, T.: Knowledge Discovery in Data Warehouses. ACM SIGMOD Record 29(3), 88–100 (2000)
2. Silva, D.R., Pires, M.T.: Using Data Warehouse and Data Mining Resources for Ongoing Assessment of Distance Learning. In: IEEE ICALT Proceedings (2002)
3. Jain, K., Murty, M.N., Flynn, P.J.: Data Clustering: A Review. ACM Computing Surveys 31(3), 264–323 (1999)
4. Berkhin, P.: Survey of Clustering Data Mining Techniques. Accrue Software Inc. (2002)
5. Kleinberg, J., Papadimitriou, C., Raghavan, P.: Segmentation Problems. Journal of the ACM 51(2), 263–280 (2004)
6. Guha, S., Rastogi, R., Shim, K.: CURE: An efficient clustering algorithm for Large Databases. In: ACM SIGMOD Proceedings, pp. 73–84 (1998)
7. Peter, W., Chiochetti, J., Giardina, C.: New unsupervised clustering algorithm for large datasets. In: ACM SIGKDD Proceedings, pp. 643–648 (2003)
8. Raymong, T.N., Jiawei, H.: Efficient and Effective Clustering Methods for Spatial Data Mining. In: 20th International Conference on Very Large Data Bases, pp. 144–155 (1994)
9. Cheng, D., Kannan, R., Vempala, S., Wang, G.: A Divide-and-Merge Methodology for Clustering. In: ACM SIGMOD Proceedings, pp. 196–205 (2005)

10. Jagadish, H.V., Lakshmanan, L.V., Srivastava, D.: Snakes and Sandwiches: Optimal Clustering Strategies for a Data Warehouse. In: ACM SIGMOD Proceedings, pp. 37–48 (1999)
11. Palmer, C.R., Faloutsos, C.: Density Biased Sampling: An Improved Method for Data Mining and Clustering. In: ACM SIGMOD Record, pp. 82–92 (2000)
12. Liu, H., Motoda, H.: On Issues of Instance Selection. Data Mining and Knowledge Discovery 6(2), 115–130 (2002)
13. Zhu, X., Wu, X.: Scalable Representative Instance Selection and Ranking. In: Proceedings of the 18th IEEE International Conference on Pattern Recognition, pp. 352–355 (2006)
14. Brighton, H., Mellish, C.: Advances in Instance Selection for Instance-Based Learning Algorithms. Data Mining and Knowledge Discovery 6, 153–172 (2002)
15. Vu, K., Hua, K.A., Cheng, H., Lang, S.: A Non-Linear Dimensionality-Reduction Technique for Fast Similarity Search in Large Databases. In: ACM SIGMOD Proceedings, pp. 527–538 (2006)
16. Zhang, D., Zhou, Z., Chen, S.: Semi-Supervised Dimensionality Reduction. In: Proceedings of the SIAM International Conference on Data Mining (2007)
17. Fodor, I.K.: A survey of dimension reduction techniques. U. S. Department of Energy, Lawrence Livermore National Laboratory (2002)
18. Hair, J.F., Anderson, R.E., Tatham, R.L., Black, W.C.: Análisis Multivariante, 5th edn., pp. 11–15. Pearson Prentice Hall, Madrid (1999)
19. Delmater, R., Hancock, M.: Data Mining Explained: A Manager's Guide to Customer-Centric Business Intelligence, ch. 6. Digital press (2001)
20. Bezdek, J.C.: Cluster Validity with Fuzzy Sets. Journal of Cybernetics (3), 58–72 (1974)
21. Kuri-Morales, A., Erazo-Rodríguez, F.: A Search Space Reduction Methodology for Data Mining in Large Databases. Engineering Applications of Artificial Intelligence, 57–65 (February 1, 2009) ISSN: 0952-1976

Towards a High Productivity Automatic Analysis Framework for Classification: An Initial Study

Thomas Ludescher[1], Thomas Feilhauer[1], Anton Amann[2,3], and Peter Brezany[4]

[1] Fachhochschule Vorarlberg, University of Applied Sciences,
Hochschulstrasse 1, 6850 Dornbirn, Austria
{thomas.ludescher,thomas.feilhauer}@fhv.at
http://www.fhv.at
[2] Breath Research Institute of the Austrian Academy of Sciences,
Rathausplatz 4, A-6080 Dornbirn, Austria
[3] Univ.-Clinic for Anesthesia, Innsbruck Medical University
Anichstr 35, A-6020 Innsbruck, Austria
[4] Research Group Scientific Computing, Faculty of Computer Science,
University of Vienna, Währinger Strasse 29, A-1090 Vienna, Austria

Abstract. Due to the recent explosion of research data based on novel scientific instruments and corresponding experiments, automatic features, in particular in data analysis, has become more essential than ever. In this paper we present a new Automatic Analysis Framework (AAF) that is able to increase the productivity of data analysis. The AAF can be used for classifications, predictions and clustering. It is built upon the workflow engine Taverna, which is widely used in different domains and there exists a large number of Taverna activities for various kinds of analytical methods. The AAF enables scientists to modify our predefined Taverna workflow and to extend it with other available activities. For the execution of the analytical methods, in particular for the computation of the results, we use our own cloud-based Code Execution Framework (CEF). It provides web services to execute problem solving environment code, such as MATLAB, Octave, and R scripts, in parallel in the cloud. This combination of the AAF and CEF enables scientists to easily conduct time-consuming calculations without the need to manually combine potential combinations of independent variables. It furthermore automatically evaluates all identified models and provides service for the scientists conducting the analysis. The framework has been tested and evaluated with real breath gas data.

Keywords: classification, clustering, prediction, Automatic Analyse Framework.

1 Introduction

Productivity has been originally defined in economics [16] as the amount of output per unit of input, or, in other words, productivity is a ratio of production output to what is required to produce it (input). Now several research communities

P. Perner (Ed.): ICDM 2013, LNAI 7987, pp. 25–39, 2013.

in various domains started to approach productivity enhancement/enrichment features in data analysis applications for two reasons (a) cost reduction, and more important (b) to achieve better results and to get faster insights into scientific discoveries. To increase the productivity means directly to reduce the cost or to generate a better result/product. The productivity directly influences the required costs.

It is necessary to be able to deal with an increasing size of data. Some reasons for the growing amount of collected data are (a) data are usually permanently stored for further analysis, and (b) the emergence of novel scientific instruments with increase of quality and/or resolution of sensor devices. For example in the breath research domain the number of detected (protonated) compounds increased from about 200 to 2000 compounds by changing from PTR-MS (Proton Transfer Reaction Mass Spectrometry) to PTR-TOF-MS (Proton Transfer Reaction Time of Flight Mass Spectrometry) [1]. In this context automatization services for data analysis play an important role.

This paper presents the Automatic Analysis Framework (AAF) for increasing productivity in e-Science applications. It is used for automatic application of classification, clustering and prediction algorithms to evaluate and analyze data. It allows to execute different analytical methods for classification, prediction and clustering in parallel. Scientists select dependent and independent variables in their input data and start the analysis workflow, which executes all time consuming calculations in a dynamic cloud infrastructure. In our first prototype we use the Taverna workflow engine [2] and our self developed Code Execution Framework (CEF) [7] to execute the analytical methods. Our CEF is implemented as a Platform-as-a-Service (PaaS) cloud service type, which can be used to execute problem solving environment (PSE) code such as R [9], Octave [11] and MATLAB [17] in parallel. The Automatic Analysis Framework defines several interfaces (e.g. for the analytical method activity). Each researcher is able to use these interfaces to extend the AAF with his own methods/activities (e.g. new analytical method).

The AAF is placed between three interacting topics, namely analysis, productivity and automatization (Fig. 1). The system tries to calculate the best (a) settings (e.g. number and type of virtual machines) for the Code Execution Framework, and (b) parameter sets of the analysis (e.g. independent variables), to optimize and increase the productivity. It further shows how these topics are influencing each other in the context of the AAF. The productivity formula 1 can be used to optimize automatization or analysis parameters (e.g., number of worker nodes). With time consuming calculations the productivity can be increased by using the AAF in several different ways as follows:

- The total costs can be reduced due to (a) reduction of license/software costs, (b) reduction of maintenance costs, (c) reduction of ownership costs, (d) reduction of the hardware cost, and (e) reduction of the personnel cost.
- The time can be reduced due to (a) provided faster algorithms for the analysis, or (b) to provide parallel executable algorithms.

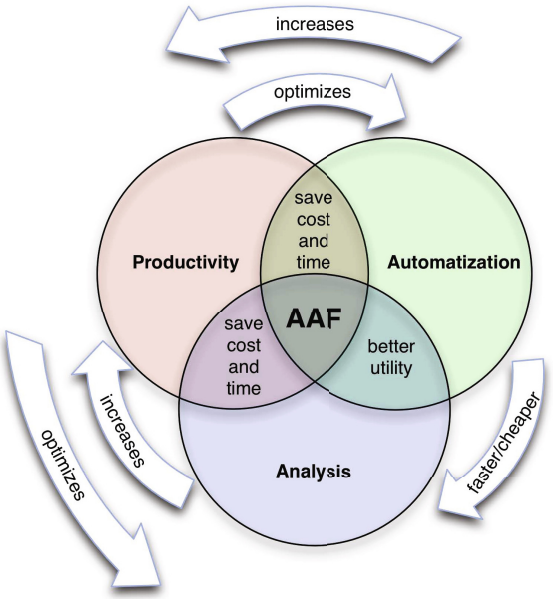


Fig. 1. AAF and the involved fields

– The Automatic Analysis Framework (AAF) helps researchers to (a) find new topics of interests or (b) get new insights in their existing research area, thus utility can be increased.

The framework has been evaluated on top of real data from the international breath research community [3,4].

The rest of this paper is organized as follows. In Section 2 we provide background information and related work, including productivity in general, the Taverna Workflow Management System and our Code Execution Framework. In Section 3 we describe our Automatic Analysis Framework (AAF) in detail and discuss its main architectural components. Section 4 exemplifies our first prototypical implementation on top of a concrete example from the breath research domain. Finally, in Section 5 we discuss further improvements and outline our next steps regarding the development of the AAF.

2 Background and Related Work

In this section we briefly discuss background material including the definition of productivity as utility over total costs, our self implemented Code Execution Framework, which is being used in AAF, and the open source Taverna workflow management system. Also related works are being discussed at the end of this section.

2.1 Productivity

Productivity is defined in economics and is a tool to evaluate the output (e.g. product) and the needed total costs to produce it [12]. In the following we briefly present a concrete example in oder to show what factors have strong influences on the productivity.

In a computer factory the productivity can be calculated by the number of computers produced divided by the needed working hours, or inversely the number of working hours to assemble a computer. The productivity can be increased by increasing the output or decreasing the needed working hours. This can exemplarily be done by (a) increasing the skills or motivation of the employees, (b) providing better working conditions (e.g., equipment), or (c) providing an automated plant. An example for increasing the utility function is to produce a better output in the same time/cost (e.g. use low power consumption electronic to increase the battery life time).

Jeremy Kepner defines the productivity as utility over the total cost [12] as listed below.

$$Productivity = \Psi = \frac{Utility}{Cost} \tag{1}$$

The Utility is the value a specific user or organization defines on getting a certain answer in a certain time [12].

2.2 Code Execution Framework

The Code Execution Framework can execute code of different problem solving environments, such as MATLAB, R and Octave, in parallel [7]. Most problem solving environments are implemented as single threaded programs; because of this constraint, the execution cannot use the power of current computers with multiple cores. The framework supports different cloud infrastructures, such as Amazon EC2 and Eucalyptus. Therefore it is possible to use hybrid cloud infrastructures, e.g. a private cloud based on Eucalyptus for general base-level computations using the available local resources and additionally a public Amazon EC2 for peak-load and time-critical calculations. The approach is to provide a secure platform that supports multiple problem solving environments, execute code in parallel with different parameter sets using multiple cores or machines in a cloud environment, and support researchers in executing code, even if the required problem solving environment is not installed locally.

The Code Execution Framework has a Kerberos based security concept for authentication[15]. The user identity will be forwarded down to cloud worker nodes.

2.3 Taverna

Taverna is an open source and domain-independent Workflow Management System [2]. Taverna provides several different activities for data analysis (e.g. classification, prediction, clustering). The common activities use the local or remote

machines for calculations. The AAF uses Taverna as workflow engine and we provide several new activities. Our statistical algorithm activities use the Code Execution Framework for the calculations and therefore a cloud based infrastructure is used. Taverna provides a lot of existing activities (e.g. CSV converter) that can easily be used within the provided AAF-workflow. With Taverna the AAF-workflow can be adopted and used for many different purposes.

2.4 Weka

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes [18]. With Weka a user can analyse different input data at a single machine, however with our AAF a researcher is enabled to use a large number of computers (e.g. within a cloud infrastructure) to allow a more detailed and faster analysis than with Weka.

3 Automatic Analysis Framework

This section describes the Automatic Analysis Framework workflow, provides a system overview of the AAF, and shows how the productivity is influenced by our system.

3.1 Workflow

The general workflow of the Automatic Analysis Framework (Fig. 2) consist of three main parts: (a) data preparation, (b) data analysis, and (c) report generation.

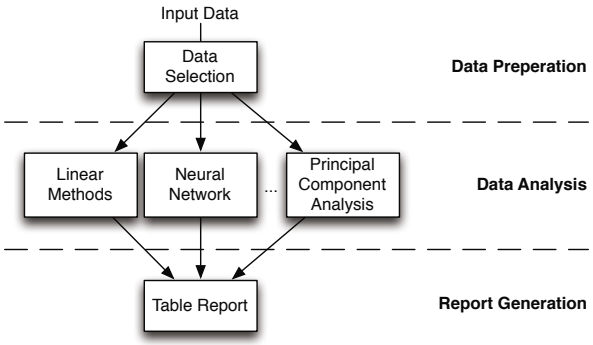


Fig. 2. AAF-Workflow

In the data preparation part the scientist is able to select the statistical analysis type (classification, prediction, or clustering), and define the independent and dependent variables. The data analysis part contains all different analytical methods (e.g. linear methods, neural network, principal component analysis). The report generation part uses the results of the analytical methods, orders the results (best classification models on top), and generates an HTML report.

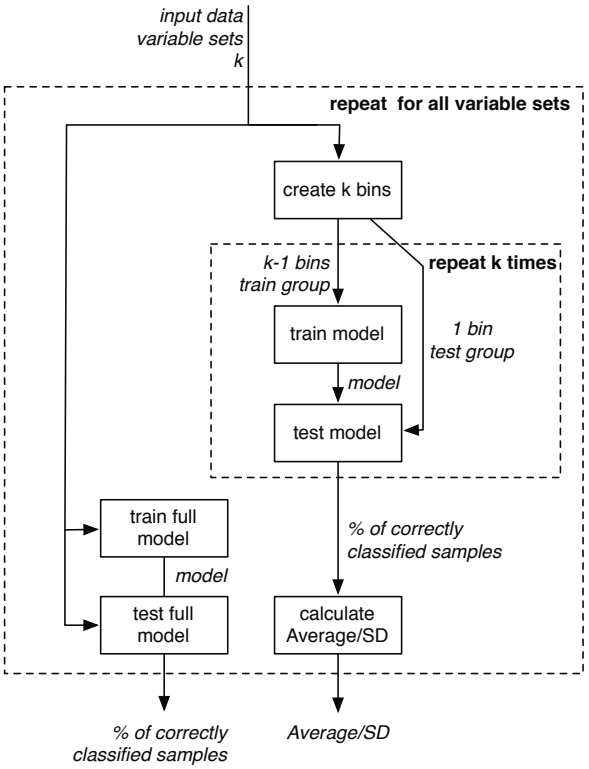


Fig. 3. Detailed evaluation workflow

For each analytical method (e.g. linear methods) a ranking of the results will be calculated (Fig. 3). To detect overfitting we decided to use k-fold cross validation [13]. Cross validation is independent of the used statistical analysis. Therefore we are able to use this evaluation for all classification algorithms. The input of such an evaluation is (a) the input data (CSV format), (b) the variable sets (one dependent variable and multiple independent variables), and (c) the k value (number of bins for the cross validation). At the beginning the input will be randomly partitioned into k equal sized sub samples. A single sub sample will be used for testing the model, the other (k-1) sub samples are used as training data. The result of one calculation is the percentage of correctly

classified samples. These steps will be repeated k-times (folds) for each sub sample. Afterwards we calculate the average and the standard deviation of the k results folds . Additionally we used the complete data set to train and test the model.

The output of one calculation is therefore (a) the percentage of correctly classified samples of the complete model, (b) the average value of the cross validation, and (c) the standard deviation of the cross validation. It is possible to detect overfitting if (a) the mean value of the cross validation is much better than the main value of the complete model or (b) the standard deviation is high. The AAF is able to calculate and evaluate a lot of different models (see Section 5). Therefore it is upmost important that the system is able to preselect/order the results.

3.2 System Overview

Presently, the Automatic Analysis Framework supports only classification with linear regression. In the future we plan to implement predication and clustering with several different methods as well.

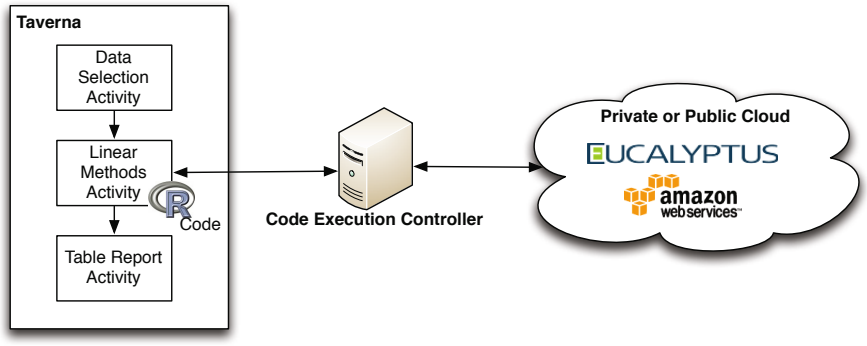


Fig. 4. System diagram

The system overview (Fig. 4) shows all involved sub systems. The Code Execution Controller (CEC) is the main component of the Code Execution Framework (CEF). It provides several different Web services (e.g. add or monitor calculations) and is responsible for (a) generating sub calculations, (b) start/stop virtual machines, (c) start new sub calculations in parallel, and (d) monitor running calculations and virtual machines. The CEF supports parallel code execution on the level of executing methods in parallel with different parameter sets. Our Code Execution Framework (CEF), which is being used for the execution of all analytical methods, is described in detail in [7]. The advantage of using CEF in this context is that the resources (Virtual Machines) can be added on the

fly, depending on the required CPUs or waiting calculations. The complete AAF workflow is implemented in Taverna, therefore we provided several different Taverna activities.

– The **Data Selection Activity** can be used to select the statistical analysis methods, such as classification, prediction or clustering. Additionally the user is able to select the independent and dependent variables (Fig. 5). The output of this activity is a list of independent semicolon separated parameter sets. Each set contains the dependent variable and a list of possible independent variables (pipe separated). The following line is an example of such a parameter set. Smoking is the dependent variable. Age, pulse, Acetonitrile, Benzene are the selected independent variables.

$$smoking; age|pulse|Aacetonitrile|Benzene$$

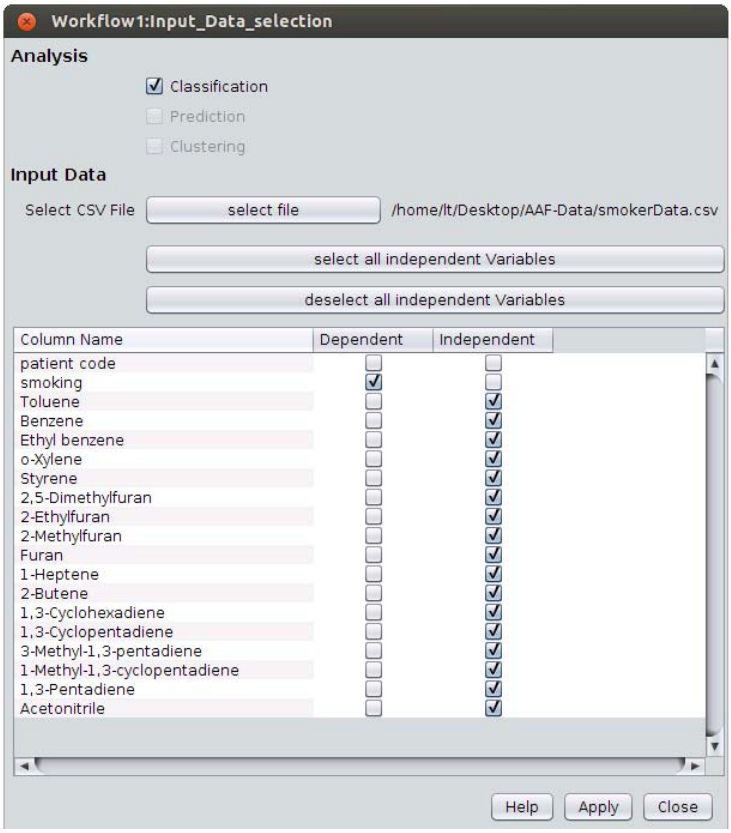


Fig. 5. Settings Data Selection Activity

– The **Linear Method Activity** provides the analytical methods. At the moment this activity supports only linear regression. The user must set the URL of the used Code Execution Framework. Additionally the user is able to define the k value for the k-fold cross validation (Fig. 6). The linear regression with cross validation is implemented in R and will be executed in the cloud based CEF. New analytical methods can easily be added by changing the already implemented R code.

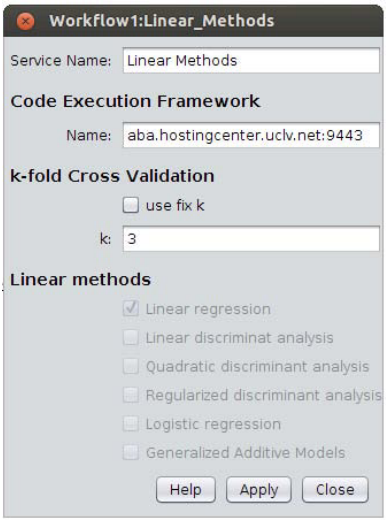


Fig. 6. Settings Linear Methods Activity

– The **Table Report Activity** uses the results of the analytical methods, orders the results (best classification models on top), and generates an HTML report.

3.3 How Productivity Is Influenced by the AAF

The AAF optimizes the productivity formula (1) by dynamically changing several parameters to increase the productivity. The researcher will be able to define some basic conditions, such as (a) the time-dependent utility function, (b) a limit for the costs for using a public cloud (e.g. Amazon EC2), (c) select the input data (Data Selection Activitiy), and (d) choose the analytical methods.

The following list shows how the AAF can automatically increase the productivity:

– The AAF can adopt the Code Execution Framework infrastructure automatically, by (a) adding a new VM while calculations are waiting, (b) using different cloud types (e.g. Eucalyptus or Amazon EC2), or (c) choosing different VM instance types (e.g. m1.small, m1.medium) [5]. These adoptions reduce the time but can affect the total costs negatively.

– If the input data consists of several independent variables, there exists a huge amount of different possibilities for analysis (e.g. all combinations of independent variables). For more information about this problem have a look to Section 4. The AAF can skip (a) not promising combinations of independent variables, or (b) even skip all analysis with a specific algorithm, e.g. if linear regression does not fit for some input data.

All above changes can automatically be handled from the AAF. The AAF dynamically adopts the cloud infrastructure and chooses the best configuration to fulfil all basic conditions.

4 Evaluation of the AAF

This section shows the usage of the AAF and the advantages of manual versus automatic analysis.

4.1 Example

The Autonomic Analysis Framework has been evaluated with data from the breath gas community [3]. This data describes concentrations of 17 different compounds in exhaled breath of 105 patients and volunteers (independent variables), and their smoking habits (smoker or non-smoker). The compounds were measured using gas chromatography with mass-spectrometric detection (GC-MS). The concentrations are given as (uncalibrated) peak areas. In our test example a researcher wants to explore the statistical results that can be achieved with this raw data.

AAF with One Single Independent Variable. First of all it is determined whether there exists a single compound that can be used to distinguish smokers from non-smokers. In this particular case the user has to (a) select the input data (URL to the CSV file), (b) select the column specifying the smoking habits as dependent variable, (c) select all other columns as independent variables, (d) choose the k-value for the cross validation (e.g. k=3), (e) set the maxIndependetVariables-input to 1 (otherwise all combinations of the independent variables are analyzed; for more information have a look at Section 5), and (f) specify the folder where the report should be stored. Subsequently the user can start the AAF workflow. To authenticate him-/herself to the CEF, the user has to enter his/her username (Kerberos principal) and the corresponding password.

The AAF generates a table (Fig. 7) that contains all test models, ordered by the number of correctly classified samples. As you can see, the AAF is able to classify smokers from non-smokers with 78.5% by using the single compound acetonitrile. The two cross validation values (mean, SD) show is the statistical quality of this result. Additionally the resulting table shows that there are several other substances that can be used to distinguish smoker and none smoker.

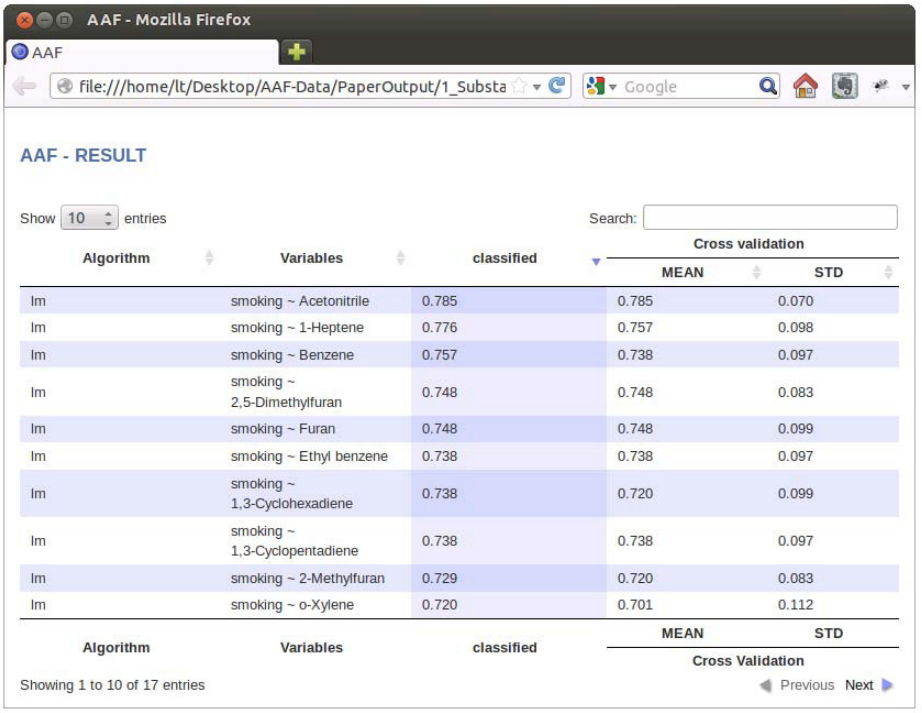


Fig. 7. AAF-Result with 1 independent variable

There exist several research studies, described in literature, that shows the same result, but use different kinds of input data, e.g. in [6,8,14] breath research data, and in [10] blood samples have been used. In all three research studies acetonitrile has been detected as a marker for classifying smokers from none-smokers.

Acetonitrile Boxplot and Histogram. As a second step the researcher may generate the boxplot or a histogram (Fig. 8) to verify the result. The boxplot compares the acetonitrile concentrations in exhaled breath of 105 smokers and non-smokers. The y-axis shows the chromatogram peak area of acetonitrile measured by gas chromatography with mass-spectrometric detection (GC-MS). The histogram shows the acetonitrile concentrations of 105 smokers and non-smokers.

AAF with Multiple Independent Variable. After the promising first results the user would like to check if it is possible to achieve better results by simultaneous analysis of two or three compounds. For this the user just has to set the maxIndependetVariables-input to 2 or 3 and rerun the AAF. The AAF now shows that it is possible to classify smokers from non-smokers with (a) two substances with 80.4%, and (b) with three substances with 82.2%.

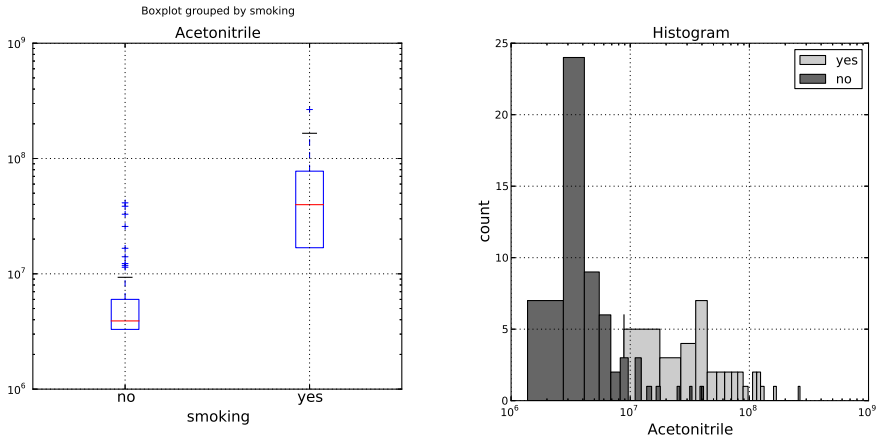


Fig. 8. Boxplot and histogram comparing acetonitrile concentrations in exhaled breath of 105 smokers and non-smokers

Optimization of the Results. As a next step the user may have a closer look at the original input data (CSV-file) and recognizes that some values are zero. To improve the result of the AAF the researcher has to identify if the value is 0 or not available and rerun the AAF cycle again.

4.2 Manual vs. Automatic Analysis

In many cases it is better to use as few independent variables as possible to reduce the likelihood of overfitting. In our test case we have 17 different independent variables, which results in 17 calculations if only a single independent variable is used at the same time. Let's imagine that the user would additionally like to analyse the models with all combinations of two independent variables. This results in 153 independent calculations (17 for a single independent variable and 136 for all combinations of two independent variables). If three variables are used, we have in total 833 independent calculations. The number of calculations is growing exponentially if the researcher would like to try more or even all different combinations. Without an automatic reduction of irrelevant combinations, a user or even a system is not able to calculate this huge amount of analysis in a reasonable time. If there is a low number of independent variables, e.g. around 10, the AAF can easily calculate all possible combinations. Having 10 independent variables results in a total of 1023 different combinations, i.e. choose 1 from 10 + choose 2 from 10 + choose 3 from 10 + ... + choose 10 from 10. With the AAF you can easily calculate the models for all 1023 different classifications, which would hardly be possible when done manually by a scientist. Scientists are further able to use the three different evaluation values, including *percentage of correctly classified samples*, *cross validation average*, and *cross validation standard deviation* to choose the best model of all.

This automatic analysis increases the productivity by reducing the time and the cost significantly. A person would need much longer for the manual configuration and initialization of all different combinations of the independent variables than tour AAF system.

5 Improvements and Future Work

There are many improvements and future research plans for this work. As a next step will adopt our Code Execution Framework to deal with these big amount of calculations. In particular we will combine several calculations to reduce the data transfer overhead.

The AAF must skip not promising combinations of independent variables. Without this improvement the required workflow execution time could be very high, depending on the input data. We will define a model that deals with exactly this problem.

We also will provide several different analytical methods including neural network, principal component analysis, cluster analysis, and so on. For each analytical method we will create a new Taverna activity. Mathematical calculations will be implemented in R to be executed with our CEF.

At the moment only classification can be done with the AAF, in the future we will support prediction and clustering as well.

Last but not least we would like to provide the possibility to choose an already trained model with new data for further usage of this model. Therefore we must store some additional data at the CEF and define a way to download an already trained model.

6 Conclusions

In this paper we have presented an Automatic Analysis Framework (AAF), which enables the user to pre-analyse existing data automatically. The main contribution of the AAF are (a) provide cloud based Taverna activities for classification, (b) define a Taverna workflow for the automatic analysis (can easily be extended with already existing activities), and (c) use a generally usable evaluation procedure (k-fold cross validation) for detecting overfitting.

AAF contributes to the recent developments of productivity enhancement frameworks in eScience applications.

The introduced Automatic Analysis Framework (AAF) is able to analyse input data (CSV). As our next step we will extend the AAF to enable classification, prediction, and clustering. At the moment it is possible to execute the linear regression (classification). The AAF tries all different combinations of independent variables to calculate the different models. Depending on the number of independent variables, the large numbers of calculations can be very time consuming. To deal with this calculation problem, we decided to use a cloud based Code Execution Framework (CEF). At the current stage the AAF can be used within the workflow engine Taverna. A user can easily define some boundaries for the

classification. Our first prototype has been evaluated by scientific applications from the breath research domain. Breath researchers are interested to receive hints and recommendations on what they can further classify, however such recommendations need to be based on some evidence. This is challenged by our AAF.

To test the AAF we used data from the breath gas community, but the whole system can be used with every other domain. For example, you can try to classify text into different categories (e.g. sports, news) or every other classification problem that can be solved with the provided algorithms.

The most important motivation for using the AAF is to improve the productivity. Therefore we must plan for all calculations to be able to (a) use unused resources in the cloud, (b) start new instances if required dynamically, (c) reduce the waiting time as far as possible, (d) generate a report, and (e) keep an eye on the total costs.

Acknowledgements. The funding of the ABA-Project (Project No. TRP 77-N13) by the Austrian Federal Ministry for Transport, Innovation and Technology and the Austrian Science Fund is key to bringing the partners together and to undertaking the research. The entire research team contributed to the discussions that led to this paper and provided the environment in which the ideas could be implemented and evaluated.

References

1. IONICON PTR-TOFMS Series (2012), http://www.ionicon.com/products/ptr-ms/ptrtofms/index.html
2. Taverna - open source and domain independent Workflow Management System (2012), http://www.taverna.org.uk
3. International Association for Breath Research (IABR), http://iabr.voc-research.at (accessed December 2012)
4. Journal of Breath Research, http://iopscience.iop.org/1752-7163 (accessed December 2012)
5. Amazon: Amazon EC2 Instance Types (2012), http://aws.amazon.com/ec2/instance-types/
6. Bajtarevic, A., Ager, C., Pienz, M., Klieber, M., Schwarz, K., Ligor, M., Ligor, T., Filipiak, W., Denz, H., Fiegl, M., Hilbe, W., Weiss, W., Lukas, P., Jamnig, H., Hackl, M., Haidenberger, A., Buszewski, B., Miekisch, W., Schubert, J., Amann, A.: Noninvasive detection of lung cancer by analysis of exhaled breath. BMC Cancer 9(1), 348 (2009), http://www.biomedcentral.com/1471-2407/9/348
7. Elsayed, I., Ludescher, T., Woehrer, A., Feilhauer, T., Brezany, P.: Data Life Cycle Management and Analytics Code Execution Strategies for the Breath Gas Analysis Domain. Procedia Computer Science 9, 156–165 (2012), http://www.sciencedirect.com/science/article/pii/S187705091200138X; Proceedings of the International Conference on Computational Science, ICCS 2012

8. Filipiak, W., Ruzsanyi, V., Mochalski, P., Filipiak, A., Bajtarevic, A., Ager, C., Denz, H., Hilbe, W., Jamnig, H., Hackl, M., Dzien, A., Amann, A.: Dependence of exhaled breath composition on exogenous factors, smoking habits and exposure to air pollutants. Journal of Breath Research 6(3), 036008 (2012), http://stacks.iop.org/1752-7163/6/i=3/a=036008
9. R Project Foundation, The R Project for Statistical Computing, http://www.r-project.org (accessed December 2012)
10. Houeto, P., Hoffman, J.R., Got, P., Dang, V., Baud, F.J.: Acetonitrile as a possible marker of current cigarette smoking. Hum. Exp. Toxicol. 16(11), 658–661 (1997), http://www.biomedsearch.com/nih/Acetonitrile-as-possible-marker-current/9426367.html
11. Eato, J.W.: Octave (2012), http://www.gnu.org/software/octave
12. Kepner, J.: High Performance Computing Productivity Model Synthesis. The International Journal of High Performance Computing Applications 4(18), 505–516 (2004)
13. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th International Joint Conference on Artificial intelligence, IJCAI 1995, vol. 2, pp. 1137–1143. Morgan Kaufmann Publishers Inc., San Francisco (1995), http://dl.acm.org/citation.cfm?id=1643031.1643047
14. Kushch, I., Schwarz, K., Schwentner, L., Baumann, B., Dzien, A., Schmid, A., Unterkofler, K., Gastl, G., Španěl, P., Smith, D., Amann, A.: Compounds enhanced in a mass spectrometric profile of smokers' exhaled breath versus non-smokers as determined in a pilot study using ptr-ms. Journal of Breath Research 2(2), 026002 (2008), http://stacks.iop.org/1752-7163/2/i=2/a=026002
15. Ludescher, T., Feilhauer, T., Brezany, P.: Security Concept and Implementation for a Cloud Based E-science Infrastructure. In: 2012 Seventh International Conference on Availability, Reliability and Security, pp. 280–285 (2012)
16. OECD: Measuring Productivity - OECD Manual. OECD Publishing (2001) http://www.oecd-ilibrary.org/content/book/9789264194519-en
17. The MathWorks: Matlab - The Language of Technical Computing, http://www.mathworks.com/products/matlab (accessed December 2012)
18. Weka 3: Data Mining with Open Source Machine Learning Software in Java (2012), http://www.cs.waikato.ac.nz/~ml/weka/

Extending Statistical Models for Batch-End Quality Prediction to Batch Control

Geert Gins, Jef Vanlaer, Pieter Van den Kerkhof, and Jan F.M. Van Impe

BioTeC, Department of Chemical Engineering, KU Leuven
W. de Croylaan 46 PB 2423, B-3001 Heverlee (Leuven), Belgium
{geert.gins,jef.vanlaer,pieter.vandenkerkhof,jan.vanimpe}@cit.kuleuven.be

Abstract. The control and optimization of batch processes is a challenging problem faced by (bio)chemical industry. Traditionally, (full) factorial tests are executed to investigate the effect of the manipulated variables (MV) on the (quality of the) process. Due to their nature, these tests are very time-consuming for batch processes. This paper investigates whether suitable data-driven models for batch optimization and control can be identified from a more limited set of tests.

Based on the results of two case studies, it is concluded that statistical inference models can predict the final quality of batches where the MV changes occur at time points not present in the training data, provided they fall inside of the time range used for training. Furthermore, the models provide accurate predictions for batches with multiple MV changes. This is a valuable result for industrial acceptance because it implies that fewer experiments are required for model identification.

Keywords: Partial Least Squares, (bio)chemical batch processes, quality prediction, optimization, process control.

1 Introduction

In this paper, it is investigated whether statistical inference models can be employed to control and optimize the production of (bio)chemical products in the chemicals and life sciences industry. The main challenge is the identification of accurate data-driven inference models from a very limited set of training data.

Chemicals and life sciences industry manufactures numerous products with a high added value (e.g., medicines, enzymes, performance polymers, additives, etc.). These products are typically produced in batch reactors in order to reduce capital investment costs and increase the flexibility of the manufacturing plant. For some goods, continuous production is impossible due to chemical constraints.

One of the major challenges in operating a batch process is accurately controlling the final quality. First of all, a batch process is dynamic by design. In addition, measurements of the quality parameter are often simply not available online: a sensor for the quality parameter might not exist (yet), it might be too expensive, or the quality might not be present yet. Hence, feedback control is not possible because the current quality of the batch cannot be easily assessed.

P. Perner (Ed.): ICDM 2013, LNAI 7987, pp. 40–54, 2013.

An open-loop control is often followed in practice: a fixed operational recipe is followed as closely as possible in an attempt to reduce variations in quality.

Statistical Process Control (SPC) provides an answer to this problem. SPC mines the information from online measurements (temperatures, pressures, flow rates, ...) to monitor the batch process. Deviations from the nominal batch behavior are used to detect abnormal quality without making actual quality predictions [1–6]. Data-driven inferential sensors have also been used to predict the final quality of a batch. Successful statistical inference of the final quality of batch processes has been reported both in simulation studies [7–11] and in actual industrial installations [12, 13].

Most research effort has been directed towards process monitoring: the current state of the batch is assessed and an alarm is raised when an abnormality is detected. In contrast, data-driven models are seldom used for batch control or optimization due to their computational complexity. In the few reported applications, disturbance rejection by close tracking of set point profiles is the main goal [14–16]. This results in smaller deviations from the nominal trajectories and, in turn, smaller variations of the batch-end quality.

Recently, McCready [17] optimized the yield of a batch process by adapting the manipulated variables (MVs) at three distinct decision moments and predicting the resulting batch-end quality with a Partial Least Squares (PLS [18]) statistical inference model. For full (online) optimization and disturbance rejection, however, more decision moments are needed. In an ideal case, changes to the MVs are made every few time points: frequently enough to tightly control the final quality and negate the effect of process disturbances, but without upsetting the batch with too frequent adjustments.

In the work of McCready, full factorial test runs were conducted for constructing the statistical inference model. This is feasible for the case with a single MV and three decision moments. For more frequent changes in the MV and/or more than one MV, this requires a very large number of test runs to identify a suitable process model. The time required for these tests is a major drawback because fast implementation with as few tests as possible is one of the major criteria for industrial acceptance.

In this paper, it is investigated whether accurate statistical inference models for quality prediction can be identified from a training set built on a limited number of MV changes. If so, not all possible decision moments and/or possible combinations of MV changes need to be tested to identify a suitable model for batch-end quality prediction, batch control, and quality optimization. This would significantly reduce the required number of test runs, tackling one of the major hurdles for industrial adoption of data-driven control strategies.

The structure of this paper is as follows. In Section 2, Multiway Partial Least Squares is described. Section 3 details the methodology for determining the system response to future set point changes. Section 4 discusses the expansion from the work of McCready to full online batch control and optimization. The case study used in this work is presented in Section 5. Results are presented and discussed in Section 6, and final conclusions are drawn in Section 7.

2 Multiway Partial Least Squares

In this paper, Multiway Partial Least Squares (MPLS [8]) is used to infer the final batch quality from online measurements. MPLS deals with the three-dimensional data structure specific to batch processes via unfolding. Next, a Partial Least Squares (PLS [18]) model links the unfolded data to the quality measurements.

2.1 Data Unfolding

A historic data set of I batches, where J sensors are sampled at K time points can be arranged in an $I \times J \times K$ three-dimensional data tensor $\underline{\mathbf{X}}$. An $I \times L$ quality matrix $\mathbf{Y}$ contains the measurements of the L quality parameters. However, PLS requires a two-dimensional data matrix. Therefore, the measurement tensor is first arranged in a two-dimensional matrix by unfolding.

When dealing with quality prediction, the entire batch history determines the final quality. Therefore, batch-wise unfolding of the data tensor $\underline{\mathbf{X}}$ is employed [7], resulting in a $I \times JK$ data matrix $\mathbf{X}$. Each row of $\mathbf{X}$ represents a single batch (observation), linked with a single batch-end quality measurement. Hence, the influence of the entire batch history on the final quality is captured.

2.2 Partial Least Squares

First, the main nonlinear and dynamic components are removed from the data by mean-centering and scaling of batch-wise unfolded data [7]. This effectively linearizes the measurements around their average trajectory, improving modelling results because PLS is multi-linear [7, 8].

Next, a standard PLS model identifies the relation between online measurements and final quality [18]. PLS is a multi-linear latent variable modelling approach that decomposes the input matrix $\mathbf{X}$ and output (quality) matrix $\mathbf{Y}$ into R uncorrelated latent variables (components).

$$\begin{cases} \mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}_\mathbf{X} \\ \mathbf{Y} = \mathbf{T}\mathbf{Q}^T + \mathbf{E}_\mathbf{Y} \end{cases} \tag{1}$$

The scores matrix $\mathbf{T}$ $(I \times R)$ contains the low-dimensional approximation of the measurements $\mathbf{X}$. The loading matrices $\mathbf{P}$ $(JK \times R)$ and $\mathbf{Q}$ $(L \times R)$ are the model weights in in- and output space, respectively. The matrices $\mathbf{E}_\mathbf{X}$ $(I \times JK)$ and $\mathbf{E}_\mathbf{Y}$ $(I \times L)$ are the residuals.

Because $\mathbf{P}$ is not invertible, a weight matrix $\mathbf{W}$ $(JK \times R)$ is used to obtain the regression matrix $\mathbf{B}$ $(JK \times R)$ for estimating the scores $\mathbf{T}$ for a given $\mathbf{X}$. $\mathbf{W}$ has orthonormal columns and is determined so that $\left(\mathbf{P}^T\mathbf{W}\right)^{-1}$ is upper triangular with ones as diagonal elements.

$$\hat{\mathbf{T}} = \mathbf{X}\mathbf{W}\left(\mathbf{P}^T\mathbf{W}\right)^{-1} \triangleq \mathbf{X}\mathbf{B} \tag{2}$$

In this work, the number of latent variables R is identified using the adjusted Wold criterion [24]. It is chosen as the smallest value r for which the following equation holds.

$$\frac{\mathrm{MSE}_{\mathrm{loo}}(r+1)}{\mathrm{MSE}_{\mathrm{loo}}(r)} > 0.9 \tag{3}$$

$\mathrm{MSE}_{\mathrm{loo}}(r)$ is the mean squared error for the MPLS model with r components obtained via leave-one-out crossvalidation on the training batches.

3 Trimmed Scores Regression

At given time k, the input matrix $\mathbf{X}_{\mathrm{new}}$ $(1 \times JK)$ of a running new batch is only partially known because the future measurements (i.e., the measurements at times $k+1$, $k+2$, ..., K) are not yet available. The PLS model requires profiles from a completed batch and can, therefore, not be used for the prediction of the final quality of $\mathbf{X}_{\mathrm{new}}$. Trimmed Scores Regression (TSR) alleviates this problem [19, 9]. It compensates for missing measurements without directly estimating the future evolution of the different measurement profiles.

The data matrix $\mathbf{X}_{\mathrm{new}}$ and regression matrix $\mathbf{B}$ are divided in two parts. The first part $\mathbf{X}_{\mathrm{new},k}$ $(1 \times Jk^{+})$ and $\mathbf{B}_k$ $(Jk^{+} \times R)$ correspond with the known measurements while the second part $\mathbf{X}_{\mathrm{new},u}$ $(1 \times J(K-k)^{-})$ and $\mathbf{B}_k$ $(J(K-k)^{-} \times R)$ contains the unknown future measurements. For batch control, the future profiles of the manipulated variables can be imposed, and are (approximately) known. Hence, the known measurements include (i) the past k values of all J measurements, and (ii) the future $K-k$ values of the manipulated variables. The unknown measurements u are the future values of the non-manipulated variables. Hence, at least Jk columns of $\mathbf{X}_{\mathrm{new}}$ are known, and at most $J(K-k)$ columns are unknown. This partitioning of $\mathbf{X}_{\mathrm{new}}$ and $\mathbf{B}$ is substituted in Equation (2) for estimating the final scores of the new batch.

$$\begin{aligned} \mathbf{T}_{\mathrm{new}} &= \mathbf{X}_{\mathrm{new}}\mathbf{B} \\ &= [\mathbf{X}_{\mathrm{new},k}\ \mathbf{X}_{\mathrm{new},u}] \begin{bmatrix} \mathbf{B}_k \\ \mathbf{B}_u \end{bmatrix} \\ &= \mathbf{X}_{\mathrm{new},k}\mathbf{B}_k + \mathbf{X}_{\mathrm{new},u}\mathbf{B}_u \\ &\triangleq \mathbf{T}^{*}_{\mathrm{new},k} + \mathbf{T}^{*}_{\mathrm{new},u} \end{aligned} \tag{4}$$

The scores of the new batch $\mathbf{T}_{\mathrm{new},k}$ $(1 \times R)$ consist of an estimation based on the known samples, $\mathbf{T}^{*}_{\mathrm{new},k}$ $(1 \times R)$, and a correction of this first estimation due to future measurements, $\mathbf{T}^{*}_{\mathrm{new},u}$ $(1 \times R)$.

Because $\mathbf{T}^{*}_{\mathrm{new},u}$ is unknown, Equation (4) cannot predict the scores of a running batch. The solution to this problem is found in the training batches $\mathbf{X}_{\mathrm{train}}$ $(I \times JK)$, for which the final scores $\mathbf{T}_{\mathrm{train}}$ and trimmed scores $\mathbf{T}^{*}_{\mathrm{train},k}$ are known.

$$\begin{aligned} \mathbf{T}_{\mathrm{train}} &= \mathbf{X}_{\mathrm{train}}\mathbf{B} \\ \mathbf{T}^{*}_{\mathrm{train},k} &= \mathbf{X}_{\mathrm{train},k}\mathbf{B}_k \end{aligned} \tag{5}$$

A simple least squares model between the final scores $\mathbf{T}_{\text{train}}$ and the trimmed scores $\mathbf{T}^*_{\text{train},k}$ is identified at each time point k. This time-varying model is used to estimate the final scores of the new batch at time k.

$$\hat{\mathbf{T}}_{\text{new},k} = \mathbf{T}^*_{\text{new},k} \left(\mathbf{T}^{*T}_{\text{train},k} \mathbf{T}^*_{\text{train},k}\right)^{-1} \mathbf{T}^*_{\text{train},k} \mathbf{T}_{\text{train}} \tag{6}$$

Exploiting Equations (1) and (2) yields the final equations for estimating the final quality variables of the new batch online.

$$\begin{aligned} \hat{\mathbf{T}}^{\text{TSR}}_{\text{new},k} = \; & \mathbf{X}_{\text{new},k} \mathbf{B}_k \left(\mathbf{B}_k^T \mathbf{X}^T_{\text{train},k} \mathbf{X}_{\text{train},k} \mathbf{B}_k\right)^{-1} \cdot \ldots \\ & \ldots \cdot \mathbf{B}_k^T \mathbf{X}^T_{\text{train},k} \mathbf{X}_{\text{train}} \mathbf{B} \\ \hat{\mathbf{Y}}^{\text{TSR}}_{\text{new},k} = \; & \hat{\mathbf{T}}^{\text{TSR}}_{\text{new},k} \mathbf{Q}^T \end{aligned} \tag{7}$$

The advantage of TSR is that a single PLS model can be used to provide batch-end quality estimates at all times of the operation. Hence, once the full PLS model is identified, online estimates are available without much extra effort.

Other options for online batch-end quality estimation include the use of *evolving models* [20–22]. However, it has been shown that TSR provides accurate estimates of the batch-end quality even when few samples are available despite its simple linear nature [19, 23].

4 Batch Control and Optimization

McCready optimized the yield of a batch process with an MPLS model by adapting manipulated variables (MVs) at three distinct decision moments [17]. For model identification, McCready used factorial experiments on the MVs. Executing these tests is feasible for a single MV and three decision moments but full online control and optimization of a batch process requires more frequent control actions to reject process disturbances and/or optimize batch-end quality. Hence, the need for factorial experiments quickly results in an unfeasibly large amount of tests for the identification of the batch-end quality MPLS model.

For continuous processes, this problem is traditionally solved by applying Pseudo-Random Binary Signal (PRBS) changes to the different MVs. These PRBS tests can also be used for profile tracking in batch processes, where the aim is to identify the influence of the MVs on other online process variables [14, 15]. In these two cases, the effect of each set of MV changes on the other online measurements is observed quasi immediately.

Batch-end quality prediction faces a few extra problems. First of all, the influence of all changes in the MVs is combined in a single quality measurement only available after batch completion. To obtain clear results, only a single combination of MV changes should be tested in each batch. This is a big difference with continuous processes, where different combinations of MV changes can be tested in rapid succession. Hence, the required tests for model identification are very time-consuming. Furthermore, the dynamics of a batch process and the (magnitude of the) influence of the MVs on the final quality change over time.

This implies that the exact decision time of each MV change is another degree of freedom. Combinations of not only different MVs but also different decision times must be tested, further increasing the number of test runs for model identification. For example, the effect of a +5% change in feed rate should be tested after 10h, 20h, 30h,... of operation to correctly capture the process dynamics. Finally, identification tests might negatively impact the final quality of the batch. For profile tracking, only a few test batches might be wasted; for quality control, this number will be much higher.

In summary, batch control and optimization requires many more tests to identify the correct MPLS model between MVs and online measurements on the one hand, and final batch quality on the other hand. This presents a major drawback for practical implementation.

This paper first investigates whether a set tests with MV changes at times $\{k_1, k_2, \ldots, k_n\}$ can be used to construct a batch-end quality MPLS model for batches where the MV change occurs at a time not included in the training set. Next, it is studied whether the training batches must include multiple changes of each MV, or if batches with a single change in MVs suffice for model training.

5 Case Studies

This work employs the same benchmark batch process as McCready [17] for two case studies. A brief description of the batch process is provided in Section 5.1. Sections 5.2 and 5.3 describe the two case studies in more detail.

5.1 Benchmark Process: `Pensim`

The `Pensim` simulator of Birol *et al.* [10] is used in this paper. It simulates an industrial-scale batch fermentative penicillin production, and is a widely-used benchmark for (bio)chemical batch processes.

Initially, the fermentation is operated in batch mode at high initial substrate concentration to stimulate biomass growth. After approximately 43 hours, the substrate concentration in the reactor drops below 0.3 g/L and penicillin production starts. In this phase, the operation is switched to fed-batch mode and substrate is fed to the reactor. The fermentation is complete once 25 L of substrate have been added. The entire process lasts approximately 460 hours. For more details, the reader is referred to Birol *et al.* [10].

In the case studies, the substrate feed rate is the manipulated variable. The quality variable is the penicillin concentration at batch completion.

A total of 11 online sensors record the variables presented in Table 1 during the fermentation. Two phases are identified: (*i*) the batch phase and (*ii*) the fed-batch phase. Due to differences in initial conditions and non-perfect controllers, not all batches evolve identically. Therefore, synchronization is required to equalize the length of each phase. The profiles are synchronized and resampled using the indicator variables proposed by Birol *et al.*: culture volume is used for the batch phase and substrate fed for the fed-batch phase [10]. To retain temporal

Table 1. Measurements of `Pensim`

Online measurement	σ_{noise}	**Online measurement**	σ_{noise}
Time [h]	—	Dissolved oxygen [mmol/L]	$6.667 \cdot 10^{-3}$
Reactor volume [L]	$5.556 \cdot 10^{-2}$	Aeration rate [L/h]	$1.389 \cdot 10^{-2}$
Feed rate [L/h]	$1.389 \cdot 10^{-4}$	Agitator power [W]	$2.778 \cdot 10^{-2}$
Feed temperature [K]	$2.778 \cdot 10^{-2}$	pH [-]	$2.778 \cdot 10^{-3}$
Reactor temperature [K]	$3.333 \cdot 10^{-1}$	Base flow [L/h]	$5.556 \cdot 10^{-7}$
Cooling water flow [L/h]	$1.389 \cdot 10^{-1}$	Acid flow [L/h]	$5.556 \cdot 10^{-8}$
Offline measurement	σ_{noise}		
Penicillin concentration [g/L]	$1.860 \cdot 10^{-2}$		

information after synchronization, time is added as a 12th measurement. After synchronization, the batch phase has a length of 101 samples while the fed-batch phase is 502 samples long.

Measurement noise on the online measurements and the penicillin concentration is simulated by Gaussian noise with zero mean and standard deviation as listed in Table 1. The measurement error on the penicillin concentration corresponds to an error of approximately 1%.

5.2 Study 1: Single Input Change

In a first case study, it is investigated how well the final quality of a batch with one change in feed rate can be predicted from batches where the change in feed rate occurs at another moment. If not all MV switch times must be included in the training set, the number of test runs for identification can be reduced.

Batches with a single change in substrate feed rate are simulated. All batches start at the nominal feed rate of 0.06 g/L at the start of the fed-batch phase. The feed rate is adjusted upwards or downwards 5% to 0.057 g/L or 0.063 g/L after 150h, 175h, 200h, 225h, 250h, 275h, 300h, 325h, 350h, 375h or 400h. The simulation is run 20 times for each step direction and change time for a total of 440 batches with a single feed rate adjustment. In addition, 20 batches without feed rate changes are simulated.

In case 1A, the model is trained on the 20 batches without feed rate changes and the 40 batches with changes in feed rate occurring after 150h of operation. To take the time-varying process dynamics into account, the training set is expanded in cases 1B (150h, 200h), 1C (150h, 250h), 1D (150h, 300h), 1E (150h, 350h) and 1F (150h, 400h). Cases 1D* (150h, 225h, 300h) and 1E* (150h, 250h, 350h) include a third set of MV changes halfway the time range of cases 1D and 1E.

5.3 Study 2: Multiple Input Changes

The second case study investigates how well the batch-end quality of batches with multiple changes in feed rate is predicted from batches with a single change in feed rate occurring at different decision times. Using batches with only a single

Table 2. Overview of available batches for study 2

	Training	Validation
Case 2A	nominal	$\pm5\%$ @ 150h, $\mp5\%$ @ 250h
	$\pm5\%$ @ 150h	$\pm5\%$ @ 150h, $\pm0\%$ @ 250h
	$\pm5\%$ @ 250h	
Case 2B	nominal	$\pm5\%$ @ 175h, $\mp5\%$ @ 275h
	$\pm5\%$ @ 175h	$\pm5\%$ @ 175h, $\pm0\%$ @ 275h
	$\pm5\%$ @ 275h	
Case 2C	nominal	$\pm5\%$ @ 200h, $\mp5\%$ @ 300h
	$\pm5\%$ @ 200h	$\pm5\%$ @ 200h, $\pm0\%$ @ 300h
	$\pm5\%$ @ 300h	
Case 2D	nominal	$\pm5\%$ @ 175h, $\mp5\%$ @ 200h
	$\pm5\%$ @ 150h	$\pm5\%$ @ 175h, $\mp5\%$ @ 250h
	$\pm5\%$ @ 300h	$\pm5\%$ @ 175h, $\mp5\%$ @ 275h
		$\pm5\%$ @ 200h, $\mp5\%$ @ 250h
		$\pm5\%$ @ 200h, $\mp5\%$ @ 275h
		$\pm5\%$ @ 250h, $\mp5\%$ @ 275h
Case 2E	nominal	$\pm5\%$ @ 175h, $\pm0\%$ @ 200h
	$\pm5\%$ @ 150h	$\pm5\%$ @ 175h, $\pm0\%$ @ 250h
	$\pm5\%$ @ 225h	$\pm5\%$ @ 175h, $\pm0\%$ @ 275h
	$\pm5\%$ @ 300h	$\pm5\%$ @ 200h, $\pm0\%$ @ 250h
		$\pm5\%$ @ 200h, $\pm0\%$ @ 275h
		$\pm5\%$ @ 250h, $\pm0\%$ @ 275h

change in feed rate instead of testing (almost) all possible combinations of MV changes greatly reduces the number of test runs needed for model identification.

Different combinations of training and validation patterns are investigated, as listed in Table 2. The model for batch-end quality prediction is trained on batches with a single change in feed rate each but occurring at different times for different batches. The validation batches exhibit multiple changes in feed rate at the moments also used for model training (cases 2A-2C) or at times inside the time range used for training (cases 2D and 2E). All validation patterns have a 5% change up or down in the feed rate at the first switch time. The first type of pattern ($\pm5\% \mp 5\%$) switches from $+5\%$ (above nominal) to -5% (below nominal) or from -5% to $+5\%$ at the second switch time (i.e., a total change of 10% in feed rate at the second switch time). The second type of validation pattern ($\pm5\% \pm 0\%$) returns to the nominal feed rate (i.e., a 5% change in feed rate). Each feed rate change pattern is simulated 20 times, resulting in 100 training and 80 validation batches in cases 2A-2C. Case 2D and 2E have 100 and 140 training batches, respectively, and 480 validation batches.

6 Results and Discussion

A model is suited for online batch optimization and control if it *accurately predicts* the final batch quality *at an early stage* of the batch. Therefore, the model

Table 3. Final prediction accuracy for batches with a single change in the feed rate. Crossvalidation results are listed between parentheses for the training batches.

	MSE·10^4							
	1A	**1B**	**1C**	**1D**	**1E**	**1F**	**1D***	**1E***
Change time	($R = 4$)	($R = 4$)	($R = 3$)	($R = 3$)	($R = 3$)	($R = 4$)	($R = 3$)	($R = 3$)
nominal	(5.99)	(5.02)	(4.65)	(5.38)	(9.40)	(22.27)	(5.16)	(6.99)
±5% @ 150h	(2.99)	(2.58)	(3.05)	(3.31)	(8.27)	(22.99)	(3.32)	(6.07)
±5% @ 175h	6.93	4.87	4.79	5.72	9.26	31.73	4.47	6.31
±5% @ 200h	5.77	(3.82)	3.55	5.19	10.89	32.84	4.32	6.97
±5% @ 225h	6.10	3.65	3.23	3.99	7.94	25.77	(3.37)	3.96
±5% @ 250h	10.35	5.32	(3.67)	5.53	10.47	32.90	4.35	(5.07)
±5% @ 275h	8.13	6.20	5.31	4.45	7.75	23.76	4.38	5.84
±5% @ 300h	10.75	7.68	6.22	(4.97)	6.20	20.49	(4.93)	3.47
±5% @ 325h	13.92	14.66	14.40	10.08	5.70	12.31	10.93	7.21
±5% @ 350h	31.74	31.77	29.85	24.14	(13.73)	17.90	24.36	(16.67)
±5% @ 375h	69.87	74.39	76.17	67.70	47.24	31.70	69.30	54.66
±5% @ 400h	108.98	116.08	119.92	107.35	85.80	(46.28)	109.49	95.91
Training	(3.99)	(3.56)	(3.62)	(4.39)	(10.68)	(32.16)	(4.06)	(8.94)
Interpolation	—	4.87	3.86	4.98	8.32	25.49	4.38	5.63

performance is evaluated via two indices in each case study. The first measure is the accuracy of the offline prediction of the batch-end quality, i.e., the model quality prediction after batch completion. If the model is unable to correctly estimate the final quality based on measurements of the entire batch, it cannot be used for online control and optimization. The second performance indicator is the convergence of the online estimate to the batch-end quality measurement. A model that only provides accurate estimates very late in the batch is again not useful for online control.

6.1 Study 1: Single Input Change

Final Prediction Accuracy. Table 3 lists the crossvalidation MSE for the training batches and validation MSE for the validation batches for different combinations of MV change times. In case 1A, the average crossvalidation MSE for the training batches is $3.99 \cdot 10^{-4}$. The validation batches exhibit higher MSE values. As the change in feed rate occurs later in the batch, the prediction accuracy degrades. For the batches with change times between 175h and 325h, the MSE gradually worsens. Once the MV change time reaches 350h, a very fast increase in MSE is observed.

Figure 1 compares the predicted and measured penicillin concentrations for the batches with a feed rate change after 400h; the crossvalidation results for the training batches are added for reference. From this plot, it is clear there is a systematic model mismatch: the process dynamics identified after 150h are not valid at the later stages of the batch. This is corroborated in Figure 2, where

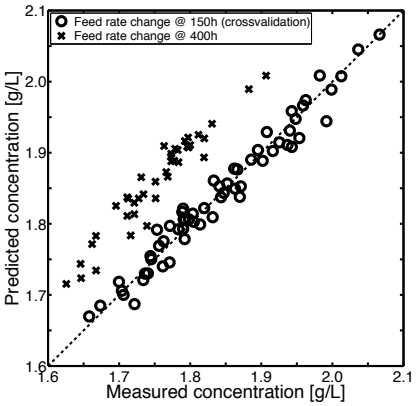


Fig. 1. Predicted vs. measured final penicillin concentration for training batches and batches with change in feed rate after 400h for case 1C

the average evolution of the penicillin concentration over time is depicted. (This is available from the simulator, but not measurable.)

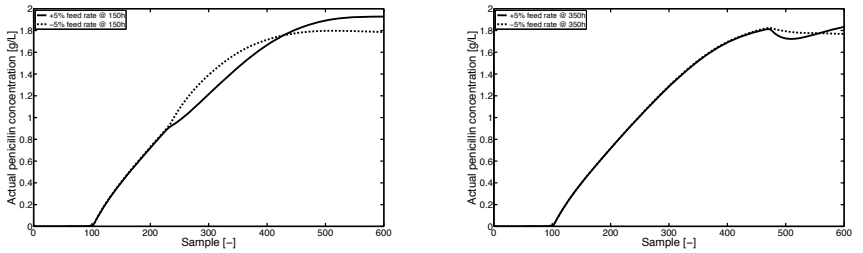
As can be seen in Figure 2(a), an increase in feed rate initially slows down the production of penicillin because the extra available substrate inhibits production and stimulates biomass growth. Once the substrate concentration is again low enough, penicillin production resumes at an increased rate due to the higher biomass concentration. By contrast, a feed rate decrease after 150h initially speeds up penicillin production owing to the lower inhibition. The lower substrate concentration also slows down biomass growth, eventually offsetting the initial production increase. A decrease in feed rate increases production if the batch were terminated between samples 225 and 425, while an increase in feed rate is advantageous for batch durations longer than 425 samples.

This behavior is not observed for the batches with feed rate changes after 350h in Figure 2(b): both an in- and decrease of the feed rate initially lead to a lower penicillin concentration, clearly illustrating the change in process dynamics.

For cases 1B-1F, the average MSE for the validation batches where the change in feed rate occurs inside the time range used for training is indicated in the last row of Table 3 ("Interpolation"). The validation performance of the time-interpolated batches is in line with the crossvalidation results for the training batches. For the batches located outside of the training range, the prediction performance gradually degrades as the change time of the feed rate is located farther from the training time range.

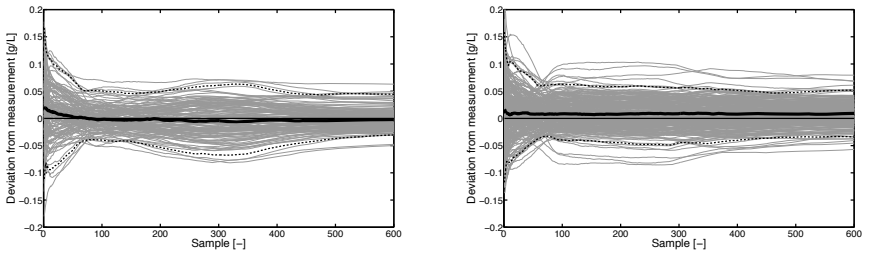
The change in process dynamics for the batches where the feed rate is changed after 350h or later is also seen in the higher crossvalidation MSE for cases 1E and 1F compared to cases 1A-1D.

The dynamics of the batch process are better captured by including the intermediate MV changes in the training set, as evidenced by the slighty lower MSEs for cases 1D* and 1E* when compared to cases 1D and 1E.



(a) (b)

Fig. 2. Evolution of the average penicillin concentration over time, for feed rate changes after (a) 150h and (b) 350h



(a) (b)

Fig. 3. Deviation of the online prediction from the final quality for (a) the time-interpolating validation batches in case 1C, and (b) the $\pm 5\% \pm 0\%$ batches in case 2D

These results imply that batch-end quality for the time-interpolating batches in this study can be predicted accurately from the training batches. Time extrapolation of the MV change is detrimental for prediction performance. It is concluded that the amount of test batches for model identification can be reduced because not all MV switch times must be included in the training set.

Online Convergence. In each of the cases in this study, the final quality is estimated online using TSR (Section 3) for all validation batches to assess the quality of the online estimate. Figure 3(a) depicts the deviation of the online batch-end quality prediction from the final quality measurement for the time-interpolating validation batches in case 1C (i.e., with feed rate changes after 175h, 200h, or 225h). The solid black line indicates the median observed deviation, the dashed black lines indicate the empiric 95% prediction interval. Similar plots are obtained for the time-interpolating batches in the other cases.

From this plot, it is concluded that the online estimate quickly converges: after less than 100 samples (i.e., before the start of the fed-batch phase), the deviation of the estimate from the final lab measurement is typically not larger

Table 4. Final prediction accuracy for batches with multiple changes in the feed rate for. Crossvalidation results are listed between parentheses for the training batches.

	MSE·10^4		
	2A	**2B**	**2C**
Pattern	($R = 3$)	($R = 3$)	($R = 3$)
nominal	(4.65)	(5.01)	(5.38)
±5% @ t_1	(3.05)	(4.34)	(4.72)
±5% @ t_2	(3.67)	(5.38)	(5.13)
±5% @ t_1, ∓5% @ t_2	6.98	43.12	88.76
±5% @ t_1, ±0% @ t_2	6.04	4.05	5.41

	2D	**2E**
Pattern	($R = 3$)	($R = 3$)
nominal	(4.16)	(5.16)
±5% @ 150h	(3.32)	(3.32)
±5% @ 225h	—	(3.37)
±5% @ 300h	(4.93)	(4.93)
±5% @ 175h, ∓5% @ 200h	5.79	5.08
±5% @ 175h, ∓5% @ 250h	5.58	5.38
±5% @ 175h, ∓5% @ 275h	24.84	27.51
±5% @ 200h, ∓5% @ 250h	4.67	4.45
±5% @ 200h, ∓5% @ 275h	14.27	16.12
±5% @ 250h, ∓5% @ 275h	3.31	3.40
±5% @ 175h, ±0% @ 200h	3.78	3.45
±5% @ 175h, ±0% @ 250h	6.00	5.71
±5% @ 175h, ±0% @ 275h	3.62	3.48
±5% @ 200h, ±0% @ 250h	4.26	3.83
±5% @ 200h, ±0% @ 275h	8.40	7.63
±5% @ 250h, ±0% @ 275h	4.91	4.55

than 0.05 g/L. This prediction accuracy is comparable to the accuracy of the offline quality estimates. Furthermore, the deviation of any single batch only changes slowly over time, resulting in a stable prediction. Hence, the batch-end quality prediction is suited for online batch control and optimization.

Summary. MPLS models can be used to predict the final quality of time-interpolating batches with the same accuracy as for the training batches. The online estimate of the final quality for these batches quickly converges towards its final value. It is, therefore, concluded that not all possible decision times for the MV must be included in the training set. This substantially reduces the amount of test runs and represents a major advantage for practical implementation.

6.2 Study 2: Multiple Input Changes

Final Prediction Accuracy. The (cross)validation MSE values for cases 2A-2E are listed in Table 4. In most cases, the validation performance for the batches

with a 10% change in feed rate at the second switch time ($\pm 5\% \mp 5\%$) is worse than that for the batches returning to the nominal feed rate ($\pm 5\% \pm 0\%$). The MSE for the $\pm 5\% \pm 0\%$ batches is in line with the training results, while the MSE for the $\pm 5\% \mp 5\%$ batches is an order of magnitude higher in cases 2B and 2C. These results are not unexpected because the training batches do not exhibit 10% changes in feed rate.

It is concluded that an MPLS model trained on batches with a single change in feed rate can accurately predict the final quality of batches with multiple changes in feed rate, provided the changes in feed rate are not larger than those used for model identification.

Online Convergence. Figure 3(b) displays the median deviation (solid line) and 95% empiric prediction interval (dashed lines) of the online estimate of the batch-end quality for the $\pm 5\% \pm 0\%$ batches in case 2D, computed with TSR. The $\pm 5\% \pm 0\%$ batches in the other cases yield similar plots.

As in the first study (Section 6.1), the online quality estimate converges towards its final value in less than 100 samples. The online prediction accuracy is in line with the offline accuracy. In cases 2D and 2E, a small bias of the estimates exists. Nevertheless, the 95% prediction interval is approximately symmetric around zero. No bias is observed in cases 2A-2C (results not shown).

It is concluded that the MPLS model identified on batches with a single change in feed rate is suited for online batch control and optimization of batches with multiple changes in feed rate.

Summary. In this case study, it was concluded that an MPLS model identified on batches with a single change in feed rate can accurately predict the batch-end quality of batches with multiple time-interpolating changes in feed rate both off- and online. However, the magnitude of the feed rate changes in the new batches should not be bigger than the magnitude of the feed rate jumps used for model identification. Hence, not all possible combinations of MV changes must be tested before training a suitable model for batch control and optimization.

7 Conclusions

Recently, a procedure for optimization of the final quality of a batch process was proposed by McCready [17]. The methodology changes the manipulated variables at three decision times during the batch. True online batch control and optimization, however, requires control actions every few time points. While forming an interesting starting point, McCready's methodology is not directly applicable for online optimization because it involves a full factorial design for model identification. This very large number of tests poses a major problem for practical implementation and industrial acceptance.

This paper investigated whether all possible decision times and/or combinations of changes of the manipulated variables must be present in the training

data. Accurate batch-end quality predictions for batches not present in the training tests greatly reduce the number of training tests by avoiding factorial design for identification.

Two case studies were performed. The first study concluded that the final quality of batches with manipulated variable switch times not present in the training data could be accurately predicted if the switch time for the new batch was located inside the time range used for training. The second case study showed that predictive models trained on batches with a single change in feed rate can also accurately predict the final quality of batches with multiple MV changes.

It is concluded that (*i*) not all possible control times and (*ii*) not all possible combinations of changes in the manipulated variables must be included in the training set. Instead, a limited set of test batches suffices for model training. This is a valuable result for practical (industrial) implementation and acceptance because the amount of test batches needed is reduced: a (full) factorial design of training batches is not required.

Future work consists of (*i*) the extension of this study towards other magnitudes of the MV changes, (*ii*) the derivation of mathematical criteria for determining the optimal number of manipulated variable switch times in the training set, and (*iii*) the validation of the results on more complicated MV patterns and other processes.

Acknowledgements. Work supported in part by Project PFV/10/002 (OPTEC Optimization in Engineering Center) of the Research Council of the KU Leuven, Project KP/09/005 (SCORES4CHEM) of the Industrial Research Council of the KU Leuven, and the Belgian Program on Interuniversity Poles of Attraction initiated by the Belgian Federal Science Policy Office. J. Vanlaer and P. Van den Kerkhof have a Ph.D grant of the Agency for Innovation by Science and Technology (IWT). J. Van Impe holds the chair Safety Engineering sponsored by the Belgian chemistry and life sciences federation essenscia. The authors assume scientific responsibility.

References

1. Nomikos, P., MacGregor, J.F.: Monitoring batch processes using multiway principal component analysis. AIChE J. 40(8), 1361–1375 (1994)
2. Chen, J., Liu, K.-C.: On-line batch process monitoring using dynamic PCA and dynamic PLS models. Chem. Eng. Sci. 57, 63–75 (2002)
3. Choi, S.W., Martin, E.B., Morris, A.J., Lee, I.-B.: Fault detection based on a maximum likelihood PCA mixture. Ind. Eng. Chem. Res. 55(7), 2316–2327 (2005)
4. Choi, S.W., Morris, A.J., Lee, I.-B.: Dynamic model-based batch process monitoring. Chem. Eng. Sci. 63(3), 622–636 (2008)
5. Simoglou, A., Georgieva, P., Martin, E.B., Morris, A.J., de Azevedo, S.F.: On-line monitoring of a sugar crystallization process. Comput. Chem. Eng. 29(6), 1411–1422 (2005)
6. Hu, K., Yuan, J.: Multivariate statistical process control based on multiway locality preserving projections. J. Process Control 18, 797–807 (2008)

7. Nomikos, P., MacGregor, J.F.: Multivariate SPC charts for monitoring batch processes. Technometrics 37(1), 41–59 (1995a)
8. Nomikos, P., MacGregor, J.F.: Multiway partial least squares in monitoring batch processes. Chemom. Intell. Lab. Syst. 30, 97–108 (1995b)
9. García-Muñoz, S., Kourti, T., MacGregor, J.F.: Model predictive monitoring for batch processes. Ind. Eng. Chem. Res. 43, 5929–5941 (2004)
10. Birol, G., Ündey, C., Çinar, A.: A modular simulation package for fed-batch fermentation: penicillin production. Comput. Chem. Eng. 26, 1553–1565 (2002)
11. Ündey, C., Ertunç, S., Çinar, A.: Online batch/fed-batch process performance monitoring, quality prediction, and variable-contribution analysis for diagnosis. Ind. Eng. Chem. Res. 42, 4645–4658 (2003)
12. Gins, G., Pluymers, B., Smets, I.Y., Espinosa, J., Van Impe, J.F.M.: Prediction of batch-end quality for an industrial polymerization process. In: Perner, P. (ed.) ICDM 2011. LNCS (LNAI), vol. 6870, pp. 314–328. Springer, Heidelberg (2011)
13. Gins, G., Van den Kerkhof, P., Van Impe, J.F.M.: Hybrid derivative dynamic time warping for online industrial batch-end quality estimation. Ind. Eng. Chem. Red. 51(17), 6071–6084 (2012)
14. Flores-Cerrillo, J., MacGregor, J.F.: Latent variable MPC for trajectory tracking in batch processes. J. Process Contr. 15, 651–663 (2005)
15. Golshan, M., MacGregor, J.F., Bruwer, M.-J., Mhaskar, P.: Latent Variable Model Predictive Control (LV-MPC) for trajectory tracking in batch processes. J. Process Control 20, 538–550 (2010)
16. Wan, J., Marjanovic, O., Lennox, B.: Disturbance rejection for the control of batch end-product quality using latent variable models. J. Process Control 22, 643–652 (2012)
17. McCready, C.: Model Predictive Multivariate Control (MPMC). In: 2nd European Conference on Process Analytics and Control Technology (EuroPACT 2011), Glasgow (United Kingdom), p. 82 (2011)
18. Geladi, P., Kowalski, B.R.: Partial least-squares regression: A tutorial. Anal. Chim. Acta 185, 1–17 (1986)
19. Arteaga, F., Ferrer, A.: Dealing with missing data in MSPC: several methods, different interpretations, some examples. J. Chemom. 16, 408–418 (2002)
20. Ramaker, H., Sprang, E.N.M., Westerhuis, J.A., Smilde, A.K.: Fault detection properties of global, local and time evolving models for batch process monitoring. J. Process Control 15, 799–805 (2005)
21. Camacho, J., Picó, J., Ferrer, A.: Bilinear modelling of batch processes. Part I: Theoretical discussion. J. Chemometr. 22, 299–308 (2008)
22. Faggiana, A., Faccoa, P., Doplicherb, F., Bezzoa, F., Baroloa, M.: Multivariate statistical real-time monitoring of an industrial fed-batch process for the production of specialty chemicals. Chem. Eng. Res. Des. 87, 325–334 (2009)
23. Gins, G., Vanlaer, J., Van Impe, J.F.M.: Online batch-end quality estimation: does laziness pay off? In: Quevedo, J., Escobet, T., Puig, V. (eds.) Proceedings of the 7th IFAC International Symposium on Fault Detection, Supervision and Safety of Technical Processes (SafeProcess 2009), pp. 1246–1251 (2009)
24. Li, B., Morris, J., Martin, E.: Model selection for partial least squares regression. Chemometr. Intell. Lab. Syst. 64, 79–89 (2002)

Pattern-Based Solution Risk Model for Strategic IT Outsourcing*

Robert Gwadera

Distributed Information Systems Laboratory
EPFL, Lausanne, Switzerland

Abstract. We present a pattern-based solution risk model for assessing risk of incurring a cost-overrun in *Strategic IT Outsourcing* (SO) by the SO provider based on historical deals and their corresponding cost overruns. The approach is based on finding co-occurring patterns of solution elements and cost-overrun elements, i.e., elements that had to be implemented as not foreseen in the project planning phase. In order to find such co-occurring patterns we apply closed itemset-mining augmented with item cost information and build corresponding association rules with risk information. Such rules can be used by project managers of SO contracts to minimize the gap between the proposed and implemented solutions. In experiments, conducted on a sample of deals of a multi-national SO provider, we show the applicability of the framework for predicting significant cost-overruns. The introduced model is a general solution risk model for service delivery, whose task is to minimize the gap between proposed and implemented service elements by the provider based on historical deals.

1 Introduction

The *Strategic IT Outsourcing* also just called *Strategic Outsourcing* (SO) is broadly defined as a process undertaken by an organization (SO client) to contract-out or to sell the organizations IT assets, staff and/or activities to a third party supplier (SO provider) who in exchange provides and manages IT assets and services for monetary return over an agreed period of time [7]. In this contractual relationship the SO provider assumes responsibility of one or more IT functions [8]. An outsourcing arrangement can be either *tactical* or *strategic*. An outsourcing is tactical when it is focused on solving a particular problem. Strategic outsourcing, on the other hand, emphasizes an alignment with the SO client's long-term strategies.

The most common reasons for outsourcing IT are financial (reducing costs, obtaining immediate cash, replacing capital outlays with periodic payments), technical (improving the quality of IT, gaining access to new and/or proprietary technology), strategic (focusing on core activities, facilitating mergers and acquisitions, specialized firms can more easily attract highly skilled professionals that are in short supply) and political motives (dissatisfaction with internal IT department, regarding IT as support function, desire to follow trends). Therefore,

* The paper was supported by the EU project OpenIoT (ICT 287305).

P. Perner (Ed.): ICDM 2013, LNAI 7987, pp. 55–69, 2013.

firms usually outsource their IT for achieving a combination of these benefits [7,13,2]. Thus, for many organizations, the goal of SO is to obtain access to the best possible technology at the lowest possible cost.

We now introduce fundamental concepts related to SO. A *Project* is a temporary (short term) tactical endeavor undertaken to produce a unique objective (product or service), within a specified scope. A SO contract usually involves the following projects (phases):

1. *Engagement*, where the client is engaged in negotiations as part of which the client gives some limited access to the existing infrastructure to the provider. Based on pre-contract *Due Diligence* project (investigation of the client's business) the provider defines and creates the transition and transformation solution and the corresponding cost case. The engagement ends with *contract signing*.
2. *Transition*, where the provider is granted a full access to the existing client's infrastructure and based on the actual evidence defines a detailed set of solution elements to transform the client's IT infrastructure. Thus, the aim of the transition project is to take over the clients service at the *service commencement date*. Conducted activities include: transfer of in-scope staff, validation of the services baseline environment, knowledge transfer from the client, set-up of a management system, workplace logistics and implementation of any interim processes and tools necessary to enable the take-over. The transition consists of a standard set of services that should typically be completed within 3-4 months. During the transition period, the provider delivers the services on a *Business as Usual* (BAU) basis (the normal execution of standard functional operations within an organization).
3. *Transformation*, where the actual transfer of client's IT infrastructure to the provider takes place. Thus, the transformation, implements the plans, processes and tools necessary to transform the services to meet the steady state service level agreements and productivity rates as committed in the contract. A distinction is made between a standard and a special transformation that may include projects that were started and not-completed by the client before the contract signing.
4. *Steady State Delivery*, where after the transformation the service level reaches the contracted objectives and the role of the provider is to provide necessary maintenance to the transferred infrastructure.

A *Project Portfolio* is a business view of projects either internal or external, that share specific common characteristics and are viewed as a group for management purposes. Thus, portfolio is the set of hardware, software or services offerings solutions or engagements within a business area that addresses a market segment. *Project Portfolio Management* is an approach to managing a business that groups projects based on their alignment and contribution to specific business objectives. Project delivery organizations focus on achieving the objectives of their particular projects within a portfolio. The portfolio manager focuses on making decisions concerning the content and performance of the portfolio as the whole.

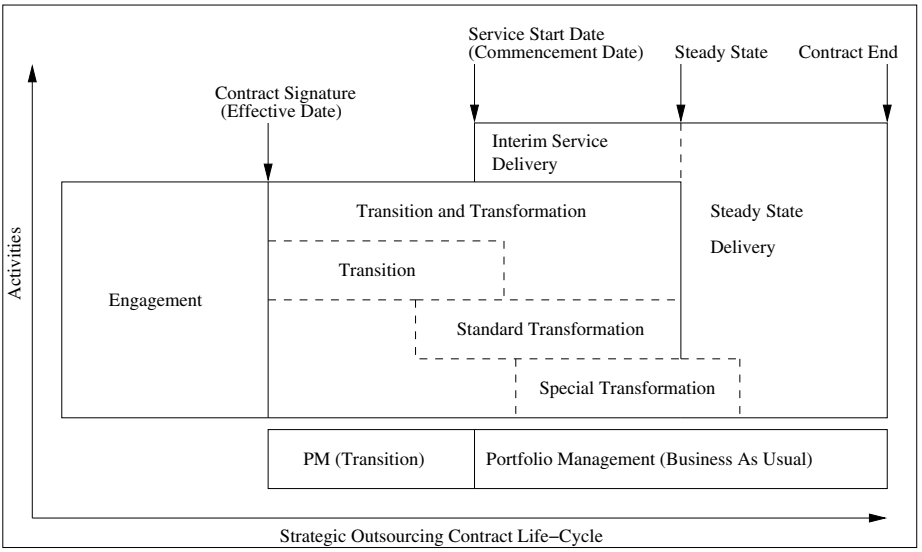


Fig. 1. Phases of the strategic IT outsourcing contract

Figure 1 shows a graphical representation of the SO contract life cycle. A standard solution consists of the following service elements: (I) hardware; (II) software; (III) customer service (e.g., a help desk) and (IV) labor (solution transformation and standardization). An important term is *Repeatable Model* that is a template from a repository related to labor (e.g., for a middleware installation).

While most of the previous research on risk modeling in SO focused on the client's side of the SO contract [14,9], we focus on the provider's side and analyze the problem of risk associated with cost-overruns, where in practice it turns out that implemented (delivered) solutions differ from proposed (contracted) solutions as a result of missing or under-costed solution elements in the contract. In particular, *the objective of our research is to design a system that minimizes the gap between proposed and implemented solutions based on risk estimation given lessons learned from previous deals.*

From the point of view of analyzing cost-overruns mainly two types of documents are of interest: (I) *Cost case*, created during the engagement phase, that provides a list of solution elements with cost information and (II) *Project change requests* (PCRs) created during the transition and transformation (T&T) phases, where a PCR is created whenever an update to the cost case in terms of a solution element is needed to carry out the project. Thus, cost-overruns are directly represented by PCRs and by comparing them with the cost case we can asses how much a proposed solution missed the actual contractual requirements. Thus, from the data mining point of view we compare solution elements with cost-overrun elements deal-wise by mining correlated itemsets and building predictive rules with risk information.

1.1 Overview of the Method

Table 1 presents a collection of deals with corresponding sets of solution elements and Table 2 presents the same collection of deals with corresponding sets of cost-overrun elements, where a 1/0 in a respective column means presence/absence of the corresponding attribute respectively. Let S_i denote the i-th solution element and R_i the corresponding cost-overrun element. Thus, S_i and R_i refer to the same i-th service element in the same underlying service alphabet (catalog) offered by the provider. For example, S_i corresponds to a server in the solution and to evidence a change in price at the time of the service delivery, R_i is created in the T&T phase. Thus, a presence of a cost-overrun element R_i (e.g., referring to a solution element "server") may indicate the following two failures in the contract depending on the presence of S_i: (I) if S_i is present in the cost case then it means that the server was under-costed and (II) if S_i is not present in the cost case then it means that the server was missing but turned out to be needed at the time of the implementation.

Table 1. A collection of proposed solution elements of the deals, where S_i is a solution element

Deal_Id	S_1	S_2	S_3	...	S_n
1	1	0	0	...	0
2	1	0	0	...	0
3	0	0	0	...	1
4	0	0	0	...	1
5	1	0	0	...	0

Table 2. A collection of cost-overrun elements, where element R_i corresponds to solution element S_i

Deal_Id	R_1	R_2	R_3	...	R_n
1	0	1	1	...	0
2	0	1	1	...	0
3	1	0	0	...	1
4	1	0	0	...	1
5	0	1	1	...	0

By comparing Tables 1 and 2 deal-wise we can discover two patterns, where items are positively correlated: (S_1, R_2, R_3) that occurs in 3 out of 6 deals and (S_n, R_1, R_n) occurs in 2 out of 6 deals. Furthermore, the conditional probabilities $P(R_2, R_3|S_1) = 1$ and $P(R_1, R_n|S_n) = 1$. Thus, we can construct two positive association rules $S_1 \Rightarrow (R_2, R_3)$ and $S_n \Rightarrow (R_1, R_n)$ that can be used for predicting cost-overrun elements given solution elements in future deals.

Rule $S_1 \Rightarrow (R_2, R_3)$ means that a presence of S_1 in the proposed solutions was frequently followed by cost-overruns on (R_2, R_3) in T&T phases, where S_1 was not under-costed (no cost-overrun element R_1) but (R_2, R_3) were missing in the corresponding solutions (no solution elements S_2 and S_3). This rule suggests that in future deals if S_1 is part of a proposed solution then S_2 and S_3 should be also included.

Rule $S_n \Rightarrow (R_1, R_n)$ means that a presence of S_n in proposed solutions was followed by cost-overruns on (R_1, R_n) in T&T phases, where S_n was under-costed (has a corresponding cost-overrun element R_n) and R_1 was missing in the corresponding solutions. This rule suggests that in future deals if S_n is part of a proposed solution then it should be higher costed and R_1 should also be part of the solution.

There is also a negatively correlated pattern $(S_1, \neg R_1, \neg R_n)$, where $\neg R_i$ means absence (not) R_i and $P(\neg R_1, \neg R_n | S_1) = 1$ leads to a negative association rule $S_1 \Rightarrow (\neg R_1, \neg R_n)$ [16]. Such negative rules can be useful for identifying sets of solution elements that generate small cost-overruns.

Table 3 summarizes the cases of causal relationships between solution elements and the corresponding cost-overrun elements. Clearly, the case $S_i = 0, R_i = 1$, means that S_i was missing in the proposed solution. The case $S_i = 1, R_i = 0$, means that although the cost case for S_i was done accurately, S_i may be involved in two sub-cases (I) it may cause cost-overruns on other components (e.g., data center relocation usually involves overruns on network hardware) so it is useful for prediction and (II) it does not cause overruns on other elements in which case it may be used as a template of proper (minimal) costing/solution (e.g., a significant pattern consisting of only solution elements).

Table 3. Possible deal-wise causal relationship between solution elements and corresponding cost-overrun elements

S_i	R_i	Description
0	1	S_i was not proposed but R_i occurred during the T&T phase of the deal
1	0	S_i was proposed but R_i did not occur during the T&T phase of the deal
1	1	S_i was proposed and R_i occurred during the T&T phase of the deal meaning S_i was under-cost.

Consider again rule $S_1 \Rightarrow (R_2, R_3)$, where $P(R_2, R_3 | S_1) = 1$ and assume, for simplicity, that every cost-overrun element has a unit cost 1 USD. Then on average (R_2, R_3) in Table 2 costs 2 USD per deal. This leads to the average risk of a cost-overrun as a result of proposing solution element S_1 that equals to $P(R_2, R_3 | S_1) \times 2 = 2$ USD.

Thus, we are interested in rules of the form [(solution element X) $\Rightarrow$ (overrun element Y),$P(Y|X)$, $Risk(Y|X)$]. For example, given a rule

$$[DataCenterRelocation \Rightarrow LAN/WAN, 1, 6.3 \cdot 10^6]$$

in historical data, we can predict that if $DataCenterRelocation$ is part of solution in a new deal then it will incur an overrun on LAN/WAN with probability 1, leading to $Risk = 6.3$ million USD.

We build the pattern-based solution risk model as follows: (I) we discover significant co-occurrence relationships between solution and cost-overrun elements and (II) we build predictive risk models and templates of proper solutions (minimum risk). Once the model is built we can use it to learn significant rules from the past deals and/or to filter rules relevant to solution elements of a new deal in order to assess the risk implied by the proposed solution elements.

1.2 Related Work and Contributions

The most related work to ours is [3] that proposed a case study in IT-operational risk measurement in the context of a network of Private Branch Exchanges

(PBXs). The approach relies on preprocessing and data mining tasks for the extraction of sequential patterns and their exploitation in the definition of a measure called expected risk.

The contributions of our work are as follows: (I) it provides the first pattern-based solution risk model for strategic IT outsourcing and (II) it introduces a general solution risk model for service delivery, whose task is to minimize the gap between the proposed and implemented service elements by the provider based on historical deals.

The paper is organized as follows. Section 2 provides foundation of our work, Section 3 presents details of the proposed pattern-based solution risk model for strategic IT outsourcing, Section 4 presents experiments on SO data of a multinational SO provider.

2 Foundations

In this section we review important concepts that the necessary to explain our framework.

2.1 Itemset Mining

Let $\mathcal{A} = \{a_1, a_2, \ldots, a_{|\mathcal{A}|}\}$ be a set of items (alphabet). A subset $\mathcal{I} \subseteq \mathcal{A}$, where $\mathcal{I} = \{a_1, a_2, \ldots, a_{|\mathcal{I}|}\}$ is called an *itemset* or *element* and is also denoted by $(a_1, a_2, \ldots, a_{|\mathcal{I}|})$, where $|\mathcal{I}|$ denotes the size of the set. Thus, $\mathcal{I}$ is a *subitemset* of $\mathcal{A}$ and $\mathcal{A}$ is the *superitemset* of $\mathcal{I}$. Given a *collection of itemsets* $\mathcal{D} = \{\mathcal{I}^{(1)}, \mathcal{I}^{(2)}, \ldots, \mathcal{I}^{(|\mathcal{D}|)}\}$ (a multiset of sequences) the *support* (frequency) of an itemsets $\mathcal{I}$, denoted by $sup_{\mathcal{D}}(\mathcal{I})$, is defined as the number of itemsets $\mathcal{I}^{(i)} \in \mathcal{D}$ that contain $\mathcal{I}$ as a subset. The *relative support* (relative frequency) $rsup_{\mathcal{D}}(\mathcal{I}) = \frac{sup_{\mathcal{D}}(\mathcal{I})}{|\mathcal{D}|}$ is the fraction of itemsets that contain $\mathcal{I}$ as a subset. Given a relative support threshold $minRelSup$ an itemset $\mathcal{I}$ is called a *frequent itemset* if $rsup_{\mathcal{D}}(\mathcal{I}) \geq minRelSup$. The problem of frequent itemset mining is to find all frequent itemsets in $\mathcal{D}$ given $minRelSup$. An itemset $\mathcal{I}$ is called a *frequent closed itemset* if none of its frequent superitemsets has the same support. Thus, mining closed itemset reduces the number of discovered patterns and provides a more compact representation.

Table 4 presents an example collection of itemsets, where for $minRelSup = 0.5$, $\mathcal{I} = (a_1, a_2)$ is a frequent itemset, where $sup_{\mathcal{D}}(\mathcal{I}) = 0.5$ and it is contained in itemsets: $id = 0, 3$. One of the most efficient algorithms for finding frequent closed itemsets is [12].

Given an itemset $\mathcal{I}$ an *itemset rule* or *itemset association rule* is defined as an implication of the form $X \Rightarrow Y$ where $X, Y \subseteq \mathcal{I}$ and $X \cap Y = \emptyset$. The sets of itemsets X and Y are called *antecedent* (left-hand-side or LHS) and *consequent* (right-hand-side or RHS) of the rule respectively. The confidence (conditional probability) of a rule is defined as follows

Table 4. A collection of itemsets

id	itemsets			
	a_1	a_2	a_3	a_4
0	1	1	0	0
1	0	1	1	0
2	0	0	0	1
3	1	1	1	0

$$conf_{\mathcal{D}}(X \Rightarrow Y) = \frac{sup_{\mathcal{D}}(X \cup Y)}{sup_{\mathcal{D}}(X)} = P(Y|X), \tag{1}$$

where $X \cup Y$ means that both X and Y are present, i.e., $sup_{\mathcal{D}}(X \cup Y) = sup_{\mathcal{D}}(\mathcal{I})$.

For example, rule $(a_1, a_2) \Rightarrow a_3$ has confidence equal to $0.25/0.5 = 0.5$ in Table 4.

2.2 Significance Measures

Although mining closed itemsets reduces the number of discovered patterns that number may still be too large for appropriately low value of the minimum relative support threshold *minRelSup*. Therefore, different measures of significance (strength of correlation between items) have been proposed for itemsets (see [15] for a review). In [4] a maximum entropy model for itemsets was presented, that does not suffer from the empty itemset phenomenon.

In general given two itemsets X and Y there are the following three correlation relationships between them

1. $P(Y|X) = P(Y)$, then Y and X are independent
2. $P(Y|X) > P(Y)$, then Y is positively dependent on X, and $X \Rightarrow Y$ is a *positive association rule*
3. $P(Y|X) < P(Y)$, then Y is negatively dependent on X and $X \Rightarrow \neg Y$ is a *negative association rule* (or $\neg Y$ is positively dependent on X)

Using confidence $P(Y|X)$ alone is not enough to determine significance of $X \Rightarrow Y$ because of the following case. Consider Table 5, where even though $P(Y|X) = 1$ and $P(Y) = \frac{3}{4}$ the rule is not interesting. We can fix the problem by requiring that $P(X|Y)$ is comparable to $P(Y|X)$ as in Table 6. This observation leads to *all-confidence* rule-wise significance measure [10] defined as follows

$$allConfidence(X \Rightarrow Y) = \min\{P(Y|X), P(X|Y)\}. \tag{2}$$

Thus, (2) leverages the rank of rules where the antecedent and consequent occur exclusively together.

All-confidence can be generalized to itemset-wise significance measure as follows

$$allConfidence(\mathcal{I}) = \min_{a_i \in \mathcal{I}} \left\{ \frac{sup(\mathcal{I})}{sup(a_i)} \right\}, \tag{3}$$

Table 5. Even though the items are positively correlated, where $P(Y|X) = 1$, $P(Y) = \frac{3}{4}$ the rule is not interesting since $P(X|Y) = \frac{1}{3}$

X	Y
0	0
1	1
0	1
0	1

Table 6. An interesting rule, where $P(Y|X) = 1$, $P(Y) = \frac{3}{4}$ but $P(X|Y) = 1$

X	Y
0	0
1	1
1	1
1	1

where the right hand side of (3) computes the confidence of the least favorable rule (when $|X| = 1$). Thus (3) leverages the rank of items that frequently co-occur.

Once significant itemsets are found using (3) they are used for generating rules for a given minimum confidence threshold $minConf$. Such rules can be generated level-wise using the fact that if for a rule $X \Rightarrow Y$, $P(Y|X) < minConf$ then clearly for any rule $X' \Rightarrow Y$, where $X' \subset X$ $P(Y|X') < P(Y|X)$ [1]. Another condensed representation are mini-max association rules [11].

3 Pattern-Based Solution Risk Model

The problem of building the pattern-based solution risk model is defined as follows.

Let $\mathcal{A}_\mathcal{S}$ be the alphabet of solution elements and $\mathcal{A}_\mathcal{R}$ the alphabet of cost-overrun elements.

Let $\mathcal{S} = \{S^{(1)}, S^{(2)}, \ldots, S^{(n)}\}$ is a collection of tuples of size $|\mathcal{S}|$, where each $\mathcal{S}^{(i)}$ has the following attributes: (I) deal id $S^{(i)}_j.deal\_id = i$; (II) set of solution elements $S^{(i)} = \{S^{(i)}_1, S^{(i)}_2, \ldots, S^{(i)}_{n^{(i)}}\}$, where $S^{(i)}_j \in \mathcal{A}_\mathcal{S}$ and (III) cost of the j-th solution element $S^{(i)}_j.cost$.

Let $\mathcal{R} = \{R^{(1)}, R^{(2)}, \ldots, R^{(n)}\}$ is a collection of tuples of size $|\mathcal{R}|$, where each $\mathcal{R}^{(i)}$ has the following attributes: (I) deal id $R^{(i)}_j.deal\_id = i$; (II) set of overrun elements $R^{(i)} = \{R^{(i)}_1, R^{(i)}_2, \ldots, R^{(i)}_{m^{(i)}}\}$, where $R^{(i)}_j \in \mathcal{A}_\mathcal{R}$ and (III) cost of the j-th overrun element $R^{(i)}_j.cost$.

Given:

- collection of deal-wise sets of solution elements of previous deals $\mathcal{S}$
- collection of deal-wise sets of overrun elements of previous deals $\mathcal{R}$
- set of solution elements $\mathcal{S}' \subseteq \mathcal{A}_\mathcal{S}$ of the current deal.

Task: find a set of association rules

$$[X \Rightarrow Y, Risk(Y|X)],$$

where $X \subseteq \mathcal{S}'$.

Note that $\mathcal{A}_{\mathcal{S}}$ and $\mathcal{A}_{\mathcal{R}}$ refer to the same underlying service alphabet but we use the two alphabets to distinguish the service elements referring to solution elements from the cost-overrun elements for the purpose of itemset mining framework that uses one underlying alphabet of items.

3.1 Deal-Wise Aggregation of Cost-Overrun Elements

Since the same service element may be a subject to several cost-overruns corresponding to multiple tuples $(R_i, R_i.cost)$ in the same deal, we aggregate such tuples deal-wise by creating a tuple $(R_i, \sum_{R_i \in \mathcal{R}} R_i.cost)$ for the corresponding deal in $\mathcal{R}$. The reason for the aggregation is to convert the data to the itemset transaction model for itemset mining.

Table 7 presents an example of such an aggregated table $\mathcal{R}$, where we also show an industry attribute. Now in order to build a risk model it is useful to compute an average cost per deal for an itemset of cost-overrun elements. Let $AvgCost_{\mathcal{D}}(\mathcal{I})$ be the average cost of an itemset in database $\mathcal{D}$ computed as follows

$$AvgCost_{\mathcal{D}}(\mathcal{I}) = \frac{1}{sup_{\mathcal{D}}(\mathcal{I})} \sum_{\mathcal{I}' \in \mathcal{S}, \mathcal{I} \subseteq \mathcal{I}'} \sum_{i \in \mathcal{I}, i \in \mathcal{I}'} cost(i). \tag{4}$$

As an example consider itemsets (R_1, R_5) and (R_2, R_3) in Table 7, where $AvgCost_{\mathcal{R}}(R_1, R_5)$ =16.3 and $AvgCost_{\mathcal{R}}(R_2, R_3) = 5$.

Table 7. An example of a table $\mathcal{R}$ where costs of multiple elements R_i are deal-wise aggregated

$Deal\_Id$	Industry	R_1	R_2	R_3	R_4	R_5
1	banking	10	0	0	0	7
2	health	0	3	2	0	0
3	banking	9	5	0	0	8
4	health	0	4	1	0	0
5	manufacturing	0	0	5	5	0
6	banking	8	0	0	0	7

Let $VarCost_{\mathcal{D}}(\mathcal{I})$ be the variance of the cost of an itemset in database $\mathcal{D}$ defined as follows:

$$VarCost_{\mathcal{D}}(\mathcal{I}) = \frac{1}{sup_{\mathcal{D}}(\mathcal{I})} \sum_{\mathcal{I}' \in \mathcal{S}, \mathcal{I} \subseteq \mathcal{I}'} \sum_{i \in \mathcal{I}, i \in \mathcal{I}'} \left(cost(i) - AvgCost_{\mathcal{D}}(\mathcal{I})\right)^2. \tag{5}$$

Given $VarCost_{\mathcal{D}}(\mathcal{I})$ we can design a ranking function called normalized average cost that ranks itemsets with respect to homogeneity of the average cost defined as follows:

$$normAvgCost_{\mathcal{D}}(\mathcal{I}) = \frac{AvgCost_{\mathcal{D}}(\mathcal{I})}{1 + \sqrt{VarCost_{\mathcal{D}}(\mathcal{I})}}. \tag{6}$$

Thus (6) leverages rank of itemsets that have homogeneous (small fluctuation) average costs.

3.2 Risk and Normalized Risk

Let $P(Y|X)$ be a conditional probability, where $X \subseteq \mathcal{I}$, $Y \subseteq \mathcal{I}$ and $X \cap Y = \emptyset$. Then clearly the risk $Risk(Y|X)$ can be expressed as follows

$$Risk(Y|X) = P(Y|X) \cdot AvgCost(Y). \tag{7}$$

As a ranking method for selecting the most actionable rules we use *normalized risk* defined as follows

$$normRisk(Y|X) = P(Y|X) \cdot normAvgCost(Y), \tag{8}$$

where normalized risk leverages rank of rules involving small fluctuations (small variance) of the cost values.

3.3 Pruning Redundant Patterns

We prune *redundant patterns* (patterns having the same semantic information and similar significance rank) by removing a pattern s if there exits another pattern s' such that $s \subseteq s'$, $allConfidence(s) > allConfidence(s')$ and

$$\frac{allConfidence(s) - allConfidence(s')}{allConfidence(s')} < pruningThreshold, \tag{9}$$

where we set $pruningThreshold = 0.05$, meaning s is contained in s' and the rank of s is not significantly greater than the rank of s' so s is redundant [6].

3.4 Algorithm for Building the Pattern-Based Risk Model

Input:

- minimum support threshold $minRelSup$
- minimum conditional probability threshold $minConf$ and
- minimum all-confidence threshold $minAllConf$
- minimum risk threshold $minRisk$
- collection of deal-wise solution elements $\mathcal{S}$ and corresponding cost-overrun elements $\mathcal{R}$.

The algorithm proceeds as follows:

1. Join $\mathcal{S}$ and $\mathcal{R}$ on $deal\_id$ attribute, i.e., $\mathcal{C} = \mathcal{S} \bowtie_{deal\_id} \mathcal{R}$
2. Obtain the set of frequent closed itemsets of $\mathcal{C}$
3. Find all frequent patterns $\mathcal{I} \in \mathcal{F}$ in $\mathcal{C}$, such that $X \cup Y = \mathcal{I}$ where:
   (a) X is a subset of solution elements in $\mathcal{S}$
   (b) Y is a subset of overrun elements in $\mathcal{R}$
   (c) prune insignificant patterns $allConfidence(\mathcal{I}) \geq minAllConf$
   (d) prune redundant patterns using Equation (9)
4. Generate actionable association rules, of the form: $[X \Rightarrow Y, Risk(Y|X)]$, where $P(Y|X) \geq minConf$ and $Risk(Y|X) \geq minRisk$
5. Rank the rules with respect to $normRisk(Y|X)$.

4 Experiments

We conducted experiments on a collection of cost case and PCR data of a multinational SO provider.

The solution elements and cost-overrun elements in the data set are organized into a *Service Catalog* (SC), that is the standard hierarchical list of services provided by the provider to customers in SO. SC is represented with a tree of height three, where every service element is characterized by three levels of nodes starting from the most general to the most specific as follows: (I) *Service*: nodes at level 1; (II) *ServiceSegment*: nodes at level 2 and (III) *ServiceElement*: nodes at level 3. Figure 2 presents a fragment of the subtree, where *Service* is equal to *DataNetworkServices*, *ServiceSegment* is equal to *Installation* and *SeriviceElement* (leaf) is equal to *Hardware* and *Software* correspondingly. In the experiments we used the *ServiceSegment* level with alphabet of size 50 as a trade-off between alphabet specificity and representation in the data set in order to guarantee a formation of representative frequent patterns. Thus, using *Service* would mean the smallest and the most general alphabet with large representation while using *ServiceElement* would mean the largest and the most specific alphabet with sparse representation.

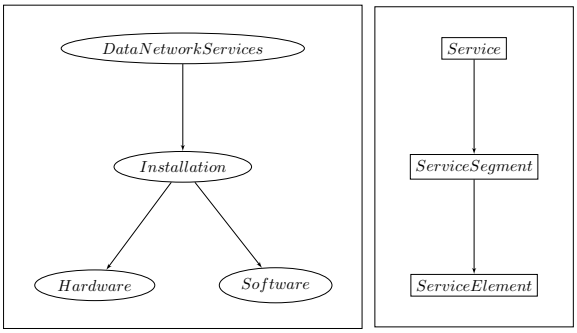


Fig. 2. Example of a *Service Catalog* hierarchy (left) and the corresponding level names (right)

The main purpose of the experiments was to demonstrate examples of actionable association rules between solution elements and cost-overrun elements in Section 4.2. However, since we had only 97 deals we also show some correlations between the cost-overrun elements themselves in Section 4.1.

4.1 Correlated Cost-Overrun Elements

The purpose of finding correlations between cost-overrun elements was to discover potential patterns that my establish right hand sides in rules involving solution elements. We start by presenting the top-10 most frequent cost-overrun

Table 8. Top-10 most frequent cost-overrun elements with their frequency and risk

	Element	$P(Y)$	$Risk(Y)$\$
1	TransitionManagementService	5.05e-01	2.82e+06
2	OverallTechnicalTransition	2.58e-01	1.53e+06
3	LAN/WAN	2.47e-01	6.96e+05
4	TransitionManagement	2.27e-01	2.11e+06
5	StandardTransformation	2.06e-01	1.57e+06
6	OtherProjects	2.06e-01	8.96e+05
7	HelpdeskTransition	1.75e-01	1.93e+05
8	TransitionManagerProjectSupportServices	1.65e-01	2.84e+05
9	CustomTransformation	1.44e-01	1.81e+06
10	ServerSecurityCompliance	1.34e-01	1.22e+06

elements with their frequency and risk in Table 8, where the most frequent pattern corresponds to cost-overruns in labor.

Table 9 presents the top-10 most correlated overrun patterns, ranked with respect to the all-confidence measure, that were discovered for $minRelSup = 0.03$. Thus, Table 9 suggests that since the most correlated patterns are of size two we should not expect lengthy consequents of the rules in Section 4.2.

Table 9. Top-10 most correlated cost-overrun elements ranked with respect to the all-confidence measure

	Pattern	$P(Y)$	$Risk(Y)$\$	allConfidence
1	ServerNTIntelAdministration SecurityMgtServerLogicalAccess	6.19e-02	5.33e+05	5.00e-01
2	GroupwareServicesTransition SecurityMgtServerLogicalAccess	6.19e-02	7.44e+05	5.00e-01
3	ServerConsoleOperations Security	3.09e-02	1.27e+05	4.29e-01
4	ProcessandProceduresManual TransitionArchitect	3.09e-02	4.54e+04	4.29e-01
5	TransitionArchitect LAN/WAN	1.03e-01	1.25e+06	4.17e-01
6	TransitionArchitect TransitionManagement	9.28e-02	5.56e+05	4.09e-01
7	CustomTransformation StandardTransformation	8.25e-02	9.16e+05	4.00e-01
8	SecurityServerLogicalAccess AssetManagementAndTracking	5.15e-02	1.73e+05	3.85e-01
9	GroupwareServicesTransition AssetManagementAndTracking	5.15e-02	9.74e+04	3.85e-01
10	TransitionManagerProjectSupportServices LAN/WAN	9.28e-02	5.06e+05	3.75e-01

4.2 Correlated Solution and Overrun Elements

We used the following parameters for building the pattern-based solution risk model: $minRelSup = 0.3$, $minConf = 0.75$, $minAllConf = 0.33$, $minRisk = 0.1e{+}06$ and we ranked the rules with respect to the normalized risk value. Table 10 presents the resulting rules involving solution and cost-overrun elements.

Table 10. Example of discovered rules $X \Rightarrow Y$ for $minRelSup = 0.3$, $minConf = 0.75$, $minAllConf = 0.33$, $minRisk = 0.1e + 06$ and ranked with respect to the normalized risk, where X is a set of solution elements and Y is a set of cost-overrun elements.

ID	X (solution elements)	Y (overrun elements)	$P(Y\|X)$	$Risk(Y\|X)$	$P(X,Y)$
1	DataNetwork, EndUserServices, ServiceSpecificInfrastructure, Server	OverallTechnical-TransitionManager	0.75	0.6e+06	0.3
2	Transition, ServiceSpecificInfrastructure, Server	HelpdeskTransition, TransitionManagement-CrossService, LAN/WAN	0.75	0.57e+06	0.3
3	TransformationManagement, AccountManagement, AssetServices, ServiceSpecificInfrastructure, TransitionBase, Server	TransitionManagement-GlobalServiceDelivery	0.75	0.56e+06	0.3
4	EndUserHelpdesk, LeadAccountManagementDeal, AccountManagement, AssetServices, ServiceSpecificInfrastructure, TransitionBase, Server	AssetManagementAnd-TrackingHardware, HelpdeskTransition	0.75	0.3e+06	0.3

5 Conclusions and Extensions

We presented a new pattern-based risk model for strategic IT outsourcing based on frequent closed itemset mining and association rule mining. In experiments, conducted on a sample of deals of a multinational SO provider we showed the applicability of the framework for predicting significant cost overruns based on solution elements.

The introduced pattern-based solution risk model is a general solution risk model for service delivery, whose task is to minimize the gap between proposed and implemented service elements by the provider based on historical deals. For example, consider a construction company that performs building renovation services in which case it is very important to design a realistic cost case that takes into account possible risks.

There are many extensions of the current framework including: (I) adding multi-level association rules [5] of the service catalog in order to discover interesting cross-level rules; (II) designing a relevant measure of interestingness for cross-level rules; (III) adding negative association rules [16] to identify sets of solution elements that generate the smallest overruns; (IV) quantizing/classifying the cost values into a number of categories by relating them to the total contract value, etc. and (V) adding client attributes (e.g., industry, number of employees, location) together with solution elements.

We envision building a system to support project managers in SO contracts in properly designing the cost case. Thus, given a set of proposed solution elements as the input the system would find relevant rules. The appeal of the multi-level association rules given the service catalog from Figure 2 would be to deal with sparseness of the data. Thus, the system would start from the lowest level (largest alphabet) and if no significant frequent itemsets were found because of the sparse alphabet then the system would move to the next upper level in the hope of finding frequent siginifcant itemsets.

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: Proceedings of the 20th International Conference Very Large Data Bases, pp. 487–499 (September 1994)
2. D'Aveni, R.A., Ravenscraft, D.J.: Economies of integration versus bureaucracy costs: Does vertical integration improve performance? The Academy of Management Journal 37(5), 1167–1206 (1994)
3. Grossi, V., Romei, A., Ruggieri, S.: A case study in sequential pattern mining for it-operational risk. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008, Part I. LNCS (LNAI), vol. 5211, pp. 424–439. Springer, Heidelberg (2008)
4. Gwadera, R., Crestani, F.: Ranking sequential patterns with respect to significance. In: Zaki, M.J., Yu, J.X., Ravindran, B., Pudi, V. (eds.) PAKDD 2010, Part I. LNCS (LNAI), vol. 6118, pp. 286–299. Springer, Heidelberg (2010)
5. Han, J., Fu, Y.: Mining multiple-level association rules in large databases. IEEE Transactions on Knowledge and Data Engineering 11(5) (1999)
6. Huang, X., An, A., Cercone, N.: Comparison of interestingness functions for learning web usage patterns. In: Proceedings of the Eleventh International Conference on Information and Knowledge Management, CIKM 2002, pp. 617–620. ACM, New York (2002)
7. Kern, T., Willcocks, L.P., Heck, E.V.: Winner's curse in it outsourcing: Strategies for avoiding relational trauma. California Management Review 44(2), 47–69 (2002)
8. Mahnke, V., Overby, M.L., Vang, J.: Strategic it-outsourcing: What do we know and need to know? In: DRUID Summer Conference 2003 on Creating, Sharing and Transferring Knowledge, Copenhagen, June 12-14 (2003)
9. Martens, B., Teuteberg, F.: Why risk management matters in it outsourcing a systematic literature review and elements of a research agenda. In: 17th European Conference on Information Systems, Verona, Italy, June 8-10 (2009)
10. Omiecinski, E.R.: Alternative interest measures for mining associations in databases. IEEE Transactions on Knowledge and Data Engineering 15, 57–69 (2003)

11. Pasquier, N., Taouil, R., Bastide, Y., Stum, G., Lakhal, L.: Generating a condensed representation for association rules. Journal of Intelligent Information Systems 24(1), 29–60 (2005)
12. Pei, J., Han, J., Mao, R.: Closet: An efficient algorithm for mining frequent closed itemsets. In: Proceedings of the 2000 ACM-SIGMOD International Workshop of Data Mining and Knowledge Discovery (DMKD 2000), pp. 21–30 (2000)
13. Quinn, J.B., Helmer, F.G.: Strategic outsourcing. Sloan Management Review, 43–55 (Summer 1994)
14. Ruzaini, S., Aris, S.: Risk management practices in it outsourcing projects. In: 2008 International Symposium on Information Technology, ITSim 2008, August 26-28, vol. 4, pp. 1–8 (2008)
15. Tatti, N.: Maximum entropy based significance of itemsets. KAIS 17(1), 57–77 (2007)
16. Wu, X., Zhang, C., Zhang, S.: Efficient mining of both positive and negative association rules. ACM Trans. Inf. Syst. 22, 381–405 (2004)

Mining Semantic Relationships between Concepts across Documents Incorporating Wikipedia Knowledge

Peng Yan and Wei Jin

Department of Computer Science, North Dakota State University
1340 Administration Ave., Fargo, ND 58102, USA
{peng.yan,wei.jin}@ndsu.edu

Abstract. The ongoing astounding growth of text data has created an enormous need for fast and efficient text mining algorithms. Traditional approaches for document representation are mostly based on the Bag of Words (BOW) model which takes a document as an unordered collection of words. However, when applied in fine-grained information discovery tasks, such as mining semantic relationships between concepts, sorely relying on the BOW representation may not be sufficient to identify all potential relationships since the resulting associations based on the BOW approach are limited to the concepts that appear in the document collection literally. In this paper, we attempt to complement existing information in the corpus by proposing a new hybrid approach, which mines semantic associations between concepts across multiple text units through incorporating extensive knowledge from Wikipedia. The experimental evaluation demonstrates that search performance has been significantly enhanced in terms of accuracy and coverage compared with a purely BOW-based approach and alternative solutions where only the article contents of Wikipedia or category information are considered.

Keywords*:* Knowledge Discovery, Semantic Relatedness, Cross-Document knowledge Discovery, Document Representation.

1 Introduction

With the explosive growth of text data and growing demand for high-quality text mining algorithms, document representation and semantic relatedness computation approaches are increasingly crucial. Traditionally text documents are represented as a Bag of Words (BOW) and relatedness between concepts are measured based on statistical information from the corpus such as the widely used tf-idf weighting scheme [12, 14]. Recently, [5, 12] introduced an interesting text mining scenario focusing on detecting links between two concepts across multiple documents. Typically, the uncovered links involving concepts A and B have the following meaning: find the most plausible relationship between concept A and concept B assuming that one or more instances of both concepts occur in the corpus, but not necessarily in the same document. For example, both may be football lovers, but maybe mentioned in different documents. However, the techniques proposed in [5, 12] were all built under the

P. Perner (Ed.): ICDM 2013, LNAI 7987, pp. 70–84, 2013.

assumption of BOW-based representation, and thus demonstrating their inherent limitations. For example, the detected links are limited to the associations occurring in the document collection; any potential relationships not appearing in the corpus cannot be discovered even though they are closely related to two concepts of interest. The semantic relatedness computing methods used in [5, 12] were mainly based on statistical information collected from the corpus and no background knowledge has been taken into account.

To alleviate all such limitations, this work proposes Semantic Path Chaining (SPC), a new model to uncover semantic paths between concepts with a focus on taking background knowledge into consideration. The approach proposed here is based on the method proposed by Srinivasan's closed text mining algorithm [13] in the biomedical domain, but we extend it to handle a more complicated query scenario where multiple-stage semantic paths are desired and also attempt to incorporate Wikipedia knowledge to enrich document representation. Motivated by the Explicit Semantic Analysis (ESA) technique introduced by Gabrilovich et al. [2], which was able to use the space of Wikipedia articles to measure the semantic relatedness between fragments of natural language text, we develop a hybrid approach and weighting scheme that combines the advantages of ESA and content based statistical analysis. Another distinct difference from the original ESA method is that [2] only focused on document-level textual analysis through mapping a given text fragment or term to a conceptual vector space spanned by all Wikipedia articles, whereas here we extend this technique by considering other valuable evidences from Wikipedia such as categories associated with each Wiki concept to further improve the semantic relatedness estimation between concepts.

Our contribution of this paper can be summarized as follows. First, compared with traditional methods mostly based on the BOW representation, the proposed technique is able to provide a much more comprehensive knowledge repository to support various queries and effectively complements existing knowledge contained in text corpus. Over 5,000,000 Wikipedia articles and more than 700,000 Wikipedia categories are considered to help measure the semantic relatedness between concepts. Therefore the relationships revealed are not limited to those appearing in the document collection literally. Also we observe for connections between rare concepts where we have little knowledge about them, the relationships discovered are often more than one level of transitivity and most of them cannot be uncovered unless integrating the knowledge from Wikipedia. Second, besides content analysis on Wikipedia articles, the new solution also integrates other valuable information, such as Wiki categories, as an effective aid in providing a better modeling of semantic relatedness estimation (based on the assumption that two concepts that share more common categories may have a closer relationship), and thus being able to boost linking concepts that are more closely related to topics of interest to higher rankings. We envision this integration would also benefit other related tasks such as question answering and cross–document summarization. Third, the discovered potential relationships have been greatly enriched by including intermediate concepts (linking terms) derived from Wikipedia, and for these newly identified connections not appearing in the text corpus, we further introduce a pruning and validation step through an application of a sequence of devised

heuristics. Last, to demonstrate the effectiveness of our new model, a significant amount of queries covering various scenarios were conducted, evaluated, and compared against the BOW based baseline. We have also further evaluated the performance of using our adapted ESA method, the approach only incorporating the Wikipedia category information, as well as the solution combining both of the above two resources, respectively. The experiments demonstrate the significant improvement achieved by our proposed method over the original ESA method and other alternative solutions.

This paper is organized as follows: Section 2 describes related work. Section 3 introduces our new semantic relatedness computation measures. In Section 4, we discuss the new model of mining semantic relationships between concepts incorporating Wikipedia knowledge in detail. Experimental results are presented and analyzed in Section 5. Section 6 concludes this work and describes future directions.

2 Related Work

Most of existing text mining algorithms for capturing relationships between concepts have built on the traditional Bag-of-Words representation and significant efforts have been paid to content analysis of document collections with no or little background knowledge being taken into account [12, 14, 15], thus resulting in a limited discovery scope. To alleviate such problems, there has been work recently reported on exploring methods of utilizing external knowledge to assist in the discovery tasks. Bollegara et al. [1] developed an approach for semantic relatedness calculation using returned page counts and text snippets produced by a Web search engine. Mehmet Ali Salahli [9] proposed another Web oriented method that calculated semantic relatedness between terms using a set of determiners (special words that are supposed to be highly related to terms of interest). However, the performance of these approaches highly relies on the generated outputs from search engines and has not reached the satisfying level. WordNet based approaches are another direction to approach this problem, especially in handling synonym, hyponymy/hypernymy relations. Hotho et al. [4] exploited WordNet to improve the BOW text representation and Martin [6] developed a method for transforming the noun-related portions of WordNet into a lexical ontology to enhance knowledge representation. Scott and Matwin [10] proposed a new representation of text based on WordNet hypernyms. These WordNet-based techniques have shown their advantages of improving the tradition BOW based representation to some degree but suffer from relatively limited coverage of Wordnet compared to Wikipedia, the world's largest knowledge base to date. Gabrilovich et al [3] applied machine learning techniques to Wikipedia and proposed a new method to enrich document representation from this huge knowledge repository. Specifically, they built a feature generator to identify most relevant Wikipedia articles for each document, and then used concepts corresponding to these articles to create new features. The experimental evaluation showed great improvements across a diverse collection of datasets. However, with the process of feature generation so complicated, a considerable computational effort is required.

In terms of improving semantic relatedness computation using Wikipedia, Milne [7] proposed a Wikipedia Link Vector Model (WLVM) for this purpose. However, only the hyperlink structure of Wikipedia and article titles were extracted to compute semantic relatedness between query terms, without any analysis of textual contents of Wikipedia articles. Gabrilovich et al. [2] presented a novel method, Explicit Semantic Analysis (ESA), for fine-grained semantic representation of unrestricted natural language texts. Using this approach, the meaning of any text can be represented as a weighted mixture of Wikipedia-based concepts (articles), called an interpretation vector [2]. [2] also discussed the problem of possibly containing noise concepts in the vector, especially for text fragments containing multi-word phrases (e.g., multi-word names like George Bush). Our proposed solution is motivated by [2, 7] and to tackle the above problems, we adapt the ESA technique to better suit our task and further develop a sequence of heuristic strategies to filter out irrelevant terms and retain only top-k most relevant concepts to the given topics. Moreover, other than content-based statistical information of Wikipedia articles being incorporated, other valuable evidence sources provided by Wikipedia, such as categories associated with each concept, are also combined into our final concept ranking scheme. The detailed experimental results and comparisons will be presented in section 5.

3 Semantic Path Chaining

Semantic Path Chaining (SPC) is attempting to mine semantic paths between two concepts (e.g., two person names) across documents incorporating Wikipedia knowledge. We propose to use the features extracted from text corpus, as well as the relationships discovered from Wikipedia to construct semantic paths which stand for potential conceptual connections between them.

3.1 Ontology Mapping and Semantic Profile Representation

To detect semantic relationships between topics of interest, we first represent each topic as a semantic profile which is essentially a set of highly related concepts to the given topic in the corpus. To further differentiate between the concepts, semantic type (ontological information) is employed in profile generation. Table 1 illustrates part of semantic type - concept mappings. Thus each profile is defined as a vector composed of a number of semantic types.

Table 1. Semantic Type - Concept Mapping

Semantic Type	Instances
Human Action	attack, killing, covert action, international terrorism
Leader	vice president, chief, governor
Country	Iraq, Afghanistan, Pakistan, Kuwait
Diplomatic Building	consulate, pentagon, UAE Embassy
Government	Bush administration, white house, national security council
Person	deputy national security adviser, chairman, executive director

$$profile(T_i) = \{ST_1, ST_2, ..., ST_n\} \quad (1)$$

Where ST_i represents a semantic type to which the concepts appearing in the topic-related text snippets belong. We used sentence as window size to measure relevance of appearing concepts to the topic term. Under this representation each semantic type is again referred to as an additional level of vector composed of a number of terms that belong to this semantic type.

$$ST_i = \{w_{i,1}m_1, w_{i,2}m_2, ..., w_{i,n}m_n\} \quad (2)$$

Where m_j represents a concept belonging to semantic type ST_i, and $w_{i,j}$ represents its weight under the context of ST_i and sentence level closeness. When generating the profile we replace each semantic type in (1) with (2). In (2), to compute the weight of each concept, we employ a variation of the *TF*IDF* weighting scheme and then normalize the weights:

$$w_{i,j} = s_{i,j} / highest(s_{i,l}) \quad (3)$$

Where $l = 1, 2, ..., r$ and there are totally r concepts for ST_i, $s_{i,j} = df_{i,j} * Log(N / df_j)$, where N is the number of sentences in the collection, df_j is the number of sentences concept m_j occurs, and $df_{i,j}$ is the number of sentences in which topic T and concept m_j co-occur and m_j belongs to semantic type ST_i. By using the above three formulae we can build the corresponding profile representing any given topic.

To summarize, the procedure of building semantic profiles for a given topic T of interest is composed of the following four steps:

1. Concept Extraction: extract all potential concepts from the document collection which co-occur with the topic T in the sentence level.
2. Semantic Type Employment: each concept will be associated with and grouped under one or more semantic types (e.g., Human Action, Country, Person) which it belongs to.
3. Weight Calculation: for each concept, a variation of the *TF*IDF* scheme is used to calculate its weight (as shown in Formula 2).
4. Weight Normalization: within each semantic type, the concept weights are further normalized by the highest concept weight observed for the semantic type as given in Formula 3, and then ranked according to the normalized weights.

3.2 Chaining Semantic Paths

In this step, we search potential conceptual connections in different levels, and use them to construct semantic paths linking two given topics (concepts). Suppose A and C are two given topics of interest, the algorithm of generating semantic paths connecting A to C from the text corpus is composed of the following sequential steps:

1. Conduct independent searches for A and C. Build the A and C profiles. Call these profiles AP and CP respectively.
2. Compute a B profile (BP) composed of terms in common between AP and CP. The corpus-level weight of a concept in BP is the sum of its weights in AP and CP. This is the first level of intermediate potential concepts generated from the text corpus.
3. Expand the semantic paths using the created BP profile together with the topics to build additional levels of intermediate concept lists DP and EP which (i) connect the topics to each concept in BP profile in the sentence level within each semantic type, and (ii) also normalize and rank them (as detailed in section 3.1).

4 Mining Semantic Relationships between Concepts Incorporating Wikipedia Knowledge

4.1 Wiki-article Content Based Measure

To utilize Wikipedia knowledge to complement existing information in the document collection, we adapt the Explicit Semantic Analysis (ESA) technique proposed by Gabrilovich et al. [2] as our underlying content-based measure for analyzing Wikipedia articles relevant to the given topics of interest. Under this measure, each article in Wikipedia is treated as a concept, and each document is represented by an interpretation vector containing related Wikipedia concepts (articles) to the document.

$$\phi(d) =< as(d,a_1),...,as(d,a_n) > \tag{4}$$

Where $as(d,a_i)$ represents the association strength between document d and Wikipedia article a_i. Suppose d is spanned by all words appearing in it, i.e., $d =< w_1, w_2, ..., w_j >$, the association strength $as(d,a_i)$ is computed as follows:

$$as(d,a_i) = \sum_{w_j \in d} tf_d(w_j) tf \cdot idf_{a_i}(w_j) \tag{5}$$

Where $tf_d(w_j)$ is the frequency of word w_j in document d, and $tf \bullet idf_{a_i}(w_j)$ is the $tf \bullet idf$ value of word w_j in Wikipedia article a_i. As a result, the vector for a document is represented by a list of real values indicating the association strength of a given document with respect to Wikipedia articles. By using efficient indexing strategies such as single-pass in memory indexing, the computational cost of building these vectors for a given term (or text fragments containing multiple terms) can be reduced to within 200-300 ms.

As discussed above, the original ESA method [2] is subject to the noise concepts introduced, especially when dealing with multi-word phrases. For example, when the input is "George Bush", the generated interpretation vector will contain a fair amount of noise concepts such as "That's My Bush", which is actually an American comedy television series. This Wikipedia concept (article) is selected and ranked in the second

place in the list because "Bush" occurs many times in the article "That's My Bush", but obviously this article is irrelevant to the given topic "George Bush". In order to make the interpretation vector more precise and relevant to the topic, a sequence of heuristics is devised to clean the vector as shown in Figure 1. More specifically, a modified Levenshtein Distance algorithm is devised to measure the relevance of the given topic to each Wikipedia concept generated in the interpretation vector with a single word as a unit for allowable edit operations, which allows the adapted algorithm to be used to compute the similarity between any two text snippets. If the topic contains only one word, then the number of its occurrences in the corresponding Wikipedia article will be used for judgement. If it occurs more than three times, this article is viewed as relevant to the given topic and is kept in the interpretation vector. If the topic contains multiple words, we will view each word as if it were a character and employ our adapted version of the Levenshtein distance algorithm to evaluate the relevance of the topic to the article text. If their Levenshtein distance is under the defined threshold, the article is viewed as relevant. Otherwise, it will be removed from the interpretation vector.

Input: a topic T of interest
an interpretation vector V representing the topic T

Output: a cleaned Wikipedia-based concept vector V' representing the topic T

1. If T is a single word topic, then count the number of occurrences of T in the article texts represented by each concept v_i in V, respectively. If T occurs more than 3 times, then keep v_i in V, otherwise, remove v_i from V.

2. If T is a multi-word topic, then the adapted Levenshtein distance algorithm applies to measure the relevance of each Wikipedia concept (article) v_i in V to topic T.

2.1. If NumOfWords(T) $\leq$ 2, then extract all text snippets TS_j within the window size NumOfWords(T) from the article text of v_i. If there exists a j such that LevenshteinDistance(T, TS_j) = 0, then keep v_i in V, otherwise, remove v_i from V.

2.2. If NumOfWords(T) > 2, then extract all text snippets TS_j within the window size NumOfWords(T)+1 from the article text of v_i. If there exists j such that LevenshteinDistance(T, TS_j) $\leq$ 1, then keep v_i in V, otherwise, remove v_i from V.

Fig. 1. Interpretation vector cleaning procedure

After the cleaning step, we are able to use the resulting interpretation vectors for computing similarities between any two concepts. In our context of mining associations between two topics, say A and C, we compute the Cosine similarity between the interpretation vectors of topic A and each concept V_i in the intermediate BP profile, as well as between topic C and each concept V_i, and take the average of two Cosine similarities as the overall similarity for each concept V_i in BP profile.

4.2 Wiki-Category Based Measure

Human edited categories associated with each Wiki concept (article), another valuable resource provided by Wikipedia, have also been integrated to better serve this task. Based on the assumption that those concepts (articles) sharing similar categories may be closer to each other in terms of semantic relatedness, a Wikipedia category

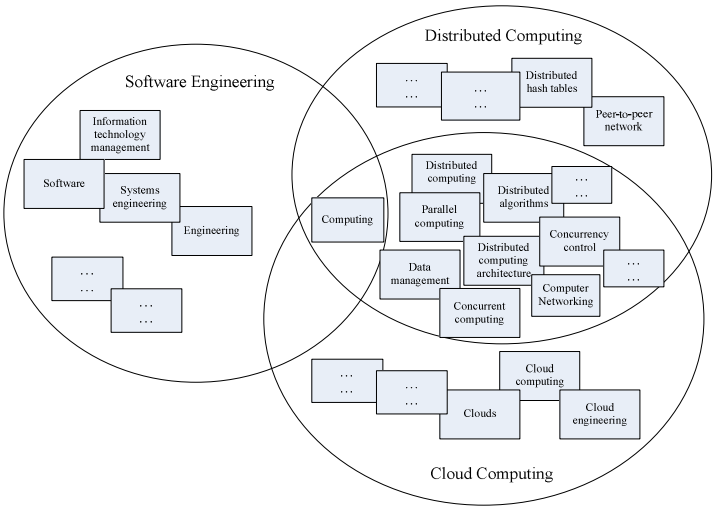


Fig. 2. Category Overlaps of the Concepts in the Interpretation Vectors of "Distributed Computing," "Cloud Computing" and "Software Engineering"

interpretation vector has been built for each desired Wiki concept and the semantic relatedness between two concepts of interest is determined by the percentage of common categories shared by the two corresponding category interpretation vectors.

Formally, suppose the interpretation vector for article a_i is $V_i =< p_1, p_2, ..., p_m >$, where p_i in V_i represents a Wiki page (or article) that is relevant to a_i, then article a_i can be further represented as a *Category Space Vector (CSV)* as follows spanning the Wikipedia category space.

$$CSV(a_i) =<< w_{i,1,1} c_{1,1}, w_{i,2,1} c_{2,1}, ... >, ..., < w_{i,1,m} c_{1,m}, w_{i,2,m} c_{2,m}, ... >> \quad (6)$$

Where $c_{x,y}$ represents category c_x that p_y in V_i belongs to, and $w_{i,x,y}$ is the weight for $c_{x,y}$. To calculate $w_{i,x,y}$, we count the number of sub-vectors within $CSV(a_i)$ in which $c_{x,y}$ appears, and then normalize it:

$$w_{i,x,y} = \frac{w_{i,x,y}}{highest(w_{i,d,y})} \quad (7)$$

Where $d = 1, 2, ..., r$ and there are totally r categories in Wikipedia. The semantic relatedness between two Wikipedia concepts (articles) can then be computed by the Cosine similarity between their corresponding CSVs. Figure 2 shows the categories built for three concepts: "Distributed Computing," "Cloud Computing" and "Software Engineering." The produced semantic relatedness between "Distributed Computing" and "Cloud Computing" is 0.715, 0.094 between "Distributed Computing" and

"Software Engineering", and 0.151 between "Cloud Computing" and "Software Engineering". This is consistent with our understanding that "Distributed Computing" and "Cloud Computing" are more semantically closely related while both further away from "Software Engineering.

4.3 Final Weighting Scheme

A final ranking for each concept generated in the intermediate profiles is calculated by linearly combining its TFIDF-based similarity, content-based similarity and category-based similarity together as below:

$$S_{overall} = \lambda_1 \cdot S_{TFIDF} + \lambda_2 \cdot S_{content} + (1 - \lambda_1 - \lambda_2) \cdot S_{category} \quad (8)$$

Where λ_1 and λ_2 are two tuning parameters that can be adjusted based on the preference on the two similarity schemes in the experiments. S_{TFIDF} refers to the similarity computed using the traditional BOW model, and $S_{content}$ and $S_{category}$ refer to the similarities computed using the content based measure and category based measure respectively.

4.4 The New Model of Mining Semantic Relationships

After defining the semantic relatedness measures between concepts, we are presenting now the new solution for building semantic paths between concepts. Suppose A and C are two given topics of interest, with Wikipedia knowledge incorporated in our model, we are able to leverage Wiki concepts to enrich the relationships (i.e., not limited to those occurring in the document collection literally). Thus the generated links would be an integration of relationships identified from the text corpus as well as from Wikipedia knowledge. The process can be summarized as the following major steps and is further illustrated in Figure 3.

1. Build ESA-based interpretation vectors for A and C. Employ the cleaning procedure illustrated in Figure 1 to remove noise concepts in the generated interpretation vectors. The concepts that survived after cleaning are ordered according to their association strength as described in Section 4.1, and will be serving as potentially novel connections between topics A and C.
2. Enrich the generated BP profile with newly identified Wiki concepts (represented by the corresponding Wikipedia article titles) by merging the cleaned interpretation vectors for topics A and C. The weight of each newly identified Wiki concept in BP is the sum of its association strengths in the cleaned interpretation vectors for topics A and C.
3. Go through the same procedures as in the above two steps to enrich DP and EP profiles that contain the intermediate concepts connecting the topics to each concept in BP profile.

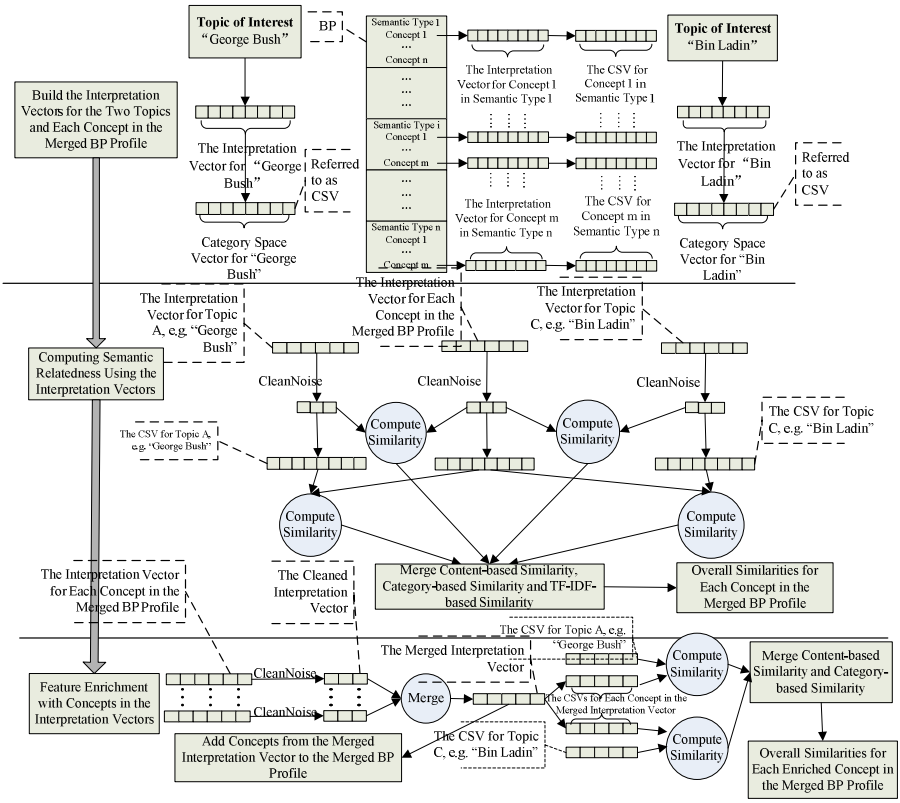


Fig. 3. The new model of mining semantic relationships

4. The BP profile is further enriched by considering relevant Wiki categories that the newly identified Wiki concepts (articles) belong to. The weight of each newly identified Wiki category in BP is the same as that of the corresponding Wiki concept.
5. Go through the same procedure as in Step 4 to enrich DP and EP profiles with the newly identified relevant Wiki categories.

5 Empirical Evaluation

5.1 Processing Wikipedia Dumps

Wikipedia offers free copies of the entire content in the form of XML files. It is an ever-updating knowledge base, and releases the latest dumps to interested users regularly. The version used in this work was released on April 05, 2011, which was separated into 15 compressed XML files and altogether occupied 29.5 GB after de-compression. An open source tool MWDumper [8] was used to import the XML dumps into our MediaWiki database, and after the parsing process, we identified 5,553,542 articles.

5.2 Evaluation Data

An open source document collection pertaining to the 9/11 attack, including the publicly available 9/11 commission report was used in our evaluation. The report consists of Executive Summary, Preface, 13 chapters, Appendix and Notes. Each of them was considered as a separate document resulting in 337 documents. The whole collection was processed using Semantex[11] and concepts were extracted and selected as shown in Table 1. A significant amount of query pairs selected by the assessors covering various scenarios [16] (e.g., ranging from popular entities to rare entities) were conducted and used as our evaluation data.

5.3 Experimental Results

Parameter Settings. As mentioned in Section 4.2, a combination of corpus-level TF*IDF-based similarity, Wiki-article content based similarity and category-based similarity is used to rank the intermediate concepts detected by the system. λ_1 and λ_2 are two parameters that need to be tuned so that the similarities between concepts best match the judgements from our assessors. To accomplish this, we first built a set of training data composed of 10 query pairs randomly selected from the evaluation set, and then generated B profiles for each of them using our proposed method. Among each B profile, we selected the top 5 concepts (links) within each semantic type, and compared their rankings with the assessors' judgements. The values of λ_1 and λ_2 were tuned in the range of [0.1, 1] and we observe the best performance was achieved with $\lambda_1 = 0.1$ and $\lambda_2 = 0.3$. This setting was also used in our later experiments.

Query Results. Tables 2 through 4 make a comparison between the search results of our baseline where the corpus-level TF-IDF based statistical information is used to generate chains without the involvement of Wikipedia and various Wiki-enabled models proposed in this work. Specifically, Table 2 shows the improvement achieved by integrating the Wiki-article content based measure over the baseline; Table 3 presents the result when the relevant Wiki categories are used to improve the discovery

Table 2. The Effect of Integrating the Adapted ESA Technique (original ESA+ Vector Cleaning)

		Baseline/Wiki-ESA					
		S_5	**S_{10}**	**S_{15}**	**S_{20}**	**S_{30}**	**S_{40}**
L_1	P	0.756/0.788	0.764/0.789	0.763/0.786	0.759/0.787	0.759/0.787	0.761/0.789
	R	0.440/0.618	0.538/0.721	0.576/0.763	0.593/0.793	0.624/0.826	0.644/0.849
L_2	P	0.845/0.855	0.844/0.855	0.843/0.853	0.843/0.852	0.842/0.850	0.841/0.849
	R	0.528/0.575	0.573/0.622	0.608/0.659	0.633/0.683	0.657/0.706	0.676/0.723
L_3	P	0.846/0.856	0.845/0.856	0.844/0.854	0.844/0.853	0.843/0.851	0.842/0.850
	R	0.530/0.575	0.573/0.620	0.608/0.658	0.634/0.681	0.657/0.705	0.676/0.722
L_4	P	0.691/0.699	0.689/0.695	0.687/0.692	0.686/0.691	0.684/0.689	0.684/0.689
	R	0.392/0.413	0.513/0.534	0.587/0.610	0.638/0.661	0.690/0.713	0.720/0.744

Table 3. The Effect of Integrating Wikipedia Categories

		Baseline/Wiki-CSV					
		S_5	**S_{10}**	**S_{15}**	**S_{20}**	**S_{30}**	**S_{40}**
	P	0.756/0.767	0.764/0.773	0.763/0.770	0.759/0.767	0.759/0.767	0.761/0.769
L_1	R	0.440/0.589	0.538/0.694	0.576/0.738	0.593/0.759	0.624/0.793	0.644/0.816
	P	0.845/0.856	0.844/0.855	0.843/0.853	0.843/0.852	0.842/0.851	0.841/0.850
L_2	R	0.528/0.580	0.573/0.628	0.608/0.663	0.633/0.687	0.657/0.710	0.676/0.728
	P	0.846/0.857	0.845/0.857	0.844/0.855	0.844/0.854	0.843/0.853	0.842/0.851
L_3	R	0.530/0.580	0.573/0.627	0.608/0.662	0.634/0.686	0.657/0.709	0.676/0.727
	P	0.691/0.702	0.689/0.699	0.687/0.696	0.686/0.694	0.684/0.692	0.684/0.691
L_4	R	0.392/0.422	0.513/0.547	0.587/0.622	0.638/0.673	0.690/0.725	0.720/0.755

Table 4. The Effect of Integrating both ESA and Wikipedia Categories

		Baseline/Wiki-ESA-CSV					
		S_5	**S_{10}**	**S_{15}**	**S_{20}**	**S_{30}**	**S_{40}**
	P	0.756/0.798	0.764/0.818	0.763/0.814	0.759/0.810	0.759/0.809	0.761/0.809
L_1	R	0.440/0.648	0.538/0.840	0.576/0.880	0.593/0.898	0.624/0.929	0.644/0.949
	P	0.845/0.864	0.844/0.865	0.843/0.862	0.843/0.861	0.842/0.859	0.841/0.865
L_2	R	0.528/0.625	0.573/0.679	0.608/0.713	0.633/0.736	0.657/0.758	0.676/0.727
	P	0.846/0.866	0.845/0.865	0.844/0.863	0.844/0.862	0.843/0.860	0.842/0.858
L_3	R	0.530/0.625	0.573/0.676	0.608/0.710	0.634/0.734	0.657/0.756	0.676/0.772
	P	0.691/0.709	0.689/0.705	0.687/0.701	0.686/0.699	0.684/0.696	0.684/0.695
L_4	R	0.392/0.443	0.513/0.570	0.587/0.645	0.638/0.696	0.690/0.748	0.720/0.778

model; Table 4 demonstrates the overall benefit when both the Wiki article contents and Wiki categories are incorporated. The table entries can be read as follows: S_N means we only keep the top N concepts within each semantic type in the searching results and L_N indicates the resulting chains of length N. The entries in the three tables stand for the precision and recall values (P for precision and R for recall). It is easy to observe that the search performance has been significantly improved with the integration of Wikipedia knowledge, and the best performance is observed when both the Wiki article contents and categories are involved.

We further use $F-measure$ to interpret the query results as a harmonic mean of the precision and recall. Figures 4 through 7 compare the search results graphically between the baseline and our new models in terms of *F-scores* for chains of different lengths. The X-axis indicates the number of concepts kept in each semantic type in the search results (S_N means the top N are kept), while the Y-axis indicates the *F-score.* We can see that the achieved F-score continues to rise as we increase the number of top concepts kept in the search results, and the most significant upward trend was observed when the number of top concepts kept increased from 5 to 10. It is also obvious that our new model consistently achieves better performances for different lengths than the baseline solution, and the approach that integrates both the Wiki article contents and categories shows the most improvement.

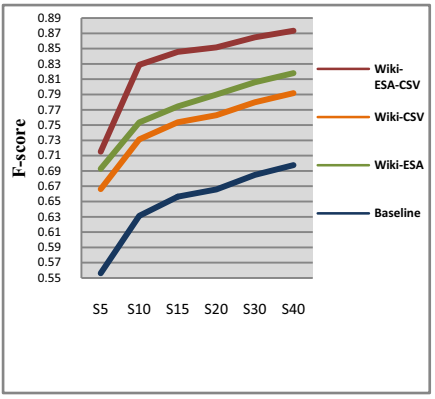


Fig. 4. The result of length 1

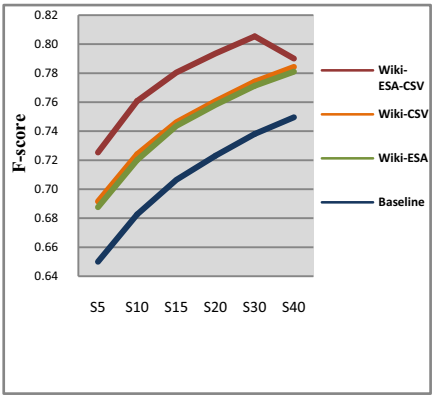


Fig. 5. The result of length 2

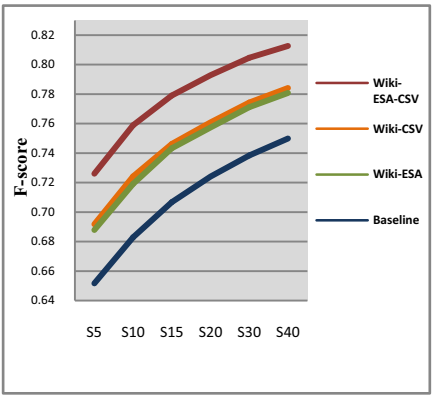


Fig. 6. The result of length 3

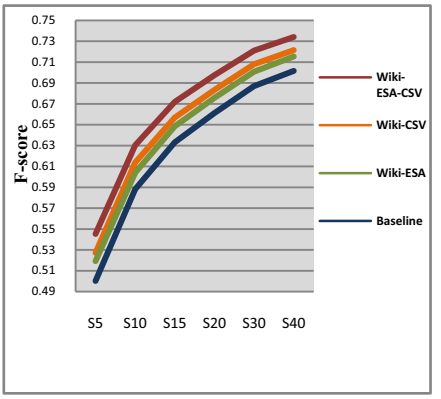


Fig. 7. The result of length 4

Table 5 shows newly discovered semantic relationships where linking concepts can only be acquired by integrating information from multiple documents or from Wikipedia knowledge (i.e., not contained in the existing document collection). For instance, for the query pair: "Atta :: dekkers," two intermediate important persons connecting them were identified: "Mohammed_Atta_al_Sayed" was an Egyptian hijacker and one of the ringleaders of the September 11 attacks and "Marwan_al-Shehhi" was the hijacker-pilot of United Airlines Flight 175, crashing the plane into the South Tower of the World Trade Center as part of the September 11 attacks.

Table 5. Instances of Enriched Semantic Relationships

Query Pair	Resulting Chain
L2 (Length 2)	
abdel_rahman :: blind_sheikh	abdel_rahman →ballistic_missile_threat_unite_state → blind_sheikh
bush :: bin_ladin	bush → east_africa_embassy_bombings →bin_ladin
adel :: ffi	adel → afghanistan → ffi
marty_miller :: oakley	marty_miller → unocal → oakley
gore :: stephen_hadley	gore → clarke → stephen_hadley
alexis :: lloyd_salvetti	alexis → janice_kephart_roberts → lloyd_salvetti
donovan :: wall_street	donovan → intelligence_group → wall_street
L3 (Length 3)	
atta :: dekkers	atta → mohammed_atta_al_sayed → marwan_al-shehhi → dekkers
amal :: sudanese	amal → bahrain → cia → sudanese
karachi :: usama_asmurai	karachi → june_14_terrorist_attack_outside_us_consulate_in_karachi → may_8_bus_attack_in_karachi → usama_asmurai
binalshibh :: pistole	binalshibh → fbi → minneapolis → pistole
martha_stewart :: saudi_arabia	martha_stewart → al-jawf,_saudi_arabia → khaled_of_saudi_arabia → saudi_arabia
L4 (Length 4)	
kenya :: mohamed	kenya → mihdhar_hazmi → afghanistan → shanksville → mohamed
gore :: stephen_hadley	gore → suicide_hijackings → white house → national_security_council → stephen_hadley
crawford :: khalilzad	crawford → bill_clinton → afghan → deputy_secretary_state_richard_armitage → khalilzad

6 Conclusion and Future Work

This paper proposes a new solution for mining semantic relationships between concepts across multiple documents by taking extensive background knowledge from Wikipedia into consideration. Specifically, we focus on detecting cross-document semantic relationships between concepts where most of them cannot be uncovered by the traditional paradigm. We also go one step further by incorporating the knowledge from Wikipedia to help identify more potential relationships that do not occur literally in the existing document corpus. The experiments were conducted using a large set of queries covering various scenarios, and compared with a purely BOW-based representation model, the original ESA method, and the approach only incorporating the Wikipedia category information. The results demonstrate the effectiveness of our proposed new hybrid solution combing all valuable resources and show the much broader and well-rounded coverage of significant relationships between concepts.

Wikipedia provides some other valuable information resources which were not used in this study. For instance, each Wikipedia article contains plenty of anchor text links which may imply potential relationships between different articles. Also, the "redirect" links pointing to a specific article may indicate synonymy and be further helpful to semantic

relatedness computing. Moreover, the infobox templates provide a good chance to increase the data quality using the ontology mapping technique. We will be exploring the usage of these resources and evaluating their performance in our future work.

References

1. Bollegara, D., Matsuo, Y., Isizuka, M.: Measuring Semantic Similarity between Words Using Web Search Engines. In: 16th International World Wide Web Conference, pp. 757–766. ACM, New York (2007)
2. Gabrilovich, E., Markovitch, S.: Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In: 20th International Joint Conference on Artificial Intelligence, pp. 1606–1611. Morgan Kaufmann, San Francisco (2007)
3. Gabrilovich, E., Markovitch, S.: Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge. In: 21st National Conference on Artificial Intelligence, vol. 2, pp. 1301–1306. AAAI Press, Menlo Park (2006)
4. Hotho, A., Staab, S., Stumme, G.: Wordnet improves Text Document Clustering. In: SIGIR 2003 Semantic Web Workshop, pp. 541–544. Citeseer (2003)
5. Jin, W., Srihari, R.: Knowledge Discovery across Documents through Concept Chain Queries. In: 6th IEEE International Conference on Data Mining Workshops, pp. 448–452. IEEE Computer Society, Washington (2006)
6. Martin, P.A.: Correction and Extension of WordNet 1.7. In: Ganter, B., de Moor, A., Lex, W. (eds.) ICCS 2003. LNCS, vol. 2746, pp. 160–173. Springer, Heidelberg (2003)
7. Milne, D.: Computing Semantic Relatedness using Wikipedia Link Structure. In: The New Zealand Computer Science Research Student Conference. Hamilton, New Zealand (2007)
8. MWDumper Software, `http://www.mediawiki.org/wiki/Manual:MWDumper`
9. Salahli, M.A.: An Approach for Measuring Semantic Relatedness between Words via Related Terms. Journal of Mathematical and Computational Applications 14(1), 55–63 (2009)
10. Scott, S., Matwin, S.: Text Classification Using WordNet Hypernyms. In: Workshop on Usage of WordNet in Natural Language Processing Systems, pp. 45–52. Association for Computational Linguistics (1998)
11. Srihari, R.K., Li, W., Niu, C., Cornell, T.: InfoXtract: A Customizable Intermediate Level Information Extraction Engine. In: HLT-NAACL 2003 Workshop on Software Engineering and Architecture of Language Technology Systems, vol. 8, pp. 51–58. Association for Computational Linguistics, Stroudsburg (2003)
12. Jin, W., Srihari, R., Ho, H.H., Wu, X.: Improving Knowledge Discovery in Document Collections through Combining Text Retrieval and Link Analysis Techniques. In: Seventh IEEE International Conference on Data Mining, pp. 193–202. IEEE Computer Society, Washington (2007)
13. Srinivasan, P.: Text Mining: Generating hypotheses from Medline. Journal of the American Society for Information Science and Technology 55(5), 396–413 (2004)
14. Swason, D.R., Smalheiser, N.R.: Implicit Text Linkage between Medline Records: Using Arrowsmith as an Aid to Scientific Discovery. Library Trends 48(1), 48–59 (1999)
15. Srihari, R.K., Lamkhede, S., Bhasin, A.: Unapparent Information Revelation: A Concept Chain Graph Approach. In: 14th ACM International Conference on Information and Knowledge Management, pp. 329–330. ACM, New York (2005)
16. Yan, P., Jin, W.: Improving Cross-Document Knowledge Discovery Using Explicit Semantic Analysis. In: Cuzzocrea, A., Dayal, U. (eds.) DaWaK 2012. LNCS, vol. 7448, pp. 378–389. Springer, Heidelberg (2012)

Estimating Risk Management in Software Engineering Projects

Jaime Santos[1] and Orlando Belo[2]

[1] ISCTE/IUL, Portugal
[2] Algoritmi R&D Centre, University of Minho, Portugal

Abstract. Independently from the nature of a project, process management variables like cost, quality, schedule, and scope are critical decision factors for a good and successful execution of a project. In software engineering, project planning and execution are highly influenced by the creative nature of all the individuals involved with the project. Thus, managing the risks of different project stages is a key task with extreme importance for project managers (and sponsors) that should be focused on control and monitoring effectively the referred variables, as well as all the others concerned with their context. In this work, we used a small "cocktail" of data mining techniques and methods to explore potential correlations and influences contained in some of the most relevant parameters related to experience, complexity, organization maturity and project innovation in Software Engineering, developing in a model that could be deployed in any project management process, assisting project managers in planning and monitoring the state of one project (or program) under its supervision.

Keywords: Software Engineering, Project Management, Data Mining, Effort Estimation, Risk Management, and Project Success Classification.

1 Introduction

The lack of success has been a generic characteristic whenever they are related to new developments or just simple enhancements in information technology projects, particularly the ones related to software engineering. Is common in the majority of the projects, delivered all over the world, to closes affecting negatively one (or more) of the main project vectors: cost, duration, quality or scope**.** Several cases presented in studies and surveys, like the ones from KPMG in 1997 [1], the Standish Group in 1995 [2], or more recently in 2009, the ratios in what we call unsuccessful are very high. In a 2009 published survey, the Standish Group concludes that just 32% of projects ended within costs, time and delivering all functionalities required. We can accept that this kind of surveys generates some controversy, but the overall conclusions are always the same: there are higher rates of cancelled projects, over budgeting, and schedule failover.

The scope of this article is focused in project estimation process, since it yields some of the most important activities in project lifecycle, but normally, with low efficiency and highly neglected, being performed based on feeling, gusts or some other political factors. Since estimating should be based on a process, with quality

P. Perner (Ed.): ICDM 2013, LNAI 7987, pp. 85–98, 2013.

standards and a time consuming on benchmark analysis of the organization and market data, we easily understand why some project managers and there organizations neglected the process. Basically, this happens because the initial estimative represents one investment without consequent returns (e.g. proposals preparation), leading the organizations or IT departments to follow simplified procedures or eventually, skipped them, even when this is a subject highly referenced in project management methodologies, like *Project Management Institute* (PMI). PMI emphasizes this procedure as a main component to calculate cost, duration, and their relations to other components, like risk management [3]. It is important that the organizations introduce new procedures and models that are able to improve and facilitate the estimating process.

This paper presents and discusses a data mining application process addressing the effort estimating activities on software engineering projects, with the objective to reach a project classification model and a project effort estimation model. This paper is organized into 3 more sections, namely: section 2 that exposes some relevant issues in project management activities; section 3 that presents and discusses the entire data mining process carried out; and finally, section 4 that presents some final remarks and conclusions, as well as a few future research lines.

2 Augmenting Effectiveness on Project Management

Having the ability to capture information, predict the uncertainty, estimate dimension and there eventual impacts, planning all activities, time and resources and then, monitoring and controlling accordingly, are the most important tasks to manage a project aiming its success. To accomplish this task, the manager should have practice and knowledge in several relevant domains like: planning, risk management, relationship management and communications, giving him the ability to plan in an accurate manner, capture the project situation and then acting proactively to take corrective actions and mitigation plans. In order to monitor and control the project it's necessary to make some estimation. Usually, the first one happens at the planning stage, so during the execution phase it could be compared with reality and then, redoing or adjusting accordingly to the current situation of the project. This task is characterized by understanding and contextualizing the project scope and their characteristics as better as we can, producing a first estimation of effort, resources, duration, cost, defects, documents, and so on. Some other task that we highlight is the ability of the project manager and his team to capture and manage the scope.

The scope volatility related to a software engineering project, follows contours of higher complexity than those characterized by a repetitive nature. Thus, their unique and non-repeatable nature, along with the team and the project sponsors creativity, are major challenges and a serious risk in the project execution, since incrementing the scope will directly impact the other project vectors: cost, quality and time. The risk of error in the estimation process, the risk management framework and its amendments, or the unpredictability of actions for each stakeholder, has a direct impact on the cost, on the quality and on the time of the project. So, during a software implementation with a high degree of risk regarding is intangible and creative nature, along the fact that good governance is characterized by a proper risk management, with a tight control of variables, it is important to use concepts and methods for data collection

and retention on analytical processing systems, making the data available to apply data mining techniques, designed to develop models for estimating and forecasting, assisting in the planning and in the risk assessment. Data mining allows us to deep our knowledge about project management, helping us to extract behavioural evidences from historical data, while understanding relations between them. It is recognized that some characteristics affects the team productivity, so new knowledge will bring direct benefits in the estimation process, assisting us on understanding the impacts on productivity as well in the detection and quantification of risks. The acquisition of new knowledge, or the simple confirmation of some ideas taken for granted, can bring to us clear benefits to improve estimates or predict future events, enhancing the management of the inherent risk and the uncertainty present on those type of projects.

3 Mining Project Management Data

3.1 Overview

Project management is a highly complex activity that pushes several techniques and knowledge, focusing on extracting information from several objectives, deliverables and surrounded features, occurring on environments of constant uncertainty, being a continuous task of checking and acting. As already referred, this paper presents and discusses a data mining application process addressing the effort estimating on software engineering projects and also, methods to explore potential correlations and influences contained in some of the most relevant parameters related to experience, complexity, organization maturity or project innovation. Those methods could then be deployed in any project management process, assisting managers in planning and monitoring the state of one project or program under is supervision. Trying to assure a systematic approach for our work, we adopted and followed the CRISP-DM methodology [4]. First, we performed a research about the business related to the project management activities and to software engineering, attempting to understand the most important features, the environmental complexity and, trying to identify the most influencing characteristics on the project events and project risks across all phases, but specially, at planning and execution phases.

3.2 Data Sets, Acquisition and Preparation

We start data acquisition and preparation tasks extracting data from all the projects we considered with relevance to this study. Therefore, from 24000 projects available on our database, we selected only those whose purpose was related to application development or major enhancement applications development. It was further selected only those on a completed state, approved and available for metrics analysis. Thus, from the initial universe we extracted approximately 5000 projects. After this first step, we proceeded adding some more variables that arise from the junction with other tables presented in the database, such as type of industry, indicators of complexity, experience and project context, sizing, resources, etc. Later, other variables were incorporated, which resulted from the aggregation and transformation processes using some of the initial variables (totals for estimated values, indicators of failure, etc.). We explore the different variables, and by doing that, we detect that some of those (quantitative) variables did not show a normal distribution (figure 1). Thus, we opt for

the logarithm transformation. Three types of sizing were used on the projects, Function Points, that consists in a certified methodology for sizing application development, Lines of Code, representing the total number of lines coded to develop the application and "Others". Since "Others" are very diffuse, not quite understandable and not comparable, we decided to discard all projects having size calculated only with this type, resulting in the final dataset with approximate 4000 projects. We also detected missing values in several variables for some projects, what imposed several treatment acts.

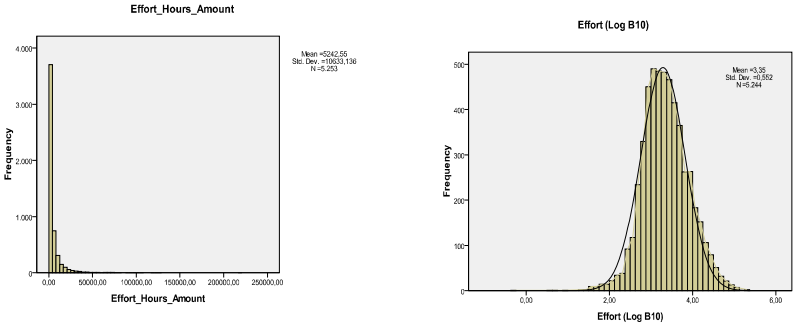


Fig. 1. Effort_Hours_Amount / *L*ogarithm histogram

The data set used contains a very broad representation in terms of geographic, technical and industry characterization. There are projects from more than 30 different countries, with major focuses in North America, followed by Australia and Europe (figure 2). The data has a hide industry representation (figure 3), from manufacturing, transport, energy, finance and even government entities. However, we can assist to the prevalence of manufacturing and finance. We also noticed that most projects were managed using some project management methodology, however, almost 20% did not use any formal methodology.

Country Region

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid		1213	20,6	20,6	20,6
	Australasia	791	13,4	13,4	34,0
	Central America	48	,8	,8	34,9
	Eastern Asia	17	,3	,3	35,1
	Eastern Europe	3	,1	,1	35,2
	North America	2786	47,3	47,3	82,5
	Northern Africa	10	,2	,2	82,7
	Northern Europe	429	7,3	7,3	90,0
	South America	158	2,7	2,7	92,7
	Southeast Asia	37	,6	,6	93,3
	Southern Africa	3	,1	,1	93,3
	Southern Asia	70	1,2	1,2	94,5
	Southern Europe	58	1,0	1,0	95,5
	Western Asia	2	,0	,0	95,5
	Western Europe	262	4,5	4,5	100,0
	Total	5887	100,0	100,0	

Fig. 2. Country Region frequency List

Industry Type

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Comms, Media & Entertainment	490	8,3	9,0	9,0
	Consumer Industries & Retail	31	,5	,6	9,6
	Internal	478	8,1	8,8	18,3
	Energy	14	,2	,3	18,6
	Financial Services	1151	19,6	21,1	39,7
	Government	471	8,0	8,6	48,3
	Healthcare	251	4,3	4,6	52,9
	Manufacturing	2134	36,2	39,1	92,1
	Multiple Industries	1	,0	,0	92,1
	Transportation	432	7,3	7,9	100,0
	Total	5453	92,6	100,0	
Missing	System	434	7,4		
Total		5887	100,0		

Fig. 3. Industry Type frequency List

During the preparation phase, and to better understand our data, we also explore some correlations between different variables; however, we didn't reach any significant correlation. We expected an immediate identification between sizing and effort, but ultimately the data showed very weak correlations, i.e., with Pearson correlation coefficient of 0.15 in relation to the functional size by 'function points' and 0.100 for the size in 'lines of code'. A wide representation of different programming languages can explain this fact. As is known, the relationship between language and effort is large, so, it is difficult to find correlations standing before such representative data. This is indeed the problem associated to the estimation, i.e. the existence of large dispersion and great amplitude in the factors that affect the productivity.

In a second attempt to demystify the foregoing, it was decided to perform an analysis on the correlation between effort and sizing for two types of programming language, having reached to values of significant Pearson correlations of 0,577 and 0, 564. Notwithstanding the foregoing, there are some interesting correlations detected, but not surprising, such as: strong correlation between the three variables of complexity classifiers - application innovation, technological innovation and BUS innovation; and also a strong correlation between some of the variables that classifies experience - computer experience, tools, language, methods and technology experience. There were a correlation of 0.988 between "FTE Amount" and "Effort Hours Amount". Being FTE (Full Time Equivalent) an expression used in business to summarize the total of man/months, then the correlation is completely acceptable. This conclusion allows us to reduce the dimensionality and complexity of the analysis by removing the variable "FTE Amount", not be considered in any following steps, in particular, at the mining tasks. It is more surprising to note the clear existence of a correlation between client complexity and the team complexity, holding a correlation of 0.388. This can show us that client as interference in how a manager constitutes his team, whether it is directly or inadvertently. Regardless the fact that we have not identified major surprises in the correlations, with the presence of very low rates, it is however possible to see differences in effort, for example, in the ratios between effort and sizing. To this end, we subdivide the data into three group types:

1) One group which sizing data was calculated according to FPA methodology.
2) Another group which sizing data was calculated according to methodology of lines of code.
3) A final group which sizing data was calculated using both methods.

Thus, we proceeded for an average comparison, according to a diverse set of deterministic variables. The OneWay ANOVA [5] method was applied to compare the means, separating each sample according to the experience, the complexity, the innovation and the maturity classifiers (e.g. table 1). The different populations were then defined according to a pre-existing data characterizing the level of each project.

Table 1. Populations characterized by experience in: project management, system, tools, programming language, methods, etc.

Code	Description
1	*Less Than 1 Year*
2	*1 - 3 Years*
3	*Greater Than 3 Years*

Before we go forward with the comparisons, there were some important preconditions to be verified for the feasibility of the chosen method. The first condition is to assure that the test variable is quantitative, which in our case the condition were guaranteed. Second, there must be a variable that defines the nominal groups. In our cases, all variables used are nominal and we can use the average function for each of its dimensions, thereby ensuring the suitability. In addition to the conditions set out above, it is still assumed that the variable under test follows a normal distribution, which was not the case. To be possible to go further, we decide to calculate a logarithm base 10 for both ratios, yielding so the tendency for the normal distribution, the desired condition for this test. Finally, to apply the OneWay ANOVA method is also assumed that there is an equality of variances in the different populations for the variable under study. For evaluating that condition we've performed a test of homogeneity of variances, which were defined by the following assumptions:

- H0: Variances are equal.
- H1: Some of the variance differs from other.
- $\alpha = 0.05$ (Alpha definition for the rejection of the null hypothesis).

Looking at the Table 2 we reject the null hypothesis, i.e. that there isn't equality of variances, since the value of Sigma is below the alpha (α).

Table 2. Test for the Homogeneity of variance

	Levene Statistic	df1	df2	Sig.
Ratio Effort by FP(LogB10)	9,894	6	2471	,000

Given this situation, generally occurred in all tests with the populations used, we were forced to abandon the ANOVA test, being resorted to a more robust equality of means test, which is the test of Brown-Forsythe and Welch [6].

Table 3. Robust Tests of Equality of Means[b]

		Statistic[a]	df1	df2	Sig.
Ratio Effort by FP(LogB10)	Welch	36,078	6	313,308	,000
	Brown-Forsythe	27,709	6	239,232	,000

a) Asymptotically F distributed.

Since in all cases our sigma (which indicates the variability) was zero, then we reject the null hypothesis, i.e. there is no equality in the average ratios. Thus, in terms of overall conclusion, the presentation of different mean values between different populations, leads us to conclude that the variables under study interfere effectively in productivity, and in turn, the effort required to produce one function point or one line of code, so, they had been considered very important variables to use in testing techniques, which would lead to the classification and estimation. As an example, in Figure 4 we can see that the mean has a tendency to increase as the complexity of the team increases, so the more complex the team is, the lower the productivity index by lines of code or by Function Points.

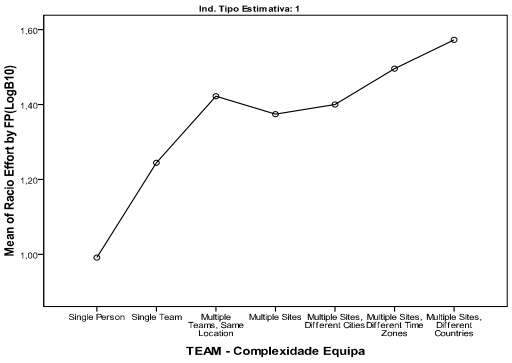


Fig. 4. Means comparison Chart: Team Complexity

Looking the variables related to innovation, the findings are as expected, i.e., there is a clear increase in the ratios as there is a higher rate of innovation in the project. However, the variable related to the application innovation presented a quite interesting behaviour, since the tendency is reversed to the initial expected, which means that the ratio decreases when the application innovation increases, so, less effort is need to produce the same size in a more innovative application. However, it's normally expected that teams struggle with some problems in innovative projects, negatively affecting their productivity. Perhaps this just occurs under the influence of programming languages, which modernization allows teams to deliver more functionality in less time! The trend found at the level of application innovation was also detected in the variables

characterizing the experience, no matter if occurs in the teams, about their system experience, language experience, or even in the project manager experience managing projects. Thus, this is a very important result, because it doesn't confirms the general knowledge that the experience as a positive impact in the productivity. What is certain is the fact that this conclusion may be itself as one of the most important in this work, since the normal thought is contrary to the one found here.

When we look to the ratios (Effort by FP or effort by LOC), separating the populations by organizational maturity (CMMI rating), we denote the benefit that a company can get by moving from a non-documented and possibly disorganized level (CMMI Level 1) to one with organizational evidences with use of standards. However, as the organization moves up the maturity, the impact in terms of productivity is achieved in a negative way. This may be caused by the existence of more bureaucracy, higher-level documentation. So, projects in organizations with higher maturity level turn out to be impacted by the amount and complexity of documents that have to bear, as well as the procedures for review and audit they are subjected.

3.3 Classification and Estimation Efforts

Considering that the target variable (Check_Sucess) was categorical binary and the fact that we are facing a world of mostly categorical or nominal data, we chose to apply two classification techniques: the C&RT classification and regression trees [7], and the C5.0 decision tree [8]. The initial universe of projects for the classification process integrated 3644 projects instances (after delete some cases to achieve balanced data set), containing a perfectly balanced data set of cases in which resulted in success or failure, remaining 1822 projects classified as failure instances. In order to ensure higher quality testing techniques, the data set was divided into three subsets, having respectively 35% for training data, 35% for test data, and 30% for validation data. Regarding the validation and for model quality evaluation purposes, we choose the misclassification error rate method [8]. In order to be able to perform the mining process, specifically the classification task, it was important to pre-characterize the target variable as to success or failure (see table 4).

Table 4. Previously classification of: Success vs. Failure

Pre-classify	Value
Success	1
Failure	0

With C&RT, the classification process was run in two modes: simple and advanced. In both cases, the results were the same, with a good ability to hit the projects targeted as success but with a high cost of misclassified cases for the ones considered with failure. Given the nature of the business, and expecting that this model help managers to anticipate risk situations in their projects, it is preferable a model that presents the best ability to classify failure to the detriment of those who had success. From the pre-classified cases of failure, the resulting output from this technique classified 1118 as success, corresponding to 61% of misclassification.

Regarding the poor results presented by this technique, it was no longer taken into account, not been used for any comparison step with other techniques. Next, we chose to train C5.0 decision tree using two modes, simple and advanced (as we did with C&RT). With this technique the situation was quite different, since there had been improvements in the classification rate, with the simple method achieving rates of 58% in the classification of cases of failure and 70% of success, but we continue to consider ineffective and with unsatisfactory results for the most important cases, the ones classified as failure. Alternatively, and after several training sessions, the execution of the advanced mode were performed with the option of 'pruning severity' equal to 50, the 'boosting' option enabled for a number of five attempts and the 'cross-validation' option also activated for a total of five folders with a minimum number of records by node of five. The advanced mode had a better ratio of good classifications, with a percentage of 65% accuracy on projects previously classified as failure, as we see on figure 5.

		$C-check_Sucesso	
check_Sucesso		0	1
0	Count	1176	646
	Row %	64.544	35.456
	Column %	65.116	35.147
1	Count	630	1192
	Row %	34.577	65.423
	Column %	34.884	64.853

Fig. 5. C5.0 Advance Mode: Classification Matrix

For the estimating task, looking to estimate the project effort, we used multiple regression [9], neuronal networks [10] and CHAID decision trees [11]. In the first trainings performed we used the complete data set, having projects whose sizing was calculated in 'functions points' or 'lines of code'. Since the results did not show a minimum quality required, we proceeded to split into two subsets, according to the method of sizing and due to time constraints it was decided to perform this task only for the universe of projects with the calculation of 'function points'. By using as first method a multiple regression, we just intended to verify if we can achieve some improvements in the final results, comparing it to more advanced techniques, such as neuronal networks.

Comparing $E-Effort_Hours_Amount with Effort_Hours_Amount

'Partition'	1 Training	2 Testing	3 Validation
Minimum Error	-31458,412	-60132,131	-79219,041
Maximum Error	71728,521	78644,846	60957,175
Mean Error	162,833	-301,841	-459,226
Mean Absolute Error	4696,258	4777,64	5065,872
Standard Deviation	8221,961	8908,419	9730,993
Linear Correlation	0,597	0,544	0,423
Occurrences	799	834	764

Fig. 6. Multiple regression – evaluation

To execute this technique, as the subsequent, the training was performed using the base variables, without any treatment, and then we always repeat the training using the size variables converted, in our case, to their logarithms or resulted from the min-max standardization method intending to achieve a normal distribution (since all

quantitative variables had a left skewed distribution). The multiple regression technique demonstrated to be incapable to result in any model when executed with the variables converted.

With neuronal networks we performed several training sessions, using several execution modes, several layers and neurons, presenting above the two with the best results obtained. In these two cases, the neuronal network was executed with the prune method [12], one with a simple mode and the other in advanced mode, using two layers, the first with three neurons and the second layer with seven neurons – Fig. 7 presents the most important variables used to estimate the effort. This figure has been extracted from the one reached with prune advance mode and it result in a simple model, with less number of variables, which is quite important for implementation purposes.

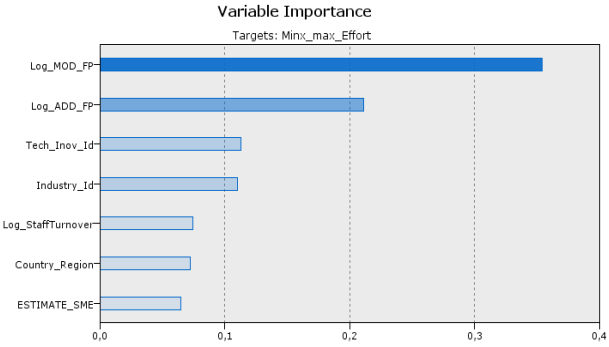


Fig. 7. Neuronal network with prune advanced mode - variable importance

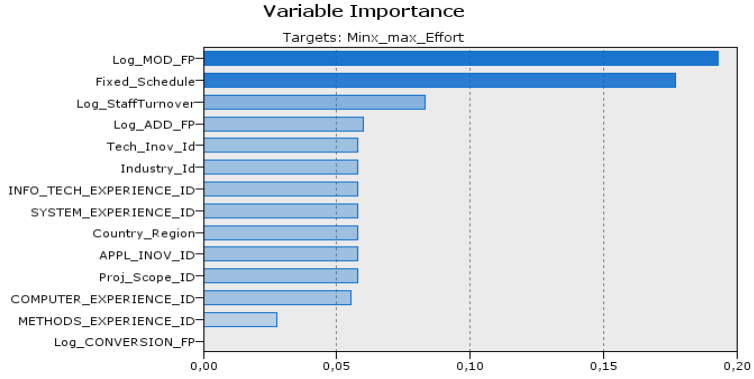


Fig. 8. CHAID decision tree - variable importance

Finally, we generated a decision tree with CHAID. It is possible to see that the values of 'sizing' hold a central importance, with CHAID method capturing some of the project context variables, those that could cause some variability in the productivity rates (figure 8).

3.4 Results Evaluation

Using the results obtained in each model, we done a small comparison process (Table 5), which allows us to evaluate the results according three indicators: overall error rate (OER), false positive rate (FPR), and false negative rate (FNR).

Table 5. Classification Task - results comparison

	OER	FPR	FNR
Decision Tree C5.0 – Simple	35,48%	29,75%	41,22%
Decision Tree C5.0 – Advanced	35,02%	34,58%	35,46%

We selected the most important variables in each model, and although there is a correspondence between models in order to the importance of the variables, we observe that: both models presented the 'client_participation', 'fixed_staff', and 'industry_id' as variables with explanatory power for failure; and the two models agree on the importance of variables related to experience, complexity and constraints of the project ('fixed team', 'fixed cost', etc.).

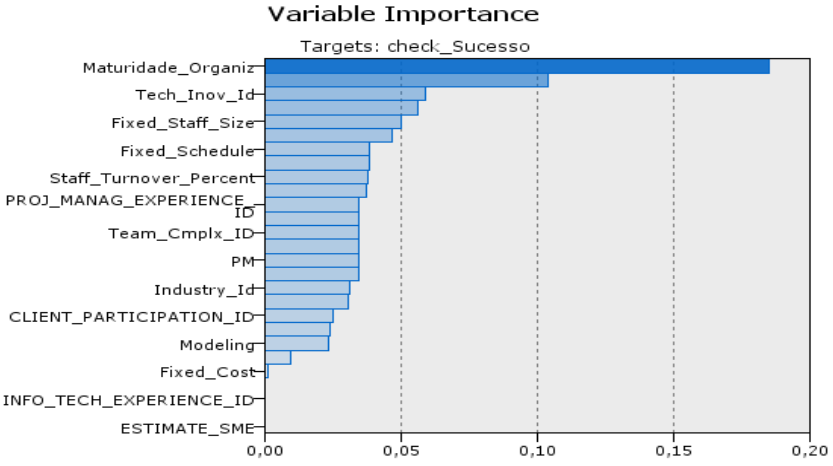


Fig. 9. C5.0 Simple - variable importance

We also observed that the tree generated by the algorithm C5.0, in advanced mode, has not only a lower overall error rate (35.02%) as a lower rate of error in the prediction of false negatives, that is, projects with failure misclassified as a success. Although the results did not reach the expected values, a good classification rate around 65% will allow the model to be implemented as a risk management tool.

In terms of the estimation tasks, we noticed the evident efficiency from all techniques comparing the results with the ones of multiple regression (Table 6) - all values were reached after applying the *min-max* normalization on the results extracted from the evaluative matrix.

Table 6. Estimation task - results comparison

	Minimum error	Maximum error	Average error	Mean Absolute Error	Standard Deviation
Multiple regression	-0,896*	0,670*	-0,015*	0,045*	0,097*
Neuronal Network – Prune Simple	-0,327	0,641	0,011	0,044	0,074
Neuronal Network– Prune Advanced	-0,358	0,639	0,003	0,051	0,090
Decision Tree – C&RT	-0,391	0,661	0,001	0,057	0,103
Decision Tree – CHAID	-0,320	0,658	-0,001	0,055	0,098

Looking at figures 7 and 8, we see that the different models agree about the variables regarded as the most explanatory. Multiple regression and C&RT decision tree are those with a simpler model, having concordance to the others in relation to the sizing variables, only. All models also feature that the sizing variables has high explanatory power for estimate effort, and there was agreement on the importance of the 'Staff Turnover' variable. There is also a widespread agreement among the various algorithms regarding the importance of variables related to experience, complexity and project constraints (team fixed and fixed cost, etc.). However, this agreement is substantially higher between the CHAID decision tree and neuronal network.

Finally, we can verify that a multiple regression showed good results in the estimation. However, it does not contain the variability resulting from the context in which a project is related. The same goes for the C&RT decision tree. It should be noted that the challenge presented in the work intend to capture this variability and thus, adjusting the estimate to be more effective, since the 20% deviations normally assumed, contains a high impact on the project's cost and the expected success. Thus, taking into account the variability of the above-mentioned issue and the results, we think that is appropriate to implement the model obtained by the decision tree CHAID.

4 Conclusions and Future Work

In this paper we presented two models that can be used in the software engineering projects management. The preliminary study of the source database and the data collection process resulted in one of the most challenging components and consuming effort of this work. The fact of having to use a relational database, with more than 200 tables, over any data mart or pre-prepared file, cause that this task became very time and resource consuming. However, this was a very important phase, because it allowed us to delve a little deeper on the business of managing software engineering projects, but mainly, it makes possible the data understanding, enabling the choice of tasks to perform, launching new challenges to the future.

This work helped us to detect within the database some important information to use in a mining process, but it was also detected some gaps and needs that should be addressed in a near future. Considering the available data, the source contains relevant information for the execution of any data mining task, not meaning that the database can't be further enriched, e.g., with detailed information about each stakeholder, like indicators of attitude, resistance to the project, level of communication, among others. The presence of many projects without the minimum quality for analysis was the major problem identified in the database, cases which have been considered by the database internal auditors. These situations were reflected in almost half of the initial data universe, and it is important to define actions aimed to improve the quality of the data. Another challenge that was put during the execution of the mining process, was due to the existence of a multitude of different programming languages, which direct influence the effort and cause data dispersion and a large deviation, so that, a segmentation task can bring clear benefits to future analytical works. It is important to note that was detected in the database some additional information related to documentation type and quantity produced, as well data about changes in the several project deliverables, which can make possible to conduct new mining processes, in particular, to estimate total number of pages of documentation or association tasks related to changes in the project. One of the most significant trends found, were, at the level of the innovation and the experience variables, for team and project manager. It was expected a trend towards an increase in productivity as well as experience increased, however this happens in reverse. We think that this occurs mainly because experience will make more positive impact in overall quality of the project (all deliverables: the software, the documentation, etc.) then it does to the productivity.

The paper demonstrates the complexity that involves a software engineering project, and although from a long time the sizing values allow us to estimate effort with a certain degree of confidence, this is not enough. On the one hand, an overvalued estimate puts at risk the victory in a competition for a project, as an undervalued estimate, causing 10% or more deviations in costs; it will direct impact on the organizations and their viability. Looking to the resulting models achieved by the tasks performed in this work, we think that the estimation model created can be implemented in various software engineering projects as an alternative tool to the techniques and methods commonly used, which representative spectrum confers a generic capabilities, while the results given a confidence in their applicability. Despite the total error rate of 35%, we think that the resulting model from the classification task can be incorporated into risk management procedures of any software engineering project, since the early detection of a disaster will allow making on time decisions and the necessary corrective actions. So, implementing this model into risk analysis, at the planning stage as well during project execution phase, will enable "what if" scenarios execution and test, enabling the manager to measure and validate several alternatives for correction or improvement, understanding how we can increase the chances of success. Another interesting mining process that can be done should aim a model resulted from stakeholder's segmentation or classification task. This could provide tools to the managers that allow him to manage each one at the most appropriate way, taking preventive actions that help to minimize impacts. This example is something that could be implemented in areas like communications management.

References

1. Whittaker, B.: What Went Wrong: Unsuccessful Information Technology Projects. KPMG Consulting, Toronto (1997)
2. Standish Group.: The Standish Group Report: Chaos (1995), http://www.projectsmart.co.uk/docs/chaos-report.pdf (acedido em January 17, 2011)
3. PMI.: A Guide to the Project Management Body of Knowledge: PMBOK Guide, 4th edn. Project Management Institute, Newton Square (2009)
4. Chapman, P., et al.: CRISP-DM 1.0: Step-by-step data mining guide. The CRISP-DM Consortium (2000)
5. Looney, S.: Biostatistical Methods, vol. 184. Humana Press, University of Louisville School of Medicine, Kentucky (2002)
6. Almeida, A., et al.: Modificações e alternativas aos testes de Levene e de Brown e Forsythe para igualdade de variâncias e médias. Revista Colombiana de Estatística 31, 241–260
7. Breiman, L., et al.: Classification and Regression Trees. Wadsworth, Belmont (1984)
8. Larose, D.T.: Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley & Sons, Inc., New Jersey (2005)
9. Cohen, J., et al.: Applied multiple regression/correlation analysis for the behavioral sciences, 2nd edn. Lawrence Erlbaum Associates, Hillsdale (2003)
10. Ripley, B.D.: Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge (1996)
11. Kass, G.V.: An Exploratory Technique for Investigating Large Quantities of Categorical Data. Applied Statistics 29(2), 119–127 (1980)
12. Cantú-Paz, E.: Prunnnig Neuronal Networks with distribution estimation algorithms. Center for Applied Scientific Computing. Lawrence Livermore National Laboratory, Livermore (2003)

Wastewater Treatment Plant Performance Prediction with Support Vector Machines

Daniel Ribeiro[1], António Sanfins[2], and Orlando Belo[1]

[1] ALGORITMI R&D Centre, University of Minho, Portugal
[2] IDITE-Minho, Portugal

Abstract. Wastewater treatment plants are essential infrastructures to maintain the environmental balance of the regions where they were installed. The dynamic and complex wastewater treatment procedure must be handled efficiently to ensure good quality effluents. This paper presents a research and development work implemented to predict the performance of a wastewater treatment plant located in the northern Portugal, serving a population of about 45,000 inhabitants. The data we used were recorded based on the daily averaged values of the measured parameters during the period of one year. The predictive models were developed supported by two implementations of Support Vector Machines methods for regression, due to the presence of two lines of treatment in the selected case of study, using two of the most relevant output parameters of a wastewater treatment plant: the biochemical oxygen demand and the total suspended solids. We describe here the wastewater treatment plant we studied as well the data sets used in the mining processes, analyzing and comparing the regression models for both predictive parameters that were selected.

Keywords: Data Mining, Regression Techniques, Wastewater Treatment Plants, Support Vector Machines, Biochemical Oxygen Demand and Total Suspended Solids Analysis.

1 Introduction

Wastewater Treatment Plants (WWTPs) are infrastructures that treat domestic and industrial wastewater with the goal of protecting public health in a commensurate manner with environmental concerns [27]. It's quite important that there exist a suitable treatment in order to avoid discharges highly polluting and therefore outside the limits regulated by law. Predicting the quality of treated water give us the possibility to measure the effectiveness of treatment and thereby obtain useful information for a better control of the entire infrastructure. However, WWTP are dynamic and complex systems, since they execute different types of treatments over wastewater, such as physical, biological and chemical. This suggests that advanced *data mining* (DM) techniques could be applied especially the ones related to non-linear predictive models, in order to successfully model the behaviour of a WWTP. Usually, in this kind of studies, there are selected some outflow variables for prediction, such as: *chemical oxygen demand* (COD), *biochemical oxygen demand* (BOD), *total suspended solids* (TSS) and nutrients (e.g. phosphorus and nitrogen).

P. Perner (Ed.): ICDM 2013, LNAI 7987, pp. 99–111, 2013.

Since these systems presents a complex and very dynamic behaviour, it's natural that non-linear prediction techniques such as *neural networks* (NN) be widely used on the field. Some other studies compared the efficacy of linear prediction models against NN models like the ones done by Gallop et al. (2004) [16], Hamed et al. (2004) [18] or Dixon et al. (2007) [13]. In its turn Belanche et al. (1999) [4] compared two NN types (classical, diffuse), while Luo et al. (2009) [26] compared the effectiveness of NN using different approaches of feature selection.

The application of decision and regression trees in predictive modelling tasks was also investigated in this area, not only in the design and implementation of regression models [3], but also in classification problems like the ones done by Atanasova and Kompare (2002a) [2] or by Cărbureanu (2010) [7]. More recently, some other studies emerged approaching techniques based on *support vector machines* (SVM) with the purpose to predict the quality of the final effluent of a WWTP. As in the case used [37], where it was selected SVM for regression in order to predict COD and TSS. Additionally, Yang et al. (2011) [40] applied a variant of SVM, called Least Square Support Vector Machine (LS-SVM) in the prediction of COD, where the obtained results were compared with the NN performance. On the other hand, Huang et al. (2009) [21] also uses the LS-SVM method to predict multiple wastewater quality parameters, whereas Hong et al. (2008) [19] in their study investigates the predictive ability of LS-SVM in classification problems.

In this study we explored several DM techniques, in particular predictive ones, in order to apply them modelling the behaviour of a WWTP. This work was based on a real case of a WWTP located in northern Portugal. In the modelling phase of the wastewater treatment system we used methods based on SVM, which were applied to predict the concentrations of two outflow parameters, namely BOD and TSS. It was intended to predict the concentrations values - knowing that these values are numerical we adopted regression methods particularly SVM for regression.

The remainder of this paper is organized as follows. In section 2, we present a brief introduction to SVM, followed by a complete description of the WWTP under study and the respective measured data (section 3). Next, in section 4, is presented the entire mining process, including all the predictive models that were designed and implemented. The paper ends with results analysis (section 5), and with the usual section of conclusions and future work (section 6).

2 Related Work

SVMs are grounded on statistical learning theory, also known as VC theory, which began to be developed in the late 1960s to the 1990s by Vapnik and Chervonenkis (1974) [36] and Vapnik (1995) [34], as cited by Smola and Schölkopf (2004) [33]. In the early 1990s, was proposed by Vapnik and its co-workers a new learning algorithm (SVM) based on VC theory. Initially, SVM emerged as a classification technique that was successfully applied in some research works, especially with special impact in pattern recognition e.g. [5, 9]. However, soon after SVM was also applied in regression problems, e.g. [14].

Basically the SVM idea consists of mapping an input vector x into a high dimensional feature space Z by a nonlinear mapping Φ [10]. In classification the problem lies in finding the best hyper-plane that separates the classes. A hyper-plane

is the optimal decision function that separates the two classes in order to maximize the margins between the support vectors, which are actually the classes on margins [9]. However when the training data are not separable, the rigid margins algorithm will find no feasible solution for the optimization problem. Therefore slack variables are introduced in the constraints allowing some classes to stay between the margins as well as some misclassifications, hence the name "the soft margin hyper-plane". A more detailed explanation of SVM for classification, including some formulations and mathematical proofs, could be found in [34], [9] or [6].

In regression problems SVM are known as *support vector regression* (SVR), having the associated idea of imagining a tube around a function line [33]. The purpose of the ε-SVR function referred in that work is to find the function that has no more than ε as the deviation over all labels from the training data and, at the same time, to get a tube as thin as possible. Similarly to the "soft margins" on SVM classification, SVR introduced slack variables in constraints, allowing for some errors in order to make the optimization problem feasible. The penalty parameter C does the trade-off between the thickness of the function and the amount up to which deviations larger than ε are tolerated. In figure 1 we can observe that for values between –ε e +ε (tube) there is no penalty. Only values outside the tube are penalized through linear loss function ε-insensitive ($|\xi|_\varepsilon$ - equation 1).

$$|\xi|_\varepsilon := \begin{cases} 0 & if \ \ |\xi| \leq \varepsilon \\ |\xi| - \varepsilon & otherwise \end{cases} \tag{1}$$

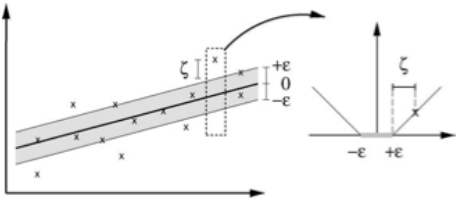


Fig. 1. Example of the ε-SVR function (left side) and the ε-insensitive loss function chart (right side) - adapted from [33]

A SVM optimization problem can be presented in dual formulation, particularly through the Lagrange function, where certain characteristics of SVM are exposed. For instance, the so-called "*Support Vector Expansion*", which demonstrates that the complexity of the function is independent of the training dataset dimensionality but instead depends only on the number of support vectors. Another feature is the sparsity of the SVM. According to the Karush-Kuhn-Tucker (KKT) conditions, the Lagrange multipliers (α_i , α_i^*) are positive only when the examples are outside the tube. This implies that the samples within the tube vanish. Therefore, only non-vanishing cases are used in the optimization problem. In fact, these are called support vectors. Since the SVM optimization problem is convex implying a solution with a unique global minimum, hence SVM are absent of the problem of local minima that affects some algorithms such e.g. NN. Another important piece is that the algorithm is described in terms of dot product, which is important for the formulation of the non-linear SVM

extension. The nonlinear SVR also use the concept of mapping Φ of an input x into the feature space Z (figure 1). Since the SVM algorithm depends solely on the dot products we have that $K(x , x') := \langle \Phi(x), \Phi(x') \rangle$, the function of nonlinear SVR is thus described as:

$$f(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) K(x_i , x) + b \quad (2)$$

The inclusion of the Kernel in SVM decision function provides a feature mentioned by Cortes and Vapnik (1995) [9] as the "Universal Machine", once several kernel functions can be used to calculate the dot product provided that they satisfy the Mercer's condition. The kernel functions most commonly adopted are the polynomial, Radial Basis Function (RBF) and sigmoidal. More details about the SVR algorithms can be found in [33], which have been summarized here in a very superficial way. The SVM variation introduced by Platt (1999) [29] is a learning algorithm called Sequential Minimal Optimization (SMO). Briefly SMO decomposes the overall quadratic programming (QP) problem into fixed-size QP sub-problems. The version of SMO for regression is also presented in [33].

3 The WWTP Case Study

The plant that served as our case study is located in northern Portugal serving a population of about 45,000 inhabitants. This plant performs the following processing stages: 1) a preliminary treatment where solid particles are removed, essentially coarse sand and fats; 2) the primary stage where suspended solids are removed; 3) the secondary phase, where it is treated mainly organic material, although are also reduced suspended solids; and, finally, 4) the tertiary stage where the nutrients are removed and pathogenic organisms disinfected.

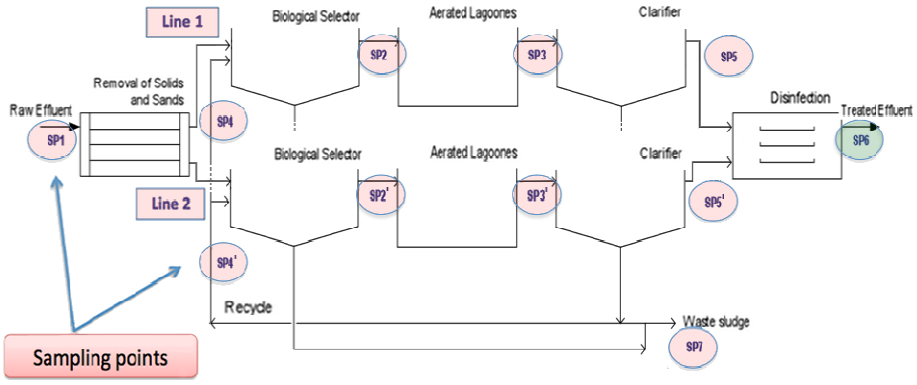


Fig. 2. A schematic diagram of the WWTP used as case study

Figure 2 shows the treatment process carried out as well as the points where the different treatment parameters are measured and gathered. After the entry of raw water in the plant (SP1) coarser matter is treated. Next, the treatment process

continues to be executed in two treatment parallel lines. Both treatments lines are processed in three main tanks, namely the biological selector, the aerated lagoon and the secondary clarifier. At the end of each treatment were collected samplings of (SP2, SP2'), (SP3, SP3') and (SP5, SP5') sampling points, respectively. Moreover, the samples from both recirculation lines (SP4, SP4') are taken. After the secondary treatment, the treatment lines converge to the final stage of tertiary treatment. After this point the treated effluent (SP6) is then measured. Parallel to this process it is also performed the treatment of the solid line, i.e. the sludge treatment (SP7).

Table 1. Treatment parameters and respective SPs gathered in the WWTP database

Parameter	Description	Sample Points
Q_{in}	Inflow rate	SP1
Q_r	Recirculation flow	SP4; SP4'
Q_p	Purge Flow	SP7
pH	Acidity or basicity of an aqueous solution	SP1; SP2; SP2'; SP3; SP3'; SP5; SP5'; SP6
P_{redox}	Redox potential	SP1; SP3; SP3'; SP4; SP4'
P_{redox_AZ}	Redox potential at aerobic zone	SP2; SP2'
P_{redox_AXZ}	Redox potential at anoxic zone	SP2; SP2'
O_2	Dissolved Oxygen	SP3
O_{2_AZ}	Dissolved Oxygen at aerobic zone	SP2; SP2'
O_{2_AXZ}	Dissolved Oxygen at anoxic zone	SP2; SP2'
COD	Chemical oxygen demand	SP1; SP5; SP5'; SP6
BOD	Biological oxygen demand	SP1; SP5; SP5'; SP6
TSS	Total solid suspended	All except SP7
VSS	Volatile solid suspended	All except SP7
P_{Total}	Total phosphorous	SP1; SP6
N_{Total}	Total Kjedldhal Nitrogen	SP1; SP6
$N\text{-}NH_4$	Ammonium	SP1; SP6
$N\text{-}NO_{3-}$	Nitrate	SP6
V30	Sludge volume after 30min of settling	SP3
ISS*	Ratio VSS/TSS (%)	SP4
SAR*	Sodium absorption ratio (Q_r/Q_{in})	SP4
SVI*	Sludge volume index	SP3; SP3'
SRT*	Solids Retention Time (sludge age)	SP4; SP4'
CBO/CQO*	Biodegradability	SP1
FM*	Food to microorganism ratio	SP3; SP3'

* Generally, process treatment variables correspond to formulas that include some measured treatment parameters.

The data we have access consist of daily averages of the measured parameters over the already referred sampling points (SP). The database contains a total of 92 examples, which corresponds to a period of one year, shown high data dimensionality with about 120 attributes. However there are many attributes with high rates of missing values. Actually the total percentage of missing values in database is about

35%. Moreover two examples with missing values almost in all the attributes were discarded, and thereby remained 90 valid examples in database. Qualitative parameters (i.e. micro fauna) were also registered. However these records, in small number, were registered in different days of the registered quantitative data. The solid phase of treatment presents a high rate of missing values (80%), and the same applies to data from septic tanks. Thus we considered only parameters related to SP1, SP2, SP3, SP4, SP5, and SP6, including flow rates and process variables. External meteorological data, as the weather and average temperature, were still added, in order to enrich the information about the treatment processes. Table 1 shows the major treatment parameters and the respective sampling points, which were selected after a first analysis of the database. The number of attributes under consideration is high, since there are several parameters and some of them are measured at various points. Since there are data from two parallel treatment lines, this factor also increases the degree of data dimensionality.

4 The Prediction Process

4.1 Data Analysis and Preparation

Before proceeding to the modeling phase, we will describe the processes of data preparation and analysis. As expected the data presented some curious aspects, some of them already reported above, as the high rate of missing values of the attributes. Treatment with two lines raised yet other issues. Sometimes one of the lines is inactive or simply no data is recorded, which in practice means more missing values. On the one hand, when considering attributes related to both treatment lines it increases the dimensionality and the missing values. On the other hand, when considering only one treatment line it reduces the dimensionality and missing values. That is, for the one line case the values are taken by the average of the two lines and in the case of inactivity of either is selected the data from the active one. However the average of the two treatment lines distorts data, even knowing that the effluent in SP6 converges from the two treatment lines. Based on these facts two approaches have been adopted and therefore investigated in this study. The first approach considers only one line of treatment, whose values are the averages of the parameters that are treated in two parallel lines (WWTP_1L). Hamed et al. (2004) [18] reports in their work that the input parameters (with parallel processing) contain the average value of the various treatment lines, hence using the same approach. The second approach considers both treatment lines, thus comprising all the parameters of both lines (WWTP_2L). In one of his studies, Dürrenmatt (2011) [12] also used this.

From all the parameters of this study (SP6), the nutrients were the only ones with about 45% of missing values. Our priority was given in the prediction of the parameters BOD, COD and TSS, which have limits defined by law, respectively 25, 125 and 35 mg/l to regulate them. Of the three parameters, since only BOD and TSS have records with values above the limits required by law, these were the first choices to be selected as prediction parameters. It is further noted that the results of the BOD measurements take 5 days, which underlines the interest to predict this parameter.

The data preparation is very important to obtain a dataset that is more easily accessed by modeling algorithms. The modeling tools used in this study (SVR) requires a numeric input data. So it was necessary to map the nominal attributes to numeric. The data from the WWTP contains only two nominal attributes, season and weather, that both were converted to integers. The normalization of the data ranges, in addition of being a pre-processing technique necessary in some algorithms, can bring considerable advantages even in algorithms that not require it. Some tests were conducted in this study showing that the normalization positively affects the performance of SVR. Also the study of Ali and Smith-Miles (2006) [1] shows that normalization increases the predictive performance of the tested datasets. Then was applied the normalization MinMax (-1 & 1) for all attributes passing to have their range from -1 to 1.

As we know, missing values usually represent a major problem in datasets. This case was no exception. Like most algorithms (including SVR) that cannot handle missing values, these have to be treated. In the study by Luengo et al. (2011) [25] is made an extensive series of tests with several missing values treatment techniques, and two interesting conclusions are drawn. The first says that the imputation methods that fill in missing values outperform the case deletion and the lack of imputation. The second is that no imputation method is best for all cases. With this in mind, in this study we tested some imputation methods. It was noted that few imputation methods outperformed the mean mode imputation (MMI), which is widely used. One of the exceptions was the imputation method with k-nearest neighbor (k-NN) that had slightly higher results and thus was chosen.

Some of the recommendations for data preparation referred in [30] were followed in this work, including the elimination of redundant attributes. Based on the correlation matrix analysis of the dataset the attributes highly correlated (i.e. TSS and VSS are about 0,991 correlated in SP1) were discarded. So VSS can be also discarded. We toke into account that these attributes are discarded only if they are very weak correlated with the prediction parameter. At this point, after the described data preparation, it was considered that both datasets were ready for the next stage (modeling).

4.2 Model Development

After we applied several cleaning and enrichment processes, the dataset we used stayed only with 90 valid examples. This required appropriate evaluation methods for the dataset size to avoid overfitted models as well to capture as much examples as possible on the model development. Having this, we seek the best practices that are generally used in similar problems. For instance, in the work of Cortez et al. (2006) [11], which also relates to a small data set problem (80 examples) was performed the evaluation of models according to the method 10-CV (*10-fold cross-validation*). The 10-CV evaluation is stated to be a good evaluation method by other several studies [23, 28], additionally repeating CV-10 reduces the variance and thus overcomes the normal evaluation 10-CV [22]. We adopted the evaluation method of 10 repeated 10-CV (10r-10CV), which makes a total of 100 different sets of training + test evaluated. Regarding the performance measures, the main measure adopted was the root mean squared error (RMSE), but is also presented the correlation coefficient (R) [39].

On modelling, feature selection brings many benefits. As we have a high dimensionality in our case, these techniques were used with two main objectives. The first one is to reduce the dimensionality in order to increase the predictive capability of the models. The second passes thereby to find the most important attributes in the prediction of BOD and TSS, thus facilitating the acquisition of knowledge about data.

There are several techniques for selecting attributes. These can be divided into wrappers, filters and embedded methods. In addition may also be used some dimensionality reduction methods - i.e. *Principal Component Analysis* (PCA) [17]. We tested some of these methods and the wrappers methods clearly outperformed the others, including filters and PCA, as in the work of Kohavi and John (1997) [24]. Among the wrappers methods tried it was chosen the *Optimize Evolutionary Selection* of RapidMiner, which is based on genetic algorithms.

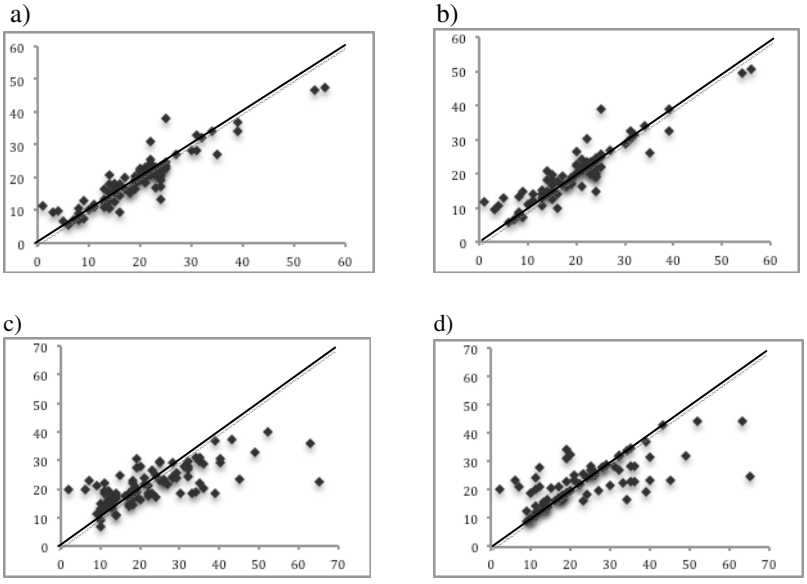


Fig. 3. Scatter plots comparing measured (x-axis) and predicted (y-axis) concentration values on *WWTP_1L* models - a) SVR-BOD; b) SMO-BOD; c) SVR-TSS; and d) SMO-TSS

Two implementations of SVM were investigated in this study, the SVR implementation of the *LibSVM* library [8] present in *RapidMiner,* and the SMOReg algorithm, which is an extension from *WEKA*, whose implementation is based on the work of Shevade et al. (2000) [32] and [33], as cited in the documentation [38]. In the case of SVR (LibSVM) there are two variants the ε-SVR and ν-SVR [31]. Here as ν-SVR presented slightly better performance hence it was the adopted one. The kernel selected in both algorithms was the RBF, as it is recommended as a good first choice and also has fewer hyper parameters to set up than other kernels. In the parameter optimization we followed the Grid Search technique. With it we seek first the values in large ranges and after we refine the search at shorter ranges [20]. The described modeling process was executed in the development of booth predictive tasks for BOD and TSS, in each of the approaches (WWTP-1L and WWTP-2L).

5 Analysis and Results Evaluation

After the modelling phase we evaluated the performance of the models produced. Similarly to previous results, we got four graphs showing the comparison between the measured and predicted concentrations (figure 6). This time the models respects to the two treatment lines approach (WWTP_2L).

Table 2. WWTP_1L results

Task	SVR		SMOReg	
	RMSE	*R*	*RMSE*	*R*
BOD	4.16	0.88	**4.02**	**0.88**
TSS	9.68	0.56	**9.41**	**0.57**

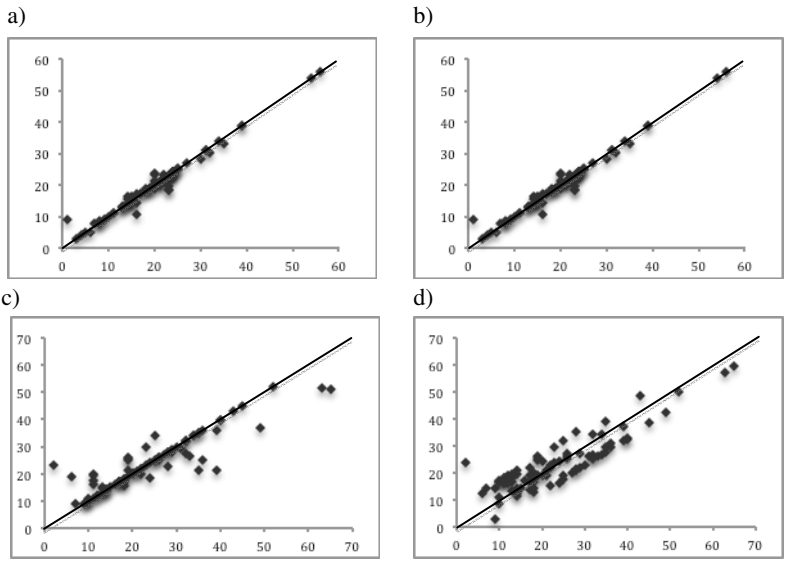


Fig. 4. Scatter plots comparing measured (x-axis) and predicted (y-axis) concentration values on *WWTP_2L* models. a) SVR-BOD; b) SMO-BOD; c) SVR-TSS; and d) SMO-TSS

As noted in the approach with one line, also in WWTP_2L the results of BOD models were higher than the TSS (table 3). In BOD prediction models the SVR and SMO methods had very similar performances. Although, SMO had better results than the SVR in the case of TSS prediction. Nevertheless, this last difference is not statistically significant, according to pairwise comparison of the statistical test (Student's t-test) with 95% confidence interval of the RMSE values [15]. Comparing the two approaches (WWTP_1L and WWTP_2L), it is clear that all the models of WWTP_2L were superior to WWTP_1L. In BOD prediction models the differences in results, between the two approaches, are statistically significant in both methods

(SVR and SMO). However in the prediction of TSS only in the results of SMO is that the difference is statistically significant between the two approaches.

The visualization of the attributes that were frequently selected in modelling provides potentially useful information in the analysis of the behaviour of the WWTP. Table 4 describes the attributes with greater influence in the prediction task of each parameter. Some observations can be drawn based on the information revealed on Table 4. For instance, BOD has many attributes related to the initial stages of treatment, as well as three process variables features.

Table 3. WWTP_2L results

Task	SVR		SMOReg	
	RMSE	*R*	*RMSE*	*R*
BOD	3.24*	0.94	**3.22***	**0.93**
TSS	8.67	0.65	**7.79***	**0.75**

* Statistically significant under pairwise comparison with the same model on WWTP_1L approach.

In TSS the nutrients were shown to be relevant, besides their high rate of missing values. Yet, a final remark for the parameter O_{2_AXZ} of SP2, which was been selected by all BOD and TSS models. However, it should be emphasized that this type of information should be analysed by an expert from the WWTP, whose knowledge of the platform determines effectively the utility and importance of this kind of information.

Table 4. Main features selected in modelling of BOD and TSS

Task	Features
BOD	SP1_COD; SP1_ BOD; SP1_ TSS; SP2_O_{2_AXZ}; SP3_V30; SP3_O_2; SP4_ P_{redox}; SP5_BOD; SP7_Q_p; SP3_SVI; SP4_SRT
TSS	Season; SP1_ N_{Total}; SP1_N-NH4; SP2_O_{2_AXZ}; SP2_P_{redox_AXZ}; SP2_TSS; SP2_VSS; SP3_P_{redox}; SP4_TSS; SP4_VSS; SP5_pH: SP5_TSS; SP5_COD; SP5_BOD; SP1_BOD/COD

6 Conclusions and Future Work

This study demonstrates that it is possible to apply predictive models successfully in the prediction of the behavior of a WWTP, especially when we are dealing with the quality parameter BOD. Despite the lowers TSS results, we found a good relationship between TSS and the other SP6 attributes, after obtaining a model that was able to predict efficiently the SP6 parameters (i.e. BOD or COD). The TSS parameter was also predicted too with accuracy. The two adopted algorithms (SVR and SMO) provided quite similar results, in spite of the models make use of different input

features. In both research approaches it was found that modeling with two WWTP lines provides superior results, which may indicate that the WWTP_1L approach distorts the data when makes the average of lines parameters. However WWTP_1L approach brings greater interpretation ease due to the lower dimensionality and consequent simplification of the problem.

In spite of using techniques known by their good generalization, it is inevitable not to assume that, because we had a small dataset. Complete information captured on the real WWTP, integrating raw sampling data, was not possible to obtain. Hence, as future work should be further investigated this problem in the presence of more detailed data. The fact that records contain daily averaged values also affected the efficiency of models, once it does not capture all the dynamics of the treatment process. A collection of data with a finer grain measurement (i.e. every 2 hours) would probably increase the effectiveness in forecasting, as was described in the work of Atanasova and Kompare [3].

Based on this real case study, we can equate new prediction models of the outflow quality in a future implementation of a decision support system for WWTP analysts at the time they insert measured data from treatment processes. Clearly, more research on this treatment plant, and even about the issues mentioned above, must be made, in order to develop a useful and reliable system.

References

1. Ali, S., Smith-Miles, K.A.: Improved Support Vector Machine Generalization Using Normalized Input Space. In: Sattar, A., Kang, B.-H. (eds.) AI 2006. LNCS (LNAI), vol. 4304, pp. 362–371. Springer, Heidelberg (2006)
2. Atanasova, N., Kompare, B.: Modelling of Wastewater Treatment Plant with Decision and Regression Trees. In: 3rd Workshop on Binding Environmental Sciences and Artificial Intelligence (2002a)
3. Atanasova, N., Kompare, B.: Modelling of waste water treatment plant with regression trees. In: Proc. of the Third International Conference on Data Mining. WIT Press, Bologna (2002b)
4. Belanche, L.A., et al.: Towards a Model of Input-Output Behaviour of Wastewater Treatment Plants using Soft Computing Techniques. Environmental Modeling and Software 14(5), 409–419 (1999)
5. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Haussler, D. (ed.) Proceedings of the Fifth Annual Workshop on Computational Learning Theory, COLT 1992. ACM Press (1992)
6. Burges, C.J.C.: A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery 2(2), 121–167 (1998)
7. Cărbureanu, M.: Pollution Level Analysis of a Wastewater Treatment Plant Emissary using Data Mining. Petroleum-Gas University of Ploieşti Bulletin Mathematics-Informatics-Physics Series LXII(1), 69–78 (2010)
8. Chang, C.-C., Lin, C.-J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2(3), 27:1–27:27 (2011)
9. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(3), 273–297 (1995)

10. Cortez, P.: Data Mining with Multilayer Perceptrons and Support Vector Machines. In: Holmes, D.E., Jain, L.C. (eds.) Data Mining: Found. & Intell. Paradigms. ISRL, vol. 24, pp. 9–25. Springer, Heidelberg (2012)
11. Cortez, P., et al.: Lamb Meat Quality Assessment by Support Vector Machines. Processing Letters 24(1), 41–51 (2006)
12. Dürrenmatt, D.J.: Data Mining and Data-Driven Modelling Approaches to Support Wastewater Treatment Plant Operation. PhD Thesis. ETH, Zurique (2011)
13. Dixon, M., et al.: Data mining to support anaerobic WWTP monitoring. Control Engineering Practice 15(8), 987–999 (2007)
14. Drucker, H., et al.: Support vector regression machines. Electronic Engineering 1, 155–161 (1997)
15. Flexer, A.: Statistical Evaluation of Neural Network Experiments: Minimum Requirements and Current Practice. In: 3th European Meeting on Cybernetics and Systems Research (1996)
16. Gallop, J.R., et al.: The use of data mining for the monitoring and control of anaerobic wastewater plants. In: 4th International Workshop on Environmental Applications of Machine Learning (2004)
17. Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection. Journal of Machine Learning Research 3(7-8), 1157–1182 (2003)
18. Hamed, M.M., Khalafallah, M.G., Hassanien, E.A.: Prediction of wastewater treatment plant performance using artificial neural networks. Environmental Modelling Software 19(10), 919–928 (2004)
19. Hong, Y., Fei, L., Yuge, X., Jin, L.: GA Based LS-SVM Classifier for Waste Water Treatment Process. In: 27th Chinese Control Conference (2008)
20. Hsu, C.-W., Chang, C.-C., Lin, C.-J.: A Practical Guide to Support Vector Classification. Bioinformatics 1(1), 1–16 (2010)
21. Huang, Z., Luo, J., Li, X., Zhou, Y.: Prediction of Effluent Parameters of Wastewater Treatment Plant Based on Improved Least Square Support Vector Machine with PSO. In: 1st International Conference on Information Science and Engineering, ICISE (2009)
22. Kim, J.-H.: Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. Computational Statistics and Data Analysis 53(11), 3735–3745 (2009)
23. Kohavi, R.: A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In: International Joint Conference on Artificial Intelligence, Montreal, Canada (1995)
24. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artificial Intelligence 97(1-2), 273–324 (1997)
25. Luengo, J., García, S., Herrera, F.: On the choice of the best imputation methods for missing values considering three groups of classification methods. Knowledge and Information Systems, 1–32 (2011)
26. Luo, F., Yu, R.-H., Xu, Y.-G., Li, Y.: Effluent Quality Prediction of Wastewater Treatment Plant Based on Fuzzy-Rough Sets and Artificial Neural Networks. In: Sixth International Conference on Fuzzy Systems and Knowledge Discovery - FSKD 2009 (2009)
27. Metcafl, Eddy: Wastewater Engineering: Treatment and Reuse, 4th edn. McGraw-Hill (2003)
28. Molinaro, A.M., Simon, R., Pfeiffer, R.M.: Prediction error estimation: a comparison of resampling methods. Bioinformatics 21(15), 3301–3307 (2005)
29. Platt, J.C.: Using Analytic QP and Sparseness to Speed Training of Support Vector Machines. Optimization 11, 1–8 (1999)

30. Pyle, D.: Data Preparation for Data Mining. Morgan Kaufmann Publishers, Inc., San Francisco (1999)
31. Schölkopf, B., Smola, A., Williamson, R., Bartlett, P.L.: New support vector algorithms. Neural Computation 12(5), 1207–1245 (2000)
32. Shevade, S.K., Keerthi, S.S., Bhattacharyya, C., Murthy: Improvements to the SMO Algorithm for SVM Regression. IEEE Transactions on Neural Networks 11(5), 1188–1193 (2000)
33. Smola, A.J., Schölkopf, B.: A tutorial on support vector regression. Statistics and Computing 14(3), 199–222 (2004)
34. Vapnik, V.N.: The Nature of Statistical Learning Theory, 2nd edn., New York (1995)
35. Vapnik, V.N.: An overview of statistical learning theory. IEEE Transactions on Neural Networks 10(5), 988–999 (1999)
36. Vapnik, V.N., Chervonenkis, A.Y.: Theory of pattern recognition. Nauka, Moscow (1974) (in Russian)
37. Wang, L.-J., Chen, C.-B.: Support Vector Machine Applying in the Prediction of Effluent Quality of Sewage Treatment Plant with Cyclic Activated Sludge System Process. In: IEEE International Symposium on Knowledge Acquisition and Modeling Workshop. KAM Workshop 2008 (2008)
38. WEKA, n.d. Class SMOReg, `http://weka.sourceforge.net/doc/weka/classifiers/functions/SMOreg.html` (accessed July 2, 2012)
39. Witten, I.H., Frank, E., Hall, M.A.: Data Minig: Pratical Machine Learnign Tools and Tecniques, 3rd edn. Morgan Kaufmann (2011)
40. Yang, B.-l., Zhao, D.-A., Zhang, J.: Prediction system of sewage outflow COD based on LS-SVM. In: 2nd International Conference on Intelligent Control and Information Processing, ICICIP (2011)

Mining Floating Train Data Sequences for Temporal Association Rules within a Predictive Maintenance Framework

Wissam Sammouri, Etienne Côme, Latifa Oukhellou, and Patrice Aknin

Université Paris-Est, IFSTTAR, GRETTIA, F-77447 Marne la Vallée, France
{wissam.sammouri,etienne.come,latifa.oukhellou,patrice.aknin}@ifsttar.fr

Abstract. In order to meet the mounting social and economic demands, railway operators and manufacturers are striving for a longer availability and a better reliability of railway transportation systems. Commercial trains are being equipped with state-of-the-art onboard intelligent sensors monitoring various subsystems all over the train. These sensors provide real-time spatio-temporal data consisting of georeferenced timestamped events that tend sometimes to occur in bursts. Once ordered with respect to time, these events can be considered as long temporal sequences that can be mined for possible relationships leading to association rules. In this paper, we propose a methodology for discovering association rules in very bursty and challenging floating train data sequences with multiple constraints. This methodology is based on using null models to discover significant co-occurrences between pairs of events. Once identified and scrutinized by various metrics, these co-occurrences are then used to derive temporal association rules that can predict the imminent arrival of severe failures. Experiments performed on Alstom's TrainTracer™ data show encouraging results.

Keywords: Association rules, Sequential data mining, Null models, Significance testing.

1 Introduction

Similar to floating car data systems which are now broadly implemented in road transportation networks [12,25], floating train data systems have also been recently developed in the railway domain [21,23]. Intelligent sensors monitoring various subsystems of the train provide a real-time flow of data consisting of georeferenced events, along with their spatial and temporal coordinates. Once ordered with respect to time, these events can be considered as one long temporal sequence for each train unit. This has created a necessity for sequential data mining techniques to be applied on such data in order to derive meaningful associations rules. Once discovered, these rules can then be used to perform an on-line analysis of the incoming event stream in order to predict the occurrence of target events, i.e, severe failures that require immediate corrective maintenance actions.

P. Perner (Ed.): ICDM 2013, LNAI 7987, pp. 112–126, 2013.

In this paper, a new methodology for discovering association rules in bursty temporal data sequences based on null models, a significance testing approach for pairwise co-occurrences, is applied on temporal sequences extracted from a floating train data system developed by Alstom Transport called TrainTracer™. The available TrainTracer™ dataset consists of a series of timestamped events where each event is identified by a unique numerical code. This paper extends the work of [9] by evaluating the performance of different null models on temporal data sequences, as well as the introduction of a post-treatment and assessement framework of the discovered couples in order to derive viable association rules. The main difficulties confronting the mining process in our application lie in the heavy presence of data bursts as well as the rareness of target events. Also, the prediction is subject to two important constraints: target events should be predicted early enough to allow logistic and preventive maintenance actions to be taken properly. Secondly, prediction accuracy should be high enough to avoid heavy intervention costs in case of false predictions.

This paper is structured as follows: In section 2, a survey of previous work in relevant literature is presented. The problem of association rule mining in temporal sequences is formulated and defined in 3 and the TrainTracer™ data are briefly discussed. In section 4, the various null models and co-occurrence scores used are explained. Experiments on synthetic data are presented in 5.1 followed by the metrics used to derive association rules from discovered event couples in 5.2. Finally, the experimental results on real TrainTracer™ data are discussed in 5.3 prior to concluding in 6.

2 Related Work

2.1 Association Rule Mining Algorithms

Association rule mining is an important data mining field that aims to discover patterns of co-occurrences and affinities between items in a transaction database or between events in a data sequence. Although the initial motivation behind the first algorithms was to tackle market basket analysis problems [1], association rule mining approaches have been extensively developed in the past years and were applied in a wide range of domains such as environmental monitoring [24], bioinformatics [9], recognition of human behavior [11,19], telecommunications [18], etc. The developed techniques are very diversified to comply with different types of problems such as mining frequent or rare patterns in transaction databases or sequences of events that can be temporal or non-temporal [18,2,26,27]. For example, an application closely related to ours is the analysis of alarms in telecommunication and sensor networks [22].

The concept of association rules was first introduced in [1,2] through the Apriori algorithm by proposing a frequency-based support-confidence statistical metrics framework. Most association rule mining techniques are variants of Apriori and focus on finding frequent itemsets and patterns. However, the disadvantage of relying strictly on frequency constraints is that it does not differentiate between significant and insignificant items or events and allows those fulfilling the

frequency threshold to survive, regardless of their informative value, which results in the discovery of numerous spurious patterns. Significant patterns that are not frequent enough can rarely be detected unless a low frequency threshold is used which would imply in its turn a very heavy computational time.

To address this problem, recent years have witnessed the uprisal of other algorithms focusing on mining significant and rare patterns such as constraint-based data mining approaches which increase the level of user engagement in the mining process [20,8] as well as weighted association rule mining techniques which value the importance of items by assigning them weights either manually (using expert knowledge) or automatically using models based on the quality of interactions and connections between items [14]. Other interesting approaches were proposed based on significance testing of pairwise co-occurrences such as null models, randomization algorithms that are followed by the calculation of different scores and a statistical hypothesis test to assess the significance of data in [9,10,15,13] and the T-patterns algorithm, which exploits the temporal dimension by investigating the statistical dependence of inter-arrival times of couples of events in order to highlight possible relationships and then build trees of hierarchical temporal dependencies [22,17].

2.2 Association Rule Mining in Railway Applications

Most of the data mining approaches applied in railway applications were based on machine learning and classification techniques. It is not until recently that association rule mining algorithms were used on railway data in an attempt to discover significant relationships between data elements which might help predict future incidents and open the doors wide for predictive maintenance strategies. For example, in [4], a closed-episode mining algorithm, CLOSEPI, was applied on a dataset containing the passage times of trains through characteristic points in the Belgian railway networks. The aim was to detect interesting patterns that will help improve the total punctuality of the trains and decrease train delays. Similary, Flier et al [5] tried to discover dependencies between train delays in the aim of supporting planners in improving timetables. The Apriori algorithm was applied in [16] on railway tunnel lining condition monitoring data in order to extract frequent association rules that might help enhance the tunnel's maintenance efforts. Various association rule mining approaches were used in [19] to analyze accident data sets of a railway network.

3 Problem Setting

3.1 Objective

The main goal of this work is to mine temporal data sequences extracted from the TrainTracer™ database for significant co-occurrences between couples of events. These co-occurrencs will then be assessed to derive association rules. We consider the input as a sequence of events, where each event is expressed by a unique numerical code and an associated time of occurrence.

Given a set E of event types, an event is defined by the pair (R,t) where $R \in E$ is the event type and $t \in \Re^+$ its timestamp, i.e, associated time of occurrence. An event sequence S is a triple (S, T_s, T_e), where S is an ordered sequence of events of the form $\langle (R_1, t_1), (R_2, t_2), ..., (R_n, t_n) \rangle$ such that $R_i \in E \ \ \forall i = 1, ..., n$ and $T_s \leq t_1 \leq t_n \leq T_e$. We define an association rule as an implication of the form $A \longrightarrow B$, where the antecedent and consequent are sets of events with $A \cap B = \phi$.

In the analysis of sequences we are interested in association rules that help predict target events. This means that mining is oriented towards association rules $A \longrightarrow B$ where B is a target event. Due to the very complex nature of the data with the heavy presence of bursts, noise as well as the rarity of target events, we have decided as a primary approach to limit our search to pairwise co-occurrences leading to length-2 association rules, where A and B consist each of a single event. Once found, these rules can be extended to length-3 and more. The null models algorithm is applied on the TrainTracer™ data extracts in disposal in order to discover significant couples of events. Each couple is then assessed in order to verify its abidance to the two constraints previously explained in 1. Significant couples respecting both constraints are considered as statistically significant association rules and are submitted to railway experts for physical assessment prior to their integration in the TrainTracer™ software. The real-time monitoring of arriving events would allow the online prediction of target events, i.e failures, using these rules. Once a target event is predicted, the maintenance teams are alerted to initiate preventive maintenance procedures.

3.2 TrainTracer™ Data

The current work was performed on a 6-month data extract from the TrainTracer™ database. TrainTracer™ is a state-of-the-art software concieved by Alstom to collect and process real-time data sent by a fleet of trains equipped with onboard sensors monitoring 31 various subsystems such as the auxiliary converter, doors, brakes, power circuit and Tilt. This data consists of series of timestamped events where each event is identified by a numerical code in addition to context variables providing physical, geographical and technical details of the train at its exact time of occurrence. There are 1112 event types existing in the data with varying frequencies and distributions. Approximately 9.1 million events have occurred in the 6-month period. Since events may vary between warnings, alarms and normative events, they have been divided into 5 intervention categories describing how critical they are. These categories in an increasing order of importance are: Status (category 1), Driver information (category 2), Driver Action Low (category 3), Maintenance (category 4) and Driver Action High (category 5).

All "Driver Action High" events require an immediate corrective maintenance actions and are thus target events. Among them, we are particularly interested in those related to tilt and traction. This is due to the fact that tilt and traction failure events are highly probable to impose a mandatory stop. In total, there are 46 tilt and traction event types requiring high action from the driver existing in the data extract within disposal, consisting 0.5% of all events. In the next section, the null models algorithm will be explained.

4 Null Models

A null model is a sequence-generating model that is based on the randomization of data sequences while preserving their general statistical characteristics. Certain elements of the data are held constant while others are varied stochastically. These models evaluate relationships between events by means of a statistical hypothesis test, where the null hypothesis refers to the significance of a particular relationship in the original data sequence. To solve this test, the initial data sequence is randomized using null models and the co-occurrence scores of each randomization are calculated. The significance of these scores is then evaluated by means of an empirical p-value which is equal to the fraction of randomizations with higher co-occurrence scores than the initial data, and comparing it to a pre-defined threshold. If it is inferior to the threshold, the event couple under scrutiny is considered to be significant. Null models have been applied in various domains such as ecology [7,6], physiology [3] and genetics [10,13].

4.1 Co-occurrence Scores

In order to quantify the degree of co-occurrence of an event couple, three different scores are used and presented below:

Given a set E of event types and $S = \{(R_1, t_1), (R_2, t_2), ..., (R_n, t_n)\}$ is a temporal sequence of length l time units, $R \in E$ and $t_i \in \Re^+$. Consider the event couple under scrutiny (A, B) where $A \in E$ and $B \in E$, let $N(A)$ be the number of times an event type A occurs in the sequence S and denote $f(A) = N(A)/n$. Divide the sequence into non-overlapping windows of width w. The total number of windows is equal to $\lceil l/w \rceil$.

- The window count score $W(A, B, S)$ for event couple (A,B) is the number of windows in which at least one event of type A and one event of type B occur. Thus, $W(A, B, S) \in \{0, ..., \lceil l/w \rceil\}$.
- The co-occurrence count score $C(A, B, S)$ is the number of events of type A that are succeeded or preceeded by at least one event of type B within distance w. Thus, $C(A, B, S) \in \{0, ..., N(A)\}$.
- The directed co-occurrence count score $D(A, B, S)$ is the number of events of type A that are succeeded by at least one event of type B within distance w. Thus, $D(A, B, S) \in \{0, ..., N(A)\}$ as well.

Figure 1 is a graphical illustration of the W, C and D co-occurrence scores for a given event sequence S of length $l = 1000$ hours with 4 event types A, B, C and D and window width, i.e, maximum co-occurrence distance parameter $w = 250$ hours. Consider the couple under scrutiny to be (A, B). In the original sequence, events A and B occur together in only one of the adjacent windows, thus W(A, B, S) = 1. The number of A events that are succeeded or preceeded by a B event within a distance w is 3, C(A, B, S) = 3. The number of A events that are succeeded by B event within a distance w is 2, hence D(A, B, S) = 2.

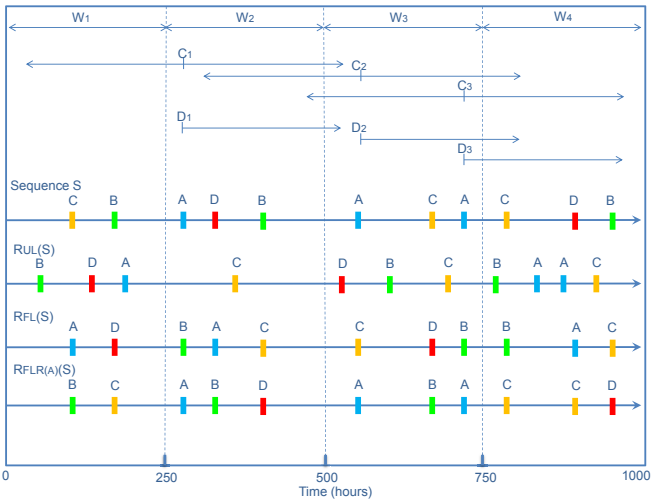


Fig. 1. Graphical illustration of the W, C and D co-occurrence scores and the UL, FL and $FL(A)$ null models for an event sequence S of length $l = 1000$ hours with 4 event types A, B, C and D and window width parameter $w = 250$ hours

4.2 Randomizing Data

In order to evaluate the significance of a co-occurrence score in the initial sequence, a null model is needed. In this paper, three different null models are used: the uniform locations (UL) model [15], the fixed locations (FL) model [10,13] and the fixed locations fixed event type ($FL(R)$) model [9]. The randomizations are generated by the following procedure:

- The Uniform Locations UL null model consists of generating sequences resulting from the randomization of both the timestamps and event codes in the sequence. That is, for an event type R, $N(R)$ events (R, t) are generated, where $N(R)$ is the number of occurrences of the event R in the original sequence. Each timestamp t is selected uniformly at random over the temporal length l of the original sequence.
- The randomized sequence $R_{FL}(S)$ is obtained by the Fixed Locations FL null model by keeping timestamps fixed, and assigning event types at random on these locations according to their frequencies in the original sequence.
- The randomized sequence $R_{FL(R)}(S)$ for a sequence S and an event type R is defined similarly to $R_{FL}(S)$, with the exception that the occurrences of events of type R are kept unchanged.

Consider again Figure 1. For the $R_{UL}(S)$ randomized sequence where event locations are completely randomized, events A and B occur together in 2 of the adjacent windows thus W(A, B, $R_{UL}(S)$) = 2, the number of events of type A that are followed or preceeded by an event of type B within distance w is 3,

hence C(A, B, $R_{UL}(S)$) = 3. Since no events of type A are followed by events of type B, D(A, B, $R_{UL}(S)$) = 0. Similarly, the co-occurrence scores for the couple (A, B) in the $R_{FL}(S)$ sequence (where timestamps of the initial sequence S are fixed and event types randomized) are W(A, B, $R_{FL}(S)$) = 2, C(A, B, $R_{FL}(S)$) = 3, D(A, B, $R_{FL}(S)$) = 1. In the $R_{FL(A)}(S)$ sequence generated similary to $R_{FL(A)}(S)$, except that events of type A are left untouched, the scores are: W(A, B, $R_{FL(A)}(S)$) = 2, C(A, B, $R_{FL(A)}(S)$) = 3, and D(A, B, $R_{FL(A)}(S)$) = 2.

4.3 Calculating p-values

For a given sequence S and a null model $M \in \{UL, FL, FL(R)\}$, the empirical p-value for an event couple (A, B) is the fraction of randomizations in which the W (or C, D) score in the randomized sequences $R_M(S)$ is superior to the W (or C, D) score of the original sequence S:

$$p_W(A, B, M, S) = \frac{\#(W(A, B, S) \leq W(A, B, R_M(S)))}{\#(W(A, B, R_M(S)))} \quad (1)$$

5 Experimental Study

5.1 Synthetic Data

The aim of this study is to tune the window size parameter w to the value that will most likely lead to optimal results on real data, as well as to assess the performance of the null models on synthetic bursty and non bursty sequences. A burst occurs when a large number of events of the same or of different types are signalled in a very short period of time mainly due to a sensor or reception error. The comparison is based on two diagnostics: (1) the efficiency to discover the planted pattern and (2) the ability to discard non-existing ones (false positives).

Generation Protocol: The generative model of the data is as follows. The timestamps of each event type are generated separately by means of a Poisson process of parameter λ_j, where $j \in \{1, ..., m\}$ over a period of l hours. λ values are unique for each event type and are generated by means of a uniform distribution on the interval $[L1_{min}; L1_{max}]$ for sparse segments and $[L2_{min}; L2_{max}]$ for dense (burst) segments, such that $L2_{max} << L1_{min}$. Lapse times between bursts (inter-burst times) and the length of each burst are generated randomly by means of a uniform distribution. Generated sequences contain a directed co-occurrence pattern between two event types A and B denoted (A, B) or $A \longrightarrow B$. That is, whenever an **A** event occurs, a Bernoulli distribution of predefined success probability p determines whether this event will be followed by a **B**-event or not. If so, a **B** event is planted with a temporal delay T_{AB} generated from a uniform distribution on an interval $[0, s]$. The reason why we focused on directed patterns is because our main goal with the real train data is to discover association rules of the form $A \longrightarrow B$ where B is a target event.

Table 1. Results of the 3 null models (UL, FL, FL(R)) with 3 co-occurrence scores (W, C, D) on sequences of length $l = 4500$ hours, $p = 0.8$ and varying values of w.

I - Mean number of discovered event couples in 100 generations									
	UL			FL			FL(R)		
w(h)	W	C	D	W	C	D	W	C	D
0.5	4	3	2	2	2	2	3	3	2
1	2	4	2	2	2	1	3	3	3
5	2	3	3	2	3	2	3	3	2
10	2	2	2	2	2	2	3	2	2.5
20	0	0	1	0	0	1	0	0	1

II - Number of generations where (a,b) was found significant									
	UL			FL			FL(R)		
w(h)	W	C	D	W	C	D	W	C	D
0.5	100	100	100	100	100	100	100	100	100
1	100	100	100	100	100	100	100	100	100
5	96	88	100	96	89	100	99	83	100
10	56	12	100	64	15	100	88	9	99
20	2	0	14	2	0	15	2	0	9

In order to find the optimal value of the maximum co-occurrence window size w, 100 sequences of length 4500 hours were generated (equivalent to 6 months, i.e, the length of the data extracts under disposal). The number of randomizations was fixed to 100 and the p-value threshold to 1%. All sequences consisted of 10 event types numbered from 1 to 10 with variable frequencies. The injected pattern was $8 \longrightarrow 9$ with $T_{89} \in [0, 1]$ hours. Table 1 shows the results.

Experimental diagnostic I in table 1 represents the mean number of pairs that were discovered by the models in the 100 generated sequences. Diagnostic II represents the number of generations where the injected pattern $8 \longrightarrow 9$ was discovered. The best results were obtained for $w = 30$ minutes and 1 hours for all models and scores except for the D score which yielded perfect results for $w = 5$ and 10 hours as well. Out of 90 possible couples ($N^2 - N$ for $N = 10$), only one was expected to be discovered (8,9). The false positive rate did not exceed 3% for all algorithms. It is clear that the efficiency of the null models decreases with larger values of w, which is due to the fact that it is more probable to have randomizations with a score as high as the initial sequence when the value of w is high, hence it is more probable to obtain a p-value higher than the threshold, leading to the couple's rejection. However, since a decrease in the value of w would imply a decrease in the scanning distance (leading to the negligence of couples with long inter-event times), a trade-off should be considered. In order to test the efficiency of null models on data with bursts, sequences containing sparse and dense segments were generated. The 100 generated sequences consisted of 12 event types numbered from 1 to 12. A directed relationship was established between events 11. and 12 existing only in dense zones, that is, $11 \longrightarrow 12$.

The bernoulli success probability p was fixed to 0.8, $w = 5$ h, number of randomizations = 100 and the inter-event time $T_{11,12} \in [0, 1]$ hours.

Varying the number of bursts between 10 and 40 with burst size $\in [1, 3]$ minutes, results in Table 2 show that the D score outperforms the C and W scores.

Repeating the experiments with bursts of size $\in [3, 6]$ minutes and with sequences of length 1000 hours reveals that the FL null model with the C and D scores outperforms both the UL and the FL(R) null models in predicting the injected pattern for all values of w, while FL(R) is more advantageous when it comes to false positive rate (the expected number of couples to be discovered was 1 out of 132 possible couples). The FL(R) model works best with the W and D scores whereas both the UL and FL null models work best with the C and D scores whether in bursty or non-bursty data sequences.

Table 2. Results of the 3 null models (UL, FL, FL(R)) with 3 different scores (W, C, D) on bursty sequences of length $l = 4500$ hours, burst size $\in [1, 3]$ minutes, $w = 5$ hours, $p = 0.8$ and varying number of non-overlapping bursts.

I - Mean number of discovered event couples in 100 generations									
	UL			FL			FL(R)		
# Bursts	W	C	D	W	C	D	W	C	D
10	2	23	28	6	23	21	6	5	5
20	0	23	33	12	25	22	10	9	6
30	0	23	37	20	26	23	13.5	11	8
40	0	22	37	28	27	23	17.5	14.5	8

II - Number of generations where (a,b) was found significant									
	UL			FL			FL(R)		
# Bursts	W	C	D	W	C	D	W	C	D
10	68	100	100	24	100	100	89	84	100
20	11	100	100	1	100	100	84	68	100
30	1	100	100	1	100	100	60	30	100
40	0	100	100	0	100	100	28	4	100

Hence, in conclusion, experiments have shown that the D score outperforms the W and C scores in bursty sequences with a clear advantage in prediction rate and the directionality aspect. The FL null model slightly outperforms the UL and FL(R) null models and maintains a high prediction rate over various sequence lengths and types.

5.2 Deriving Association Rules from Discovered Significant Couples

In this section, we discuss the metrics used to analyze the significant couples discovered by the null models in order to derive robust association rules. As mentioned in section 1, rules in our application should respect two major constraints: a high global accuracy on one hand and a sufficiently large warning time

on another. Prior to explaining the choice of metrics, it is important to define formally the notion of a correct prediction. A prediction is correct if a target event occurs within its prediction period [W, M], also called critical interval, which is defined by a warning time, W, and a monitoring time, M. The warning time is the time delay before a target event becomes highly probable to occur. The monitoring time determines how far into the future the prediction extends.

Modeling Inter-event Times

Railway experts from Alstom have asked for an inter-event time of 30 minutes and a prediction period extending to 24 hours. Thus, the critical interval was fixed to [30 minutes , 24 hours]. The approach adopted in this work to evaluate the inter-event time between events of discovered couples is the following: The T_{AB} vector of each (A, B) couple discovered by the null models is computed, which is equivalent to the vector of the time distance between every occurrence of an A-event and the first succeeding B-event. If the median of this vector is ≥ 30 minutes, the couple is considered to have a sufficiently acceptable inter-event time and thus abides the inter-event time constraint.

Interestingness Measures

In order to scrutinize the robustness of the discovered couples, two interestingness measures are computed: Recall and Precision. They are defined by:

$$Recall = \frac{\#Predicted\ target\ events}{\#Total\ target\ events} \tag{2}$$

$$Precision = \frac{\#True\ predictions}{\#Total\ predictions} \tag{3}$$

The recall represents the percentage of target events that were predicted while the precision represents the percentage of correct predictions. Intuitively, a high recall value implies a low rate of false negatives, i.e, only few target events were missed and left unpredicted. A high precision value however reflects a high predictive capability and indicates a low rate of false positives, i.e, wrong predictions. Due to the fact that some events have occured very frequently in a limited number of trains only, the calculation of the recall/precision measures of couples consisting of at least one of such events is effected negatively and leads to erroneous high values. To overcome this inconvenience, a filter was introduced prior to the calculation of interestingness measures of couples discovered by the three null models. This filter identifies trains where the frequency of an event is superior to $\overline{x} + 3\sigma$, where $\overline{x}$ refers to the mean frequency of an event among all trains and σ its standard deviation. For each (A, B) event couple, trains in which the frequency of A or B events exceeds that threshold are neglected. This procedure renders recall and precision values more robust. In the end, discovered couples abiding the inter-event time constraint and having acceptable recall and precision values are considered as significant association rules.

5.3 Real Data

In this section, the results obtained by the UL, FL and FL(R) null models with three different co-occurrence scores W, C and D will be presented and discussed. Results on synthetic data have shown that the efficiency of null models decreases with high values of w. However, knowing that a small value signifies a short co-occurrence scanning distance, this would mean that all pairwise co-occurrences with an inter-event time longer than w will be neglected. Thus, a trade-off was considered and the null models were applied on the real TrainTracer™ data sequences with $w = 5$ hours. Table 3 shows the number of significant couples discovered by each null model.

Table 3. Number of significant event couples discovered by the UL, FL and FL(R) null models (p-value threshold = 1%) respectively in the TrainTracer™ data sequences

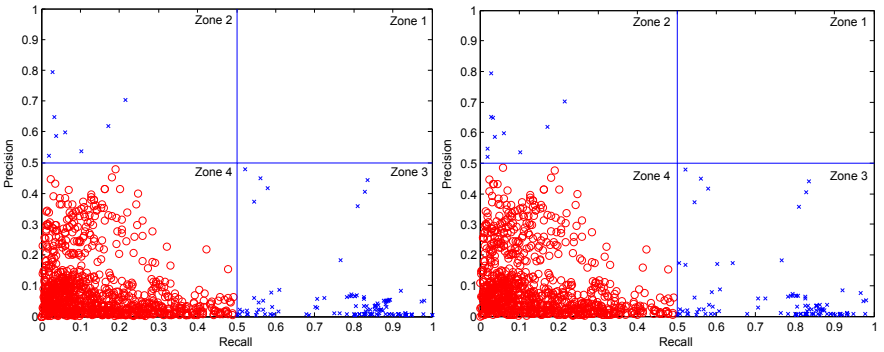
UL			FL			FL(R)		
W	C	D	W	C	D	W	C	D
879	2023	1760	629	1650	1454	1057	771	638

As explained in 5.2, all of the discovered couples were subjected to multiple evaluation processes in order to determine those satistfying the two constraints introduced in section 1 and thus might be considered as reliable association rules. The above mentioned processes consisted of modeling inter-event times in addition to the calculation of recall and precision measures. Since railway experts from Alstom have asked for a critical interval of [30 min , 24 hours], mining was focused on couples with inter-event times at least equal to 30 minutes. Table 4 shows the number of discovered couples abiding the inter-event time constraint.

Table 4. Number of significant event couples abiding the inter-event time constraint discovered by the UL, FL and FL(R) null models (p-value threshold = 1%) respectively in the TrainTracer™ data sequences

UL			FL			FL(R)		
W	C	D	W	C	D	W	C	D
632	1468	1335	446	1180	1079	806	579	471

Since the results obtained in 5.1 have shown that the FL null model with the C and D scores (FL-C and FL-D respectively) have preserved a high efficiency with various sequence lengths and types of bursts better than the other models, only the couples discovered by this model were further assessed and post-treated. Figure 2 shows the Recall/Precision scatter plots for couples discovered by FL-C and FL-D. Four zones can be defined according to 50% thresholds on both recall and precision.



(a) FL-C (b) FL-D

Fig. 2. Recall/Precision scatter plot of all couples with T_{AB} median value ≥ 30 minutes discovered by the FL null model with the C-score (FL-C) (a) and with the D-score (FL-D) (b). Recall threshold = 50%, Precision threshold = 50%

Event couples belonging to zone 1 are statistically very relevant and hence can be considered as plausible association rules with high interestingness. Zones 2 and 3 contain all couples with either a high recall or precision value. Couples belonging to these two zones are considered to be possibly relevant enough to be association rules because the weakness of one of the two measures might be a result of the complexities occurring in the data such as redundancy or bursts. Zone 4 contains couples that are considered insignificant due to their low recall and precision values. The scatterplots show that FL-C and FL-D have discovered the same interesting couples (977 common couples in total). Table 5 shows the number of discovered couples per zone.

Table 5. Number of event couples per zone discovered by the FL null model with the C-score (FL-C) and D-score (FL-D). Recall and Precision thresholds = 50%

Zone	*FL-C*	*FL-D*
1	0	0
2	8	10
3	143	115
4	1029	954

Due to the lack of the ground truth on the real existence of rules in the TrainTracer™ data extract under disposal, the analysis of the discovered association rules had to be both statistical and physical with the help of railway maintenance experts in order to identify among them those having a real physical meaning.

Consider the following association rule as an example:

Tilting system isolated by permanent tilt isolation switch $\longrightarrow$ Train tilt system defect
Recall: 52.1%, Precision: 47.8%

The recall value indicates that 52.1% of the "Train tilt system defect" events (a category 5 failure event) have been predicted by "Tilting system isolated by permanent tilt isolation switch" events (category 3 event). However, only 47.8% of the "Tilting system isolated by permanent tilt isolation switch" events have led to a "Train tilt system defect" within a time window of [30 min , 24 h]. This rule belongs to zone 3 with a high precision and low recall and has a real physical meaning as both events are related to the tilt process.

Both recall and precision values may be negatively effected by data bursts in a specific train or at a certain period of time where failures were frequent due to infrastructure-related factors, hardware/software problems, etc. That is why, prior to presenting the rule to technical experts, it is meaningful to consider the recall and precision values of the association rule per train as well as the distribution of the two events of the couple constituting the rule amongst trains over the 6-month observation period (example Figure 3). The observation of unusual distributions may decrease the chances of a rule to be credible.

In order to improve the robustness of the discovered rules, data should be further cleaned from redundancy and bursts. Also, increasing the length of rules to be discovered might reveal hidden valuable information. For instance, association rules between events that seem physically irrelevant may be explained more logically by extending to length-3 and more. For example, the two events in the above rule might be the extremities of a longer and more physically relevant one.

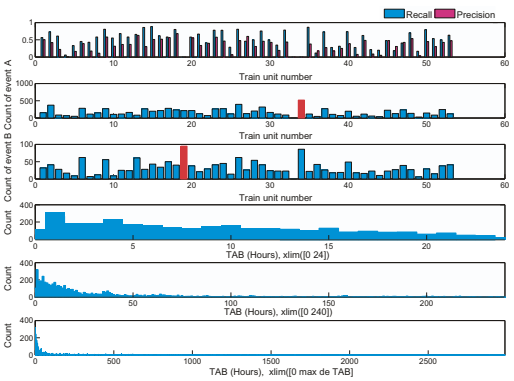


Fig. 3. Example of the distribution of Recall/Precision values of an association rule A $\longrightarrow$ B per train as well as the distribution of both events per train and histograms of T_{AB} values of the rule visible within variable time scales

6 Conclusion and Future Work

In this paper, a new methodology based on a significance testing approach for pairwise co-occurrences in bursty temporal data sequences is presented and applied. Three different null models with three different co-occurrence scores were discussed and confronted on both synthetic and real floating train data. The aim is to discover association rules leading to rare target events requiring immediate maintenance actions within very complex and challenging data sequences with multiple constraints. These rules, once integrated in an online analysis process of the incoming event stream will allow railway operators to predict severe failures in the future. The choice of a significance testing algorithm to derive association rules was mainly motivated by the rareness of temporal rules to be discovered within a large temporal sequence of events with bursts. Experiments were carried out on real floating train data extracts from Alstom's TrainTracer™ software. The obtained rules were evaluated using classical metrics such as recall and precision, as well as by railway maintenance experts in order to incorporate physical expertise into the analysis process. The preliminary results show a promising potential of null models to highlight rare but relevant co-occurrences between couples of events leading to significant association rules.

This work can be extended in several directions. Future work will mainly involve further investigations to render the null models more robust against bursty data in order to minimize their high false positive rate. Finally, increasing the length of discovered rules can be carried out to broaden the spectrum of results.

Acknowledgements. The authors wish to express their thanks to Alstom Transport for providing the TrainTracer™ data extracts. The authors particularly thank Kevin Prendergast and Charles-Eric Fonlladosa for their valuable help and fruitful discussions.

References

1. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proceedings of ACM SIGMOD 1993, pp. 207–216 (1993)
2. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: Proceedings of the 20th International Conference on Very Large Data Bases (VLDB 1994), pp. 487–499 (1994)
3. Bellwood, D., Wainwright, P., Fulton, C., Hoey, A.: Assembly rules and functional groups at global biogeographical scales. Functional Ecology 16, 557–562 (2002)
4. Cule, B., Goethals, B., Tassenoy, S., Verboven, S.: Mining train delays. In: Gama, J., Bradley, E., Hollmén, J. (eds.) IDA 2011. LNCS, vol. 7014, pp. 113–124. Springer, Heidelberg (2011)
5. Flier, H., Gelashvili, R., Graffagnino, T., Nunkesser, M.: Mining railway delay dependencies in large-scale real-world delay data. In: Ahuja, R.K., Möhring, R.H., Zaroliagis, C.D. (eds.) Robust and Online Large-Scale Optimization. LNCS, vol. 5868, pp. 354–368. Springer, Heidelberg (2009)
6. Gotelli, N.J., Graves, G.R.: Null models in ecology. Smisonian Inst. Press (1996)
7. Gotelli, N.: Null model analysis of species co-occurrence patterns. Ecology 81, 2606–2621 (2000)

8. Grahne, G., Wang, X., Laksgmanan, L.: Efficient mining of constrained correlated sets. In: International Conference on Data Engineering, p. 512 (2000)
9. Haiminen, N., Mannila, H., Terzi, E.: Determining significance of pairwise co-occurrences of events in bursty sequences. BMC Bioinformatics 9(1) (2008)
10. Hannenhalli, S., Levy, S.: Predicting transcription factor synergism. Nucleic Acids Res. 30(19) (2002)
11. Honda, S., Fukui, K., Moriyama, K., Kurihara, S., Numao, M.: Extracting human behaviors with infrared sensor network. In: Proceedings of the 4th International Conference on Networked Sensing Systems, INSS 2007, pp. 122–125 (2007)
12. Kerner, B., Demir, C., Herrtwich, R., Klenov, S., Rehborn, H., Aleksi, M., Haug, A.: Traffic state detection with floating car data in road networks. In: Proceedings of the International IEEE Conference on Intelligent Transportation Systems 2005, pp. 700–705 (2005)
13. Klein, H., Vingron, M.: Using transcription factor binding site co-occurrence to predict regulatory regions. Genome informatics. International Conference on Genome Informatics 18, 109–118 (2007)
14. Koh, Y.S., Pears, R., Yeap, W.: Valency based weighted asssociation rule mining. In: Zaki, M.J., Yu, J.X., Ravindran, B., Pudi, V. (eds.) PAKDD 2010, Part I. LNCS (LNAI), vol. 6118, pp. 274–285. Springer, Heidelberg (2010)
15. Levy, S., Hannenhalli, S., Workman, C.: Enrichment of regulatory signals in conserved non-coding genomic sequence. Bioinformatics 17(10), 871–877 (2001)
16. Liu, Y., Xu, W., Du, H.: The method of test for state of railway tunnel lining based on association rules (May 2011)
17. Magnusson, S.: Discovering hidden time patterns in behavior: T-patterns and their detection. Behakvior Research Methods, Instruments, & Computers 32s, 93–110 (2000)
18. Mannila, H., Toivonen, H., Verkamo, A.: Discovery of frequent episodes in event sequences. Data Mining and Knowledge Discovery 1, 259–289 (1997)
19. Mirabadi, A., Sharifian, S.: Application of association rules in iranian railways (rai) accident data analysis. Safety Science 48(10), 1427–1435 (2010)
20. Ng, R., Lakshmanan, L., Han, J., Pang, A.: Exploratory mining and pruning optimizations of constrained association rules. SIGMOD 27(2), 13–24 (1998)
21. Onboard diagnosis transforms maintenance: Alstom's onboard diagnosis system traintracer is changing the way maintenance is planned and conducted. International Railway Journal (2009)
22. Salah, A., Pauwels, E., Tavenard, R., Gevers, T.: T-patterns revisited: Mining for temporal patterns in sensor data. Sensors 10(8), 7496–7513 (2010)
23. Sammouri, W., Côme, E., Oukhellou, L., Aknin, P., Fonlladosa, C.E., Prendergast, K.: Temporal association rule mining for the preventive diagnosis of onboard subsystems within floating train data framework. In: Proceedings of the IEEE International Conference on Intelligent Transportation Systems, ITSC 2012, pp. 1351–1356 (2012)
24. Tan, P., Steinbach, M., Kumar, V., Potter, C., Klooster, S., Torregrosa, A.: Finding spatio-temporal patterns in earth science data. In: Proceedings of the KDD Workshop on Temporal Data Mining (2001)
25. Van Zuylen, H., Chen, Y., Zheng, F.: Using floating car data for traffic state estimation in signalized urban networks. In: IWTDCS Barcelona 2008 (2008)
26. Weiss, G.: Timeweaver: A genetic algorithm for identifying predictive patterns in sequences of events, pp. 718–725 (1999)
27. Zaki, M.: SPADE: An efficient algorithm for mining frequent sequences. Machine Learning Journal 42, 31–60 (2001)

Online Shopping Customer Data Analysis by Using Association Rules and Cluster Analysis

Serhat Güden and Umman Tugba Gursoy

Institute of Business Administration, Istanbul University, Istanbul, Turkey
serhatguden@yahoo.com, tugbasim@istanbul.edu.tr

Abstract. Data Mining is the process of exploration and analysis of large quantities of data in order to discover meaningful patterns and rules. Data mining is considered as the only solution towards efficient use of increasing amounts of data worldwide. The process of converting data into information is achieved by means of data mining. In this study, first the concept of data mining is presented, then CRISP-DM process are described. In this paper Cluster Analysis and Association Rules are used to analyze the data. k-means Algorithm, Confidence and Support Ratios are theoretically explained and these techniques applied to a data set obtained from 314 customers from 7 regions of Turkey to identify their profile.

Keywords: Data Mining, Cluster Analysis, Association Rules Analysis, k-means, A priori Algorithm.

1 Introduction

Data mining is a rapidly growing interdisciplinary area of tools for extracting models from, and identifying patterns in data. These utilize aspects from statistics, machine learning, neural networks, plus many other emerging methodologies. In order to comprehend the complexities of data mining, having an understanding of both mathematical modeling and computational algorithms is very important. [1] Without computational algorithms it would be impossible to mine the huge quantities of data being generated today.

Data Mining is the process of exploration and analysis of large quantities of data in order to discover meaningful patterns and rules. In the last few years Data Mining becomes widespread and recently, it has become more common and important. [2].

2 The Cross-Industry Standard Process for Data Mining (Crisp-Dm)

Data mining is the process of selection, exploration and modeling of large databases in order to discover models and patterns that are unknown a priori. [3]

A wide range of organizations in various industries are making use of data mining including manufacturing, marketing, chemical, aerospace to take advantages over

P. Perner (Ed.): ICDM 2013, LNAI 7987, pp. 127–136, 2013.

their competitors. The needs for a standard data mining process therefore increased dramatically. The data mining process must be reliable. In 1990, a Cross-Industry Standard Process for Data Mining (CRISP-DM) first published after going through a lot of workshops, and contributions from over 300 organizations.[4]

CRISP-DM consists of 6 phases:

1- Business Understanding
2- Data Understanding
3- Data Preparation
4- Modeling
5- Evaluation and
6- Deployment

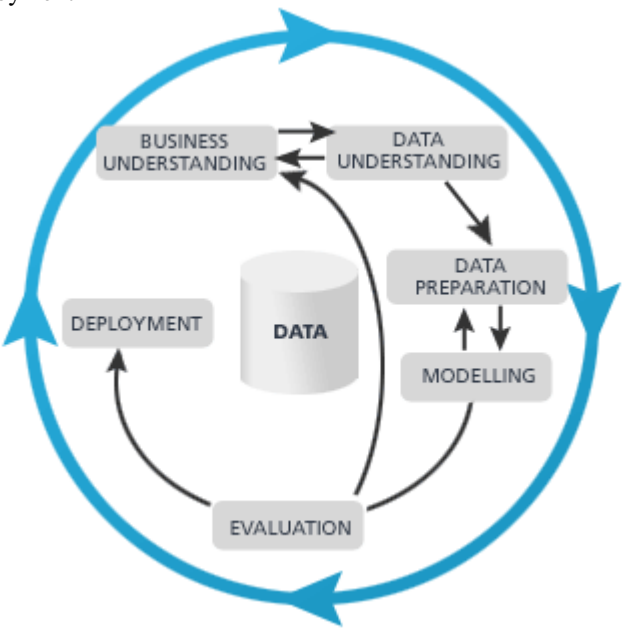


Fig. 1. The CRISP-DM Process

3 Association Rules Analysis

This data mining technique is used to identify the behavior, specific events or processes. Association discovery links occurrences within a single event. For example;

- Men who purchase premium brands of coffee are three times more likely to buy imported cigars than men who buy standard brands of coffee.

Retail stores use this data mining technique to find buying patterns in grocery stores. Association discovery is sometimes called market basket analysis.

The strength of an association rules measure with Support and Confidence Ratios.

Support and Confidence Ratios

Support shows the frequency of the patterns in the rule; it is the percentage of transactions that contain both A and B.

Support = Probability(A and B)

Support = (Transactions involving A and B) / (Total number of transactions).

Confidence is the strength of implication of a rule; it is the percentage of transactions that contain B if they contain A,

Confidence = Probability (B if A) = P(B/A)

Confidence = (Transactions involving A and B) / (Total number of transactions that have A).

4 Cluster Analysis

Clustering is the process of grouping the data into clusters, so that objects within a cluster have high similarity in comparison to one another but are very dissimilar to objects in other clusters.

Dissimilarities are assessed based on the attribute values describing the objects. Often, distance measures are used. Cluster analysis has been extensively studied for many years, focusing mainly on *distance- based cluster analysis*.

In data mining, efforts have focused on finding methods for efficient and effective cluster analysis in large databases. The most well-known and commonly used methods are *k-means, k-medoids* and their variations. In this paper k-means algorithm is used to analyze the data. [5]

4.1 K-Means Algorithm

The k-means algorithm takes the input parameter **k,** which is the number of clusters desired.

First **k** initial centroids are chosen. Each point is then assigned to the closest centroid, and each collection of points assigned to a centroid is a cluster. The centroid of each cluster then updated based on the points assigned to the cluster. The assignment is repeated and updated until no point changes clusters, or until the centroids remain the same. [6]

Basic k-means algorithm described below:

1- Select **k** points as initial centroids
2- Repeat
3- Form **k** clusters by assigning each point to the closest centroid
4- Recompute the centroid of each cluster
5- Until Centroids do not change.

5 Application

5.1 Problem Definition

This study was carried out for identifying online shopping customer behaviors by using cluster and association rules analysis.

5.2 Data Preparation

The first step in the data preparation stage is collecting the data. The data collection process was conducted by survey. The survey form consists of 14 questions, conveyed to 400 customers in 7 regions of Turkey via cargo and e-mail between 05.11.2012 and 23.11.2012. Due to different problems, 77 customers did not submit their Q&A forms. Hence, 323 customers have been analyzed in the study.

The data collected from 323 customers was carefully coded on SPSS. As a result of data cleaning process, 9 samples were evaluated as invalid and 314 data were analyzed.

The data set consists of 39 variables and 314 records. Data for 10 records is given in Table 1.

Table 1. - Data Set

	Gender	Age	Marital_status	Education	Frequency_of_shopping	Shopping_on_internet	Reliable	Limango	Morhipo
1	0	2	1	1	4	1	2	0	0
2	1	2	1	3	4	1	3	0	0
3	1	1	1	1	4	1	2	0	0
4	1	2	1	3	4	1	2	1	0
5	1	1	1	4	2	1	2	0	0
6	0	3	1	2	4	1	3	0	0
7	1	3	1	1	2	1	4	0	0
8	0	1	1	1	4	1	2	0	0
9	1	1	1	1	3	0	.	.	.
10	0	2	1	3	2	1	1	0	0

5.3 Modeling

The variables have been examined by using IBM Modeler.

Gender: 57.01% of the customers are Female and 42.99% are Male.

Value	Proportion	%	Count
1		57,01	179
0		42,99	135

Fig. 2. Distribution of Gender

Age:

63.06% of the customers are between26-35.
17.83% of the customers are between16-25.
14.33% of the customers are between36-45.
4.78% of the customers are 46 years old and above.

Marital Status:

47.45% of the customers are single and 52.55% are married.

Education:

61.46% of the customers have a BA degree.
18.15% of the customers are high school graduates.
10.19% of the customers have a graduate degree.
9.55% of the customers are two-year vocational school graduates.
0.64% of the customers are doctoral graduates.

Shopping Frequency:

57.32% of the customers shop when they had the opportunity.
22.93% of the customers shop once a week.
12.10% of the customers shop once in fifteen days.
7.64% of the customers shop once a month.

Do you do your shopping online?

87.12% of the customers shop online and 16.88% do not.

16.88% of the customers answered the question "Do you shop online?" (6th question) as NO. Therefore; 261 customers data analyzed and evaluated.

Is It safe to shop online?

46.82% of the customers agree with this idea. 24.84% is uncertain. There is 16.56% null value. Since it has been assumed that customers who do not shop online did not answer this question, it will not generate a problem during data analysis. There is an 8.6% of customers that find online shopping absolutely safe. 2.87% find online shopping unsafe and 0.32% find it absolutely unsafe.

Shopping Websites:

154 customers shop from Markafoni.com,
152 customers shop from Hepsiburada.com,
106 customers shop from Trendyol.com,
99 customers shop from Gittigidiyor.com,
91 customers shop from Limango.com,
68 customers shop from Sahibinden.com,
43 customers shop from Morhipo.com,
33 customers shop from Avon's website and
15 customers shop from Amway's website.

Among others, the websites with considerable customers are;

Biletix.com with 18 customers,
Mybilet.com with 6 customers,
And VIPdukkan.com with 8 customers.

What do you shop online for?

179 customers shop for Clothing,
139 customers shop for Electronics,
127 customers shop for Social Activity Tickets,

84 customers shop for Cosmetics,
56 customers shop for Other needs and
27 customers shop for Food.

Payment Method:

229 customers make their payment Online by Credit Card,
38 customers make their payment At the Door by Credit Card,
19 customers make their payment Cash at the Door and
14 customers make their payment by Money Order.

Number of Installments:

94 customers choose Single Payment,
85 customers choose 4-6 Installments,
82 customers choose 2-3 Installments,
34 customers choose 10-12 Installments and
10 customers choose 7-9 Installments.

It can be seen that the installment options of customers vary between 1-6 installments. Customers prefer to make short-term payments.

Special Offers:

75.48% of the customers choose online shopping due to prices being below the market.

Problems after Shopping:

144 customers have encountered Late Delivery problems,
72 customers have encountered Missing Items,
63 customers have not encountered any problems,
42 customers have faced with Customer Service Negligence.

It is apparent that shopping websites will have significantly overcome customer dissatisfaction by investing in the logistics area for solving the "Late Delivery" issue.

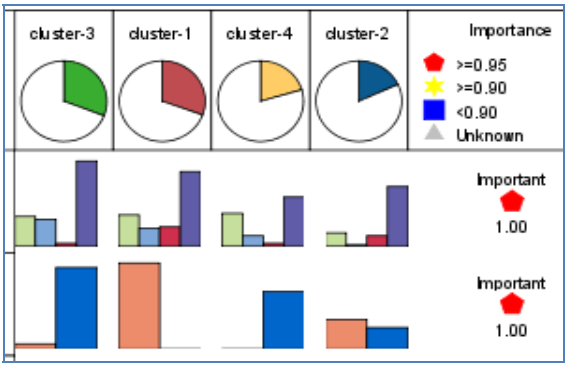
Income:
31.53% of the customers have an income over 2.401 TL,
19.43% of the customers have an income below 1.300 TL,
17.52% of the customers have an income between 1.301 – 1.800 TL,
17.52% of the customers have an income between 1.801 – 2.400 TL and
14.01% of the customers are not included in the income section.

5.4 Modeling

5.4.1 Cluster Analysis

For cluster analysis, the most commonly used k-means algorithm was chosen. As a result of performed tests, the cluster number was approved as 4. The acquired output is given in Table 2.

Table 2. Cluster Analysis



When Table 2 examined, cluster profiles can be identified as follows:

Cluster 1: Single males with BA degrees, with an age range of 26-35 who have over 2401 TL income are included in this cluster. The customers in this cluster tend to shop online when they have the opportunity.

Cluster 2: High school graduate married males with an age range of 36-45 who have over 2401 TL income are included in this cluster. The customers in this cluster tend to shop online when they have the opportunity.

Cluster 3: Single females with BA degrees, with an age range of 26-35 who have an income below 1300 TL are included in this cluster. The customers in this cluster also tend to shop online once a week.

Cluster 4: Married females with BA degrees, with an age range of 26-35 who have over 2401 TL income are included in this cluster. The customers in this cluster tend to shop online when they have the opportunity.

5.4.2 Association Rules Analysis

The data set collected via questionnaires coded as 0-1 for Association Rules Analysis. Data set can be seen in Table 3.

Table 3. Data Set

	A	B	C	D	E	F	G	H	I
1	lim	morh	mark	trend	avon	heps	gitti	sahib	amway
2	0	0	1	1	1	0	1	0	0
3	0	0	0	0	0	0	0	0	0
4	0	0	1	0	0	0	0	0	0
5	1	0	1	0	0	0	0	0	0
6	0	0	1	0	0	0	0	1	0

IBM Modeler program and Apriori algorithm was used to analyze the data.

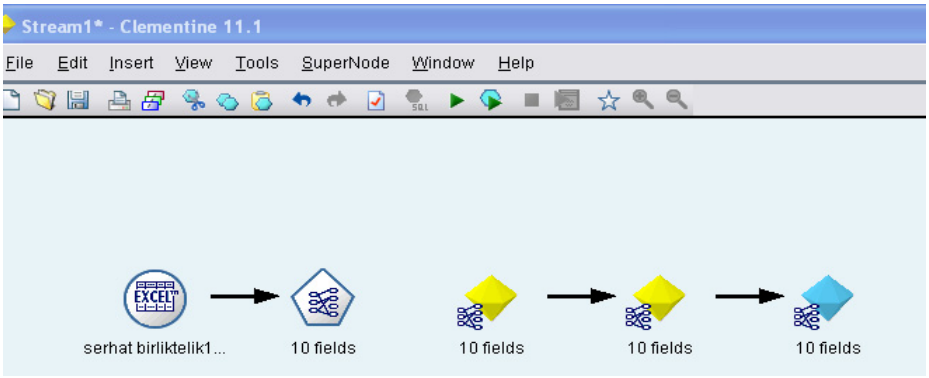


Fig. 3. Model

The validity of the rule sets obtained by Association Rules Analysis, evaluated by Support and Confidence Ratios. For this application Support ratio selected as 15% and Confidence ratio selected as 60%.

Rule sets can be seen in Table 4.

Table 4. Rule sets

10 fields

File Generate

Sort by: Confidence % 18 of 18

Consequent	Antecedent	Support %	Confidence %
markafoni	limango trendyol	17,516	96,364
markafoni	trendyol hepsiburada	19,745	90,323
markafoni	trendyol	33,758	87,736
markafoni	limango hepsiburada	16,242	82,353
markafoni	limango	28,981	81,319
hepsiburada	gittigidiyor markafoni	20,701	80,0
hepsiburada	gittigidiyor	31,529	79,798
trendyol	limango markafoni	23,567	71,622
hepsiburada	sahibinden.com	21,656	70,588
gittigidiyor	sahibinden.com hepsiburada	15,287	68,75
markafoni	gittigidiyor hepsiburada	25,159	65,823
markafoni	gittigidiyor	31,529	65,657

First rule set: 96,364% of the customers who shop on Limango and Trendyol website, also shop on Markafoni website. The support ratio for this rule is 17,516%.

7th rule set: 79,798% of the customers who shop on **Gittigidiyor** website, also shop on hepsiburada website. The support ratio for this rule is 31,529%.

6 Conclusions

In the last 10 years, markets have become more globalized due to improvements in the communication technologies. It is possible to say that retailers who intend to compete in the global markets have quickly become aware of the internet's importance and begun operating in this area. Significant and continuing changes in technology have affected markets in the same direction and e-commerce has rapidly become popular. Its development over time shows that e-commerce has usually been included in an integrated channel strategy in both Turkish and global markets.

Table 5. Internet Usage Rate in Turkey (www.internetworldstats.com)

YEAR	Users	Population	% Pop.
2000	2,000,000	70,140,900	2.9 %
2004	5,500,000	73,556,173	7.5 %
2006	10,220,000	74,709,412	13.9 %
2010	35,000,000	77,804,122	45.0 %

In spite of the e-commerce sector's rapid growth, there are still negative impressions by some customers concerning safety issues. The result of our study says that 46.82% of online shoppers agree with the internet's safety and 8.60% strongly agrees with it. Therefore, when we adapt survey samples to the main mass, it appears that 55.42% of customers in Turkey find online shopping safe. Observations regarding world-wide safety are also included in the result of our study. Results show that 44.58% find internet usage unsafe. E-commerce companies should consider this 44.58% group and search for answers.

Results also show that marital status and gender variables have a significant effect on online shopping. 90 single female customers and 89 married female customers show that women have an important effect on online shopping. The study results prove that e-commerce companies have substantially expanded their range of products for women. It is recorded that the use of internet shopping increases after marriage among male customers. While the number of single male customers is 59, there are 76 married male customers.

It is expected that male customers prefer websites which include both buying and selling. 49 male customers shop on Sahibinden.com; and 80 of them shop for electronics. There is not a big difference between female customers who shop for cosmetics and who do not.

Another important result of the study is the social status of customers. Male customers shop online for event tickets and electronics needs. Female customers usually shop online for books and general needs.

According to survey results, online shopping customers are usually married with a high level of income who work in the private sector. Businesses that have e-markets can develop their strategies based on this profile. Majority of customers also complain

about late delivery due to weak logistic networks of e-tailing companies. Companies can increase their revenues by minimizing this problem. Special offers for interest-free payments on short-term (1-6 months) installment options for pos machines contracted to banks or special offers that will increase business profit in short-term installments will increase customer numbers as well.

References

1. Hand, D., Mannila, H., Smyth, P.: Principles of Data Mining. Artificial Intelligence in Medicine 26(1-2), 175–178 (2002)
2. Berry, M.J.A., Linoff, G.S.: Mastering Data Mining, p. 7. Wiley (2000)
3. Giudici, P.: Applied Data Mining, p. 1. Wiley & Sons (2003)
4. Data Mining Processes (December 6, 2012), http://www.zentut.com/data-mining/data-mining-processes/
5. Han, J., Kamber, M.: Data Mining Concepts and Techniques, p. 228. Morgan Kaufmann (2006)
6. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining, pp. 327–497. Pearson Education (2006)

A Study on Multi-label Classification

Clifford A. Tawiah and Victor S. Sheng

Department of Computer Science, University of Central Arkansas,
Conway, Arkansas, USA 72035
{ctawiah2,ssheng}@uca.edu

Abstract. Multi-label classifications exist in many real world applications. This paper empirically studies the performance of a variety of multi-label classification algorithms. Some of them are developed based on problem transformation. Some of them are developed based on adaption. Our experimental results show that the adaptive Multi-Label K-Nearest Neighbor performs the best, followed by Random k-Label Set, followed by Classifier Chain and Binary Relevance. Adaboost.MH performs the worst, followed by Pruned Problem Transformation. Our experimental results also provide us the confidence of existing correlations among multi-labels. These insights shed light for future research directions on multi-label classifications.

Keywords: multi-label classification, Multi-Label K-Nearest Neighbor, Random k-Label Set, Adaboost.MH, Classifier Chain, Binary Relevance, Pruned Problem Transformation.

1 Introduction

Multi-label classifications deal with multiple labels being assigned to every instance in a dataset. That is, an instance can be assigned more than one class simultaneously. It is concerned with learning a model that outputs a bipartition of a set of labels into relevant and irrelevant with respect to a query instance. This type of classification differs in some respect from traditional single label classifications in that one of multiple labels is allocated to an instance in the dataset. In single-label classifications each instance is associated with a single label and a classifier learns to associate each new test instance with one of these known labels [1, 4].

Multi-label classification tasks exist in many real-world applications, such as, gene classification in bioinformatics [8], medical diagnosis, document classification, music annotation, image recognition, and so on. All these applications require effective and efficient multi-label classification algorithms. There exist a variety of multi-label classification algorithms [10]. For data mining practitioners, it is very important for them to know the general knowledge on which algorithms perform the best, such that they can try to apply it first. For data mining researchers, it is important to investigate the performance of the existing algorithms to get insights, such that they can use the clue to guide their research on developing more effective multi-label classification algorithms.

P. Perner (Ed.): ICDM 2013, LNAI 7987, pp. 137–150, 2013.

Current existing multi-label classification algorithms are developed based on two basic approaches: algorithm adaptation and problem transformation. Problem transformation is easy to understand. We discuss it first.

Before we discuss the procedure of problem transformation, let us review traditional classifications a bit. On the contrast, traditional classifications can be called single label classifications. Giving a set L of labels, traditional single label classifications choose one label from the set to assign to an instance. If $|L| = 2$, then the problem is binary classification. Otherwise, if $|L| > 2$, it becomes a multi-class classification problem. On the contrary, multi-label learning is to assign multiple different labels to a test instance simultaneously [9].

Problem transformation is to transfer multi-label classifications into multiple traditional single label classifications, specifically, multiple binary classifications. Figure 1 shows an example of the process of transferring a multi-label classification (movie classification) into multiple binary classifications (yes or no).

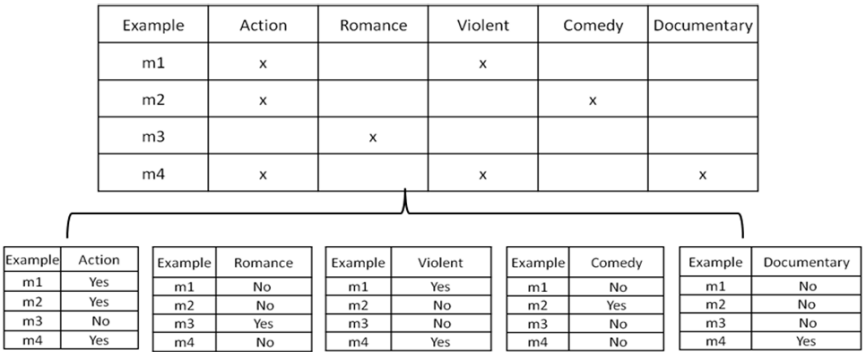


Fig. 1. An example of multi-label problem transformation

After a multi-label classification problem is transferred into multiple binary classification ones. All the traditional classification algorithms can be applied directly to build a classifier for each binary dataset and make prediction for its correlated test instances. The prediction for a multi-label instance is made by aggregating outputs from autonomous binary classifiers. Binary Relevance (BR) [3, 11], Classifier Chain [3], Random *k*-Label Set [7], and Pruned Problem Transformation [12] are the examples of classifiers, which use problem transformation. We will briefly review these algorithms in the following section and make comparison among them in Section 4.

Different algorithms have their method for classifying multi-label instances. For example a classifier which employs problem transformation for its classification may apply a probability distribution over the transformed dataset to rank and assign labels to the test instances. In ranking, the task is to order labels, so that the topmost labels have a greater probability to assign to the test instance. This way the class with the highest probability will be ranked first, the class with the second best probability will be ranked second, and so on.

The second method used in multi-label classifications is algorithm adaptation. It extends existing traditional classification algorithms to perform multi-label classifications directly, for example, Adaboost.MH [15] and Multi-label K-Nearest Neighbor (MLKNN) [5]. Adaboost.MH is an adaptation of Adaboost [13] for multi-label classifications. MLKNN is an adaption of the traditional k-NN algorithm for multi-label classifications.

Algorithm adaption completely differs from problem transformation in that no binary transformation made. Instead, the algorithm learns the structure and co-relations that exist among labels to classify a test instance after being trained. Thus, it is very useful to investigate the performance of multi-label classification algorithms, which are developed based on the two approaches. The investigating results will guide data mining researchers in their future research on developing better multi-label classification algorithms.

The rest of the paper is organized as follows. Section 2 introduces the popular multi-label classification algorithms which we will make comparison empirically. Section 3 introduces five popular performance metrics specifically designed for multi-label classifications. In Section 4, we describe the experiments we have conducted. They consist of the setting of the experiments, the experimental results, and the analysis of the experimental results. Section 5 concludes with a summary of our work and a discussion of future work.

2 Popular Multi-label Classification Algorithms

We briefly review the six popular multi-label classification algorithms (i.e. Binary relevance [3, 11], Classifier Chain [3], Random K-Label Set [7], Pruned Problem Transformation [12], AdaBoost.MH [15], and Multi-label K-Nearest Neighbor [5]) in this section, which are used in our experiments in Section 4.

2.1 Binary Relevance (BR)

As we introduced before, Binary Relevance [3, 11] is one of the popular problem transformation approaches. It transfers a multi-label classification into multiple binary classifications (Figure 1). After the transformation, a traditional classification algorithm is applied to build multiple binary classifiers. Each classifier is responsible for predicting the presence or absence of each corresponding label which belongs to *L*. For example, if the total number of labels for a dataset is 70, then 70 binary classifiers would be trained for each label, with a 0 or 1 association. When classifying a new instance, this approach outputs the union of the labels that are predicted by all the binary classifiers.

2.2 Classifier Chain (CC)

Classifier Chain (CC in short) [3] is also a problem transformation method for multi-label classifications. It combines the computational efficiency of binary relevance and

label dependency for classifications [3]. Classifiers are linked along a chain, where each classifier deals with the binary relevance problem associated with a label in |L|. This is how classifier chain works. When a training set is passed to the algorithm, it creates a chain of classifiers C_1, C_2, C_3, ..., $C_{|L|}$, where |L| is the total number of labels for the dataset. Each of the multi-labels of the dataset is transformed into a binary problem. Thus, each classifier (C_1, ..., $C_{|L|}$) is responsible for learning and predicting binary associations (0 or 1) for each label. If a test instance *X* is introduced to the trained model, the classification process for *X* starts from C_1 and runs down along the chain of classifiers. Each classifier determines the probability of *X* to be classified into L_1, L_2, L_3, ..., $L_{|L|}$. It sort of builds a binary tree, where each link in the chain is extended with a 0 or 1 label association. This chaining method passes label information between classifiers, allowing CC to take into account label correlations and thus overcoming the label independence problem.

2.3 Random *k*-Label Set (RA*k*EL)

RA*k*EL [7] was proposed to solve performance issues of Label Powerset (LP). LP is a simple problem transformation method, but time consuming. It produces all subsets *LP(L)* of the multi-label set *L*, and treats each subset as a label. Then, it transforms the multi-label classification into traditional single label classification. Because LP produces a great number of labels, it could cause imbalanced issues, considering the number of labels versus the number of instances. In addition, the size of the generated labels incurs considerable computational costs [7]. RA*k*EL improves it by randomly selecting *k*-sized label sets from *L*, and then performing the typical LP approach on these generated subsets.

2.4 Pruned Problem Transformation (PPT)

Pruned Problem Transformation [12] is also an improvement of LP. The only improvement is that PPT prunes away the power subsets *LP(L)* that occur fewer times than a small user-defined threshold (usually 2 or 3). Removing certain information might skew or cause information to be lost. In order to avoid this issue, PPT optionally split each of the multi-label set into subsets. Thus, the subsets could occur more than the user-defined threshold.

2.5 AdaBoost.MH (AD)

Adaboost [13] is a short form of adaptive boosting. Boosting is a meta-algorithm, meaning it can be used in conjunction with other learning algorithms for improving their performance. It combines inaccurate and rough rules to produce accurate results. Given a base learning algorithm, Adaboost works by initially setting the weights of all training instances to be equal. Then it calls the base learning algorithm several times. For each call, the weight of incorrectly classified instances is increased. This is to help the base learning algorithm focus on the misclassified instance until it is correctly classified. AdaBoost.MH [15] is an adaptation of Adaboost for multi-label classifications. Adaboost.MH works similarly like the adaboost algorithm, except that

it breaks the multi-label classification down into a binary problem where each test instance is classified according to its label association (either 0 or 1).

2.6 Multi-Label *K*-Nearest Neighbor (MLKNN)

MLkNN [5] is an adaption of the traditional k-NN algorithm for multi-label classifications. It is one of lazy learning algorithms. The algorithm identifies, for each unseen test instance, the k nearest neighbors in the training set. It calculates prior probabilities from the k nearest training instances, and then finds the maximum posteriori probability to determine the label set for the test instance.

3 Performance Metrics

The performance evaluation of multi-label classification is much more complicated than traditional classification. In this Section, we introduce five popular performance metrics specifically designed for multi-label classifications [6, 15, 16], i.e., Hamming Loss, Average Precision, One-Error, Coverage, and Ranking Loss. Before describing their definitions, we explain the related notations first. Supposing there is a multi-label dataset D = $\{(x_i, Y_i) \mid 1 \leq i \leq p\}$, notations presented in the formulas below are: $h(x_i)$ represents a set of proper labels for x_i ; Δ represents symmetric difference; $h(x_i, y)$ represents the value of confidence for y to be a label of x_i; $rank^h(x_i, y)$ returns the rank of y from $h(x_i, y)$; and $\overline{y_i}$ represents the complementary of y_i [14].

3.1 Hamming Loss (HL)

Hamming Loss takes into account prediction errors, which labels are incorrectly predicted and which labels were not predicted at all [14]. It takes into account how many instance-label pairs are misclassified. The smaller the value, the better the performance is.

$$\mathrm{HL} = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{|y|}|h(x_i)\Delta y_i| \tag{1}$$

3.2 Average Precision (AP)

Average Precision evaluates the average fraction of labels ranked above a particular label which belongs to L [14]. It is often used for the evaluation of information retrieval tasks. The bigger the value of Average Precision, the better the classification performance is.

$$\mathrm{AP} = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{|y_i|}\bullet\frac{|P_i|}{rank^h(x_i, y)}, \text{ where} \tag{2}$$

$$P_i = \{y' \mid rank^h(x_i, y^i) \leq rank^h(x_i, y), y^i \in y_i\}$$

3.3 One-Error (OE)

This measure calculates the number of times that the top-ranked label predicted is not in the original label set of an instance. So it checks whether the top-ranked label is relevant, and ignores the relevancy of all other labels. The smaller the value of One- Error, the better the performance is.

$$\mathrm{OE} = \frac{1}{p}\sum_{i=1}^{p}[[\arg\max_{y\in Y} h(x_i, y)] \notin y_i] \tag{3}$$

3.4 Coverage (CV)

Coverage measures how far we need, on average, to go down the list of labels in order to cover all the possible labels assigned to an instance [14]. The goal of coverage is to assess the performance of a classifier for all the possible labels of instances. The smaller the value of Coverage, the better the performance is.

$$\mathrm{CV} = \frac{1}{p}\sum_{i=1}^{p}\max rank^{h}{}_{y\in Y_i}(x_i, y) - 1 \tag{4}$$

3.5 Ranking Loss (RL)

Ranking Loss computes the average fraction of label pairs which are not correctly ordered [14] for an instance. The smaller the value of Ranking Loss, the better the performance is.

$$\mathrm{RL} = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{\left|Y_i\right|\left|\overline{Y_i}\right|}\bullet\left|R_i\right|, \text{ where} \tag{5}$$

$$R_i = \{(y_i, y_2) \mid h(x_i, y_1) \le h(x, y_2), (y_1, y_2) \in y_i \times \overline{y_i}\}$$

4 Experiments

In this section, we will make thorough comparisons among the six multi-label classification algorithms introduced in Section 2, by applying them on eleven datasets. We evaluate their performance using the five popular metrics described in Section 3.

4.1 Experimental Setup

In our experiments, we try to conduct experiments on all available datasets listed in MULAN [10] website[1]. We have not obtained experimental results for some datasets,

[1] http://mulan.sourceforge.net/datasets.html

because of the limitation of our computer memory. The specifications of the computer used is a 64-bit Operating System, x64-based processor, Intel(R) Core(TM) i5-2467M with 8GB memory. We succeeded in the eleven multi-label datasets: Cal500 (human-generated musical annotations), Corel5k (learning a lexicon for a fixed image recognition), Emotions (music emotion detection), Enron (email messages) [2], Genbase (classification of protein families), Medical (variables consisting of illnesses and treatments), Scene (semantic indexing of still images), Yeast (gene function classification), Bookmarks (text tagging suggestion), Mediamill (multimedia analysis), and Bibtex (text tagging suggestion). The detail characteristics of the eleven datasets are listed in Table 1.

Table 1. Description of the datasets used in the experiments

Name	#Instances	#Training Inst.	#Test Inst.	Nominal	Numeric	Labels
Cal500	502			0	68	174
Corel5k	5000	4501	499	499	0	374
Emotions	593	391	202	0	72	6
Enron	1702	1123	579	1001	0	53
Genbase	662	463	199	1186	0	27
Medical	978	333	645	1449	0	45
Scene	2407	1211	1196	0	294	6
Yeast	2417	1500	917	0	103	14
Bookmarks	87856			2150	0	208
Mediamill	43907			0	120	101
Bibtex	7395	4880	2515	1836	0	159

Each dataset comes along with an xml header file specifying the names of the labels and hierarchical relationships among them [10]. Except Cal500, Bookmarks, and Mediamill, eight datasets in Table 1 are already separated into training and test sets. For each of these datasets, we loaded two files, an XML file, and the ARFF files for the train and test set.

We conduct experiments on the six classification algorithms described in Section 2 for each dataset. If a dataset is separated into a training and test set already, we only run each classification algorithm once on its training set, and report its performance over its test set. We have to explain how we conduct experiments on Bibtex. Because of memory limitation, we could not obtain experimental results using its original training and test set directly. Thus, we have to sample our training set (2000 instances, the maximum number of instances our computer can handle) and testing set (1200 instances) from its original training and test set respectively to conduct experiments.

There are three datasets: Cal500, Bookmarks, and Mediamill, which are not separated into train and test sets. We resample them using randomization and repeat the process ten times. The average results are presented in the paper. For Cal500, we split its whole dataset into 70% for training and 30% for testing each time. For Bookmarks and Mediamill, we sample 2000 instances as the training set and 1200

instances as the test set each time. Again, the size 2000 is the proper number of instances that our computer can process.

Notice that in our experiment the default base learner is used for the six multi-label classification algorithms. Specifically, J48 is used in conjunction with Binary Relevance and Classifier Chain. LabelPowerset, in conjunction with J48 is used as the base learner for RA*k*EL. J48 pruning tree is used for Pruned Problem Transformation. Adaboost.MH uses AdaBoostM1 as its base learner, which in turn uses decision stump as its base learner [17].

4.2 Experimental Results

Tables 2 through 9 show the experimental results of all of the six multi-label classification algorithms on eight datasets, i.e., Emotions, Enron, Genbase, Medical, Scene, Yeast, Mediamill, and Bibtex. Tables 10 through 12 show the experimental results of some of the six multi-label classification algorithms on three datasets: i.e., Cal500, Corel5k, and Bookmarks. This is because some of the algorithms run out of memory. For Cal500, we will show the experimental results for the algorithms except PPT. For Corel5k, we will show the experimental results for the algorithms except Adaboost.MH. For Bookmarks, we will show the experimental results for the algorithms except PPT and RA*k*EL. Based on the experimental results, we highlight the best in bold, highlight the second in italic, and underline the worst, for each of the five performance metrics for each dataset. Again, only for Average Precision, a higher value is better. For the rest four metrics, a lower value is better.

Table 2. Experimental results on Emotion

	HL	AP	OE	CV	RL
CC	0.2896	0.6726	0.4455	2.8267	0.3383
RAKEL	*0.2228*	*0.7841*	*0.2871*	*2.0545*	*0.1841*
AD	0.3160	0.5906	0.5248	3.0842	0.4295
PPT	0.3127	0.6928	0.4208	2.6832	0.3135
BR	0.2599	0.6959	0.3663	2.8069	0.3103
MLKNN	**0.2087**	**0.7965**	**0.2822**	**1.8762**	**0.1586**

Table 3. Experimental results on Enron

	HL	AP	OE	CV	RL
CC	0.0530	0.5722	0.4162	*23.2919*	*0.1794*
RAKEL	**0.0509**	*0.6051*	*0.2815*	24.7841	0.2030
AD	0.0619	0.4574	0.4594	27.7789	0.2370
PPT	0.0727	0.4703	0.5043	27.8152	0.2613
BR	0.0540	0.5746	0.4059	24.7997	0.1861
MLKNN	*0.0514*	**0.6345**	**0.2798**	**13.1813**	**0.0934**

Table 4. Experimental results on Genbase

	HL	AP	OE	CV	RL
CC	**0.0011**	*0.9918*	0.0050	*0.3166*	*0.0018*
RAKEL	**0.0011**	0.9900	*0.0101*	**0.2965**	**0.0016**
AD	0.0456	0.3184	0.7286	14.0704	0.5281
PPT	*0.0026*	0.9864	**0.0000**	0.5176	0.0063
BR	**0.0011**	*0.9918*	0.0050	*0.3166*	*0.0018*
MLKNN	0.0052	**0.9931**	**0.0000**	0.5779	0.0062

Table 5. Experimental results on Medical

	HL	AP	OE	CV	RL
CC	**0.0103**	**0.8268**	**0.1907**	4.4341	0.0756
RAKEL	0.0113	0.8126	0.2031	*4.0946*	*0.0732*
AD	0.0276	0.3571	0.7395	13.8512	0.2858
PPT	0.0173	0.7476	0.2713	5.9364	0.1051
BR	*0.0106*	*0.8226*	*0.1969*	4.4450	0.0763
MLKNN	0.0188	0.7266	0.3535	**3.5442**	**0.0586**

Table 6. Experimental results on Scene

	HL	AP	OE	CV	RL
CC	0.1392	0.7295	0.3712	1.3094	0.2365
RAKEL	*0.1150*	*0.8150*	*0.2977*	*0.6873*	*0.1166*
AD	0.1810	0.4302	0.7860	2.5819	0.4898
PPT	0.1623	0.7248	0.4130	1.1714	0.2134
BR	0.1389	0.7115	0.4264	1.2809	0.2307
MLKNN	**0.0953**	**0.8513**	**0.2425**	**0.5652**	**0.0925**

Table 7. Experimental results on Yeast

	HL	AP	OE	CV	RL
CC	0.2638	0.6295	0.3490	9.0349	0.3286
RAKEL	*0.2328*	*0.7102*	0.2901	*7.6696*	*0.2240*
AD	0.2330	0.5930	*0.2497*	9.2454	0.3821
PPT	0.2947	0.6470	0.3391	8.5627	0.3130
BR	0.2588	0.6164	0.4024	9.2857	0.3206
MLKNN	**0.1980**	**0.7574**	**0.2421**	**6.3642**	**0.1707**

Table 8. Experimental results (in average) on Mediamill

	HL	AP	OE	CV	RL
CC	0.0518	0.4324	0.5086	*45.1207*	*0.1661*
RAKEL	0.0481	*0.4969*	0.3147	46.0259	0.1745
AD	**0.0382**	0.4503	*0.2543*	59.2112	0.2395
PPT	0.0448	0.4222	0.3362	54.9310	0.2380
BR	0.0534	0.3757	0.6336	51.9397	0.2107
MLKNN	*0.0404*	**0.5682**	**0.2457**	**29.8017**	**0.1060**

Table 9. Experimental results on sampled Bibtex

	HL	AP	OE	CV	RL
CC	0.0145	**0.3784**	0.5415	*66.2824*	**0.2496**
RAKEL	**0.0136**	0.3602	**0.5183**	72.6379	0.3097
AD	0.0149	0.0859	0.8505	104.302	0.5181
PPT	0.0196	0.2938	0.6445	76.4153	0.3393
BR	0.0141	*0.3781*	*0.5316*	68.3189	*0.2615*
MLKNN	*0.0138*	0.2753	0.6611	**66.2425**	0.2744

Table 10. Experimental results (in average) on Cal500

	HL	AP	OE	CV	RL
CC	0.1789	0.3053	0.7616	169.8146	0.3746
RAKEL	0.1700	*0.3829*	0.3709	*166.3377*	*0.2983*
AD	*0.1423*	0.2319	**0.0927**	169.3974	0.5642
BR	0.1659	0.3198	0.8013	170.1391	0.3466
MLKNN	**0.1375**	**0.4916**	*0.1060*	**128.9934**	**0.1823**

Table 11. Experimental results on Corel5k

	HL	AP	OE	CV	RL
CC	0.0101	0.2392	0.7240	161.74	0.1826
RAKEL	*0.0096*	0.1296	0.7260	333.23	0.6381
PPT	0.0163	0.1767	0.8040	268.01	0.4332
BR	0.0098	*0.2550*	*0.7080*	*124.91*	*0.1429*
MLKNN	**0.0093**	**0.2656**	**0.7060**	**113.04**	**0.1297**

Table 12. Experimental results (in average) on bookmarks

	HL	AP	OE	CV	RL
CC	0.0112	0.2159	0.8168	84.1414	0.2834
AD	*0.0104*	0.1166	0.9529	109.528	0.4230
BR	0.0118	*0.2299*	*0.7853*	*82.5759*	*0.2818*
MLKNN	**0.0096**	**0.2423**	**0.7644**	**80.5183**	**0.2704**

In order to see clearly and to show the general knowledge of the performance of the six multi-label classification algorithms, Table 13 shows the average values of the five performance metrics over the eight datasets (Emotions, Enron, Genbase, Medical, Scene, Yeast, Mediamill, and Bibtex), from Table 2 to 9. The experimental results of other three datasets (Cal500, Corel5k, and Bookmarks) are not included in Table 13, because we did not obtain results for some of the six multi-label classification algorithms, which ran out of memory. Partial experimental results for the three datasets are shown in Tables 10 through 12.

We further summarize the comparisons among the six multi-label classification algorithms through ranking over the eight datasets. For each of the eight datasets, we ranked the performance of the six algorithms from 1 (best) to 6 (worst) on each metric. The average rank of the six algorithms on each metric is shown in Table 14. Further, we average the average ranks for each of the six algorithms across the five performance metrics in the last column of Table 14. Again, the experimental results of other three datasets (Cal500, Corel5k, and Bookmarks) are not included.

In Table 13 and 14, we also highlight the best in bold, highlight the second in italic, and underline the worst for each of the five performance metrics and the overall average ranks (Table 14 only) for the six algorithms.

Table 13. Average Results of the eight datasets for different algorithms

	HL	AP	OE	CV	RL
CC	0.1029	0.6541	0.3534	*19.077*	0.1969
RAKEL	*0.0869*	*0.6967*	**0.2753**	19.781	*0.1608*
AD	0.1147	0.4103	0.5741	29.265	0.3887
PPT	0.1158	0.6231	0.3661	22.254	0.2237
BR	0.0988	0.6458	0.3710	20.399	0.1997
MLKNN	**0.0789**	**0.7003**	*0.2883*	**15.269**	**0.1200**

Table 14. Average ranks of different algorithms on the eight datasets

	HL	AP	OE	CV	RL	**Average**
CC	3.375	3.00	3.5	3.125	3.125	3.225
RAKEL	**2.0**	*2.625*	*2.375*	*2.375*	*2.5*	*2.375*
AD	4.5	4.5	4.625	5.75	5.75	5.025
PPT	4.75	4.125	3.625	4.375	4.375	4.25
BR	3.25	3.5	3.625	4.0	3.25	3.525
MLKNN	**2.0**	**2.0**	**2.0**	**1.5**	**1.5**	**1.8**

Table 14 shows that MLKNN performs the best. Its average ranks on all five performance metrics are the best (lowest values). Its overall rank value across the five performance metrics is 1.8, much lower than the second position RA*k*EL, whose overall rank value is 2.375. This is also supported by its ranks on each of the eight datasets, from Table 2 to 9. The average performance over the eight datasets shown in Table 13 also supports this. Table 13 shows that MLKNN achieves the lowest values

on the lowest-best metrics: Hamming Loss (HL), Coverage (CV), and Ranking Loss (RL), and achieves the highest value on the highest-best metric Average Precision (AP). Although we did not include the experimental results of three datasets (Cal500, Corel5k, and Bookmarks) into the summarization, their experimental results shown in Tables 10 through 12 completely support the conclusion.

Table 14 shows that RA*k*EL occupies the second position on all performance metrics, including the tied best in Hamming Loss (HL). Its overall rank 2.375 is also in the second position. Thus, we can conclude that RA*k*EL is the second best among the six algorithms. This is supported by its ranks on each of the eight datasets, from Table 2 to 9. The average performance over the eight datasets shown in Table 13 also supports this. Table 13 shows that RA*k*EL achieves the lowest value on the lowest-best metric One-Error (OE). It also achieves the second lowest in the lowest-best metrics: Hamming Loss (HL) and Ranking Loss (RL), and achieves the second highest value on the highest-best metric Average Precision (AP).

Table 14 shows that Adaboost.MH (AD) performs the worst. Its average ranks on all five performance metrics are the worst (highest values). Its overall rank value across the five performance metrics is 5.025, close to the maximum value 6.0. This is supported by its ranks on each of the eight datasets, from Table 2 to 9. The average performance over the eight datasets shown in Table 13 also supports this. Table 13 shows that AD achieves the highest values on the lowest-best metrics: One-Error (OE), Coverage (CV), and Ranking Loss (RL), and achieves the lowest value on the highest-best metric Average Precision (AP).

Table 14 shows that Classifier Chain (CC), Binary Relevance (BR), and Pruned Problem Transformation (PPT) take the middle positions. We can see that CC is the best, followed by BR, followed by PPT, among the three algorithms. Table 13 also shows this relationship. CC combines the binary relevance and the multi-label dependency. The better performance from CC shows that the label dependency exists, and needs to be utilized in multi-label classifications. That RA*k*EL performs the second also supports this. The potential drawback of Binary Relevance is the multi-label independency assumption.

In general, MLKNN performs the best, followed by RA*k*EL, followed by Classifier Chain and Binary Relevance. Classifier Chain improves the performance of Binary Relevance. Adaboost.MH performs the worst, followed by Pruned Problem Transformation. Why does Adaboost.MH perform the worst? We conjecture that the possible reason is its default base learner, decision stump. In the future, we will further investigate this.

5 Conclusion

In this paper, we provide an empirical comparison on six multi-label classification algorithms using eleven datasets. Our experiments show that the adaptive multi-label learning algorithm MLKNN performs the best, followed by RA*k*EL, followed by Classifier Chain and Binary Relevance. Adaboost.MH performs the worst, followed by Pruned Problem Transformation. This provides the guide for multi-label classification practitioners and saves their time to try and to estimate the possible

achievement. This also stimulates us to study adapting traditional single label classification algorithms for multi-label problems. Our experimental results also provide us the confidence on the conjecture: there exist correlations among multi-labels. The multi-label independency assumption is not succeeded in most of datasets. How to utilize the correlations among these labels will shed a light for our future research.

We will continue to evaluate the performance of existing multi-label classification algorithms. In the same time, we are going to design novel algorithms for multi-classifications with the insights found in the experiments.

Acknowledgement. We thank the anonymous reviewers for the valuable comments. The work was supported by the National Science Foundation (IIS-1115417).

References

1. Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining Multi-label Data. In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook, 2nd edn. Springer (2010)
2. Klimt, B., Yang, Y.: Introducing the Enron corpus. In: First Conference on Email and Anti-Spam, CEAS (2004)
3. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. Machine Learning 85(3), 333–359 (2011)
4. Tsoumakas, G., Ioannis, K.: Multi-label Classification: An Overview. International Journal of Data Warehousing and Mining (2007)
5. Min-Ling, Z., Zhou, Z.: ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognition 40(7), 2038–2048 (2007)
6. Arunadevi, J., Rajamani, V.: An Evolutionary Multi Label Classification Using Associative Rule Mining. International Journal of Soft Computing 6(2), 20–25 (2011)
7. Tsoumakas, G., Katakis, I., Vlahavas, I.: Random k-labelsets for multilabel classification. IEEE Transactions on Knowledge and Data Engineering 23(7), 1079–1089 (2011)
8. Yu-Yin, S., Zhang, Y., Zhi-Hua, Z.: Multi-label learning with weak label. In: Twenty-Fourth AAAI Conference on Artificial Intelligence (2010)
9. Alvares, C.E., Monard, M.C., Metz, J.: Multi-label Problem Transformation Methods: A Case Study. CLEI Electronic Journal 14(1), 4 (2011)
10. Tsoumakas, G., et al.: Mulan: A java library for multi-label learning. Journal of Machine Learning Research 1, 1–48 (2010)
11. Tao, L., Zhang, C., Zhu, S.: Empirical studies on multi-label classification. In: The Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2006 (2006)
12. Huang, D.-S., McGinnity, M., Heutte, L., Zhang, X.-P. (eds.): ICIC 2010. CCIS, vol. 93. Springer, Heidelberg (2010)
13. Colin, C., Mohammad, S.M., de Bruijn, B.: Binary classifiers and latent sequence models for emotion detection in suicide notes. Biomedical Informatics Insights 5(suppl. 1), 147 (2012)
14. Jesse, R.: A pruned problem transformation method for multi-label classification. In: Proc. 2008 New Zealand Computer Science Research Student Conference, NZCSRS 2008 (2008)

15. Yoav, F., Schapire, R., Abe, N.: A short introduction to boosting. Journal-Japanese Society for Artificial Intelligence 14, 771–780 (1999): 1612
16. Min-Ling, Z., Zhang, K.: Multi-label learning by exploiting label dependency. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM (2010)
17. Schapire, R.E., Singer, Y.: Boostexter: A boosting-based system for text categorization. Machine Learning 39(2/3), 135–168 (2000)
18. Tsoumakas, G., Zhang, M.-L., Zhou, Z.-H.: Tutorial on learning from multi-label data. In: ECML/PKDD 2009, Bled, Slovenia (2009), http://www.ecmlpkdd2009.net/wp-content/uploads/2009/08/learningfrom-multi-label-data.pdf
19. Kumar, V., Wu, X.: Adaboost, The top ten algorithms in data mining, ch. 7, pp. 127–144. CRC Press (2009)

Robust Feature Selection for SVMs under Uncertain Data

Hoai An Le Thi[1,2], Xuan Thanh Vo[1], and Tao Pham Dinh[3]

[1] Laboratory of Theoretical and Applied Computer Science EA 3097
University of Lorraine, Ile de Saulcy, 57045 Metz, France
hoai-an.le-thi@univ-lorraine.fr, xuan-thanh.vo5@etu.univ-lorraine.fr
[2] Lorraine Research Laboratory in Computer Science and Its Applications
CNRS UMR 7503, University of Lorraine, 54506 Nancy, France
[3] Laboratory of Mathematics, National Institute for Applied Sciences-Rouen
Avenue de l'Université- 76801 Saint-Etienne-du-Rouvray cedex, France
pham@insa-rouen.fr

Abstract. In this paper, we consider the problem of feature selection and classification under uncertain data that is inherently prevalent in almost all datasets. Using principles of Robust Optimization, we propose a robust scheme to handle data with ellipsoidal model uncertainty. The difficulty in treating zero-norm ℓ_0 in feature selection problem is overcome by using an appropriate approximation and DC (Difference of Convex functions) programming and DCA (DC Algorithm). The computational results show that the proposed robust optimization approach is more performant than a traditional approach in immunizing perturbation of the data.

Keywords: Feature selection, SVM, Robust Optimization, DC programming, DCA.

1 Introduction

Data uncertainty is common in real-world applications due to various causes, including imprecise measurement, outdated sources and implementation errors. These kinds of uncertainty must be handled cautiously, or else the results could be highly unreliable even with very small perturbations of the nominal data. Consequently, there exists a real need of a methodology capable to detect the cases when data uncertainty can heavily effect the quality of the nominal solution, and to generate a robust solution in such cases, one that is immunized against the effect of data uncertainty. For this need, we address here the *Robust Optimization* approach studied by Ben-Tal, El Ghaoui and Nemirovski [1].

1.1 A Robust Optimization Approach

A generic mathematical programming problem is of the form

$$\min_{x\in\mathbb{R}^n}\{f_0(x,u) : f_i(x,u)\leq 0, i=1,\ldots,m\}, \qquad \text{(P[u])}$$

P. Perner (Ed.): ICDM 2013, LNAI 7987, pp. 151–165, 2013.

where $x \in \mathbb{R}^n$ is a vector of decision variables, the function f_0 (the objective function) and $f_1, \ldots, f_m$ are *structural elements* of the problem, and u stands for the *data* specifying a particular problem instance. Since the data u can not be determined exactly, it is assumed to take arbitrary values in an *uncertainty set* $\mathcal{U}$ in the space of data. Then, we have to deal with an *uncertain optimization problem* defined as a collection of the usual ("certain") optimization problems

$$\{(P[u]) \,|\, u \in \mathcal{U}\}. \tag{1}$$

For the purpose of immunizing against the effect of data uncertainty, a meaningful candidate solution x of the uncertain problem (1) is required to be feasible for all realizations of the disturbances u within $\mathcal{U}$. That is, x is required to satisfy the semi-infinite system of constraints

$$f_i(x, u) \leq 0, i = 1, \ldots, m \quad \forall u \in \mathcal{U}.$$

Moreover, to quantify robustly the quality of the uncertain problem, we minimize the largest value $\hat{f}_0(x) = \sup_{u \in \mathcal{U}} f_0(x, u)$ – say *robust value* – of the "true" objective $f_0(x)$ over all realizations of the data from the uncertainty set. These lead us to *Robust Counterpart* of the uncertain problem (1),

$$\min_{x \in \mathbb{R}^n} \{\sup_{u \in \mathcal{U}} f_0(x, u) : f_i(x, u) \leq 0, i = 1, \ldots, m \quad \forall u \in \mathcal{U}\}. \tag{2}$$

The feasible/optimal solutions to the Robust Counterpart are called *robust feasible/optimal* solutions of the uncertain problem (1).

1.2 Robust Optimization for Feature Selection in SVMs

In this paper, we focus on feature selection in the context of Support Vector Machines (SVMs) learning with two-class linear models as well as presence of uncertainty data. In traditional feature selection, the patterns (assumed as vectors $x \in \mathbb{R}^n$) belong to one of two classes (labeled by +1 or -1), and we seek to discriminate them by a hyperplane $\mathcal{H} = \{x \,|\, x \in \mathbb{R}^n, \langle w, x\rangle + b = 0\}$, $(w \in \mathbb{R}^n, b \in \mathbb{R})$ which uses as few features as possible. This aims to select a subset of relevant features while preserving the discriminative ability and improving the performance of classifier. For this traditional approach, the input data (x, δ)–patterns with corresponding labels–are given *exactly*. However, as mentioned above, this is unrealistic because uncertainty data is ubiquitous in many real world applications. In the context of this paper we will assume that the uncertainty is only in the patterns x and the labels $\delta \in \{+1, -1\}$ are known precisely whenever given. Motivated by robust optimization approach by Ben-Tal, El Ghaoui and Nemirovski [1], the notion of uncertainty is made explicit by specifying the allowable values of a data point via an ellipsoid.

We take a look at existing works on classification under data uncertainty. Bhattacharyya et al. [5] develops Second Order Cone Programming (SOCP) SVM formulation to design a robust linear classifier when the uncertainty was

described by multivariate normal distributions. This work has been generalized in [23] by proposing a SOCP formulation for designing robust binary classifier for arbitrary distributions having finite mean and covariance. The latter approach can be interpreted as ellipsoidal bounded uncertainty. Bi and Zang [6] provided Total Support Vector Classification (TSVC) formulation for bounded uncertainties. A similar work was developed in Trafalis et al. [24] for robust SVM classification of imbalanced and noisy data. In the above works, the authors use ℓ_2-norm for regularization. Works on the feature selection problem and uncertainty are rarely encountered. In [5], the authors developed robust sparse hyperplanes based on ellipsoidal data uncertainty model to uncertain molecular profiling data using the sparsity-inducing regularizer ℓ_1.

In this work we consider the ellipsoidal data uncertainty model with ℓ_0 regularizer term representing the sparsity and investigate an efficient nonconvex programming approach to tackle the robust feature selection SVM problem. Our method is based on DC (Difference of Convex functions) programming and DCA (DC Algorithms) that were introduced by Pham Dinh Tao in a preliminary form in 1985. They have been extensively developed since 1994 by Le Thi Hoai An and Pham Dinh Tao (see [12,25,26] and the references therein). Using an appropriate approximation of ℓ_0-norm we reformulate the considered problem as a DC program and then develop a DCA based algorithm for solving it.

We now describe the notations used in this paper. All vectors will be column vectors unless transposed to a row vector by a superscript T. For a scalar $s \in \mathbb{R}$, s_+ is defined by $\max\{0, s\}$. For vectors $x, y \in \mathbb{R}^n$ and $1 \leq p < \infty$, the inner product and ℓ_p-norm are $\langle x, y\rangle = \sum_{i=1}^n x_i y_i$, $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{\frac{1}{p}}$ respectively.

The rest of this paper is organized as follows. The next section states the problem of feature selection and classification with uncertain data, and the specific ellipsoidal model. In section 3 we present DC programming and DCA for general DC programs, and show how to apply DCA to solve our robust feature selection and classification problem. Finally, the numerical experiments are presented in section 4 and section 5 concludes the paper.

2 Feature Selection for SVMs under Uncertain Data

2.1 Feature Selection for Linear Two-Class SVM Models

Consider a two-class dataset consisting of N data points as well as labels, $\{(x^i, \delta_i)\}_{i=1}^N \subset \mathbb{R}^n \times \{-1, 1\}$. We suppose that $N = m + k$, and the first m data points belong to the class with label $+1$ while the last k data points belong to the class -1. The feature selection for SVM problem is formulated in [2] as follows:

$$\min_{w,b} (1-\lambda) \left(\sum_{i=1}^N \sigma_i \left[1 - \delta_i(\langle w, x^i\rangle + b)\right]_+ \right) + \lambda \|w\|_0, \tag{3}$$

where $\sigma_i = \frac{1}{m}$ if $\delta_i = 1$ and $\sigma_i = \frac{1}{k}$ if $\delta_i = -1$, and the parameter $\lambda \in [0, 1]$ is a measure of trade-off between misclassification and sparsity, and the ℓ_0-norm of

the vector w is defined as

$$\|w\|_0 = \text{cardinality}\{j \,|\, w_j \neq 0\}.$$

2.2 Data Uncertainty Model and Robust Counterpart

Assume that each input data x^i $(i = 1, \ldots, N)$ varies in a given *uncertainty set* $\mathcal{U}_i$. Then, *uncertain problem* corresponding to (3) is a collection of the form

$$\left\{ \min_{w,b} (1-\lambda) \left(\sum_{i=1}^{N} \sigma_i \left[1 - (\langle w, x^i \rangle + b)\right]_+ \right) + \lambda \|w\|_0 \right\}_{\substack{x^i \in \mathcal{U}_i \\ i=1,\ldots,N}}. \tag{4}$$

Since uncertainty on data points x^i is separable, the *Robust Counterpart* of the uncertain problem (4) is given by

$$\min_{w,b} \left\{ (1-\lambda) \left(\sum_{i=1}^{N} \sigma_i \sup_{x^i \in \mathcal{U}_i} \left[1 - (\langle w, x^i \rangle + b)\right]_+ \right) + \lambda \|w\|_0 \right\},$$

or equivalently,

$$\begin{array}{l} \min\limits_{w,b,\xi} (1-\lambda) \left(\dfrac{1}{m} \sum\limits_{i=1}^{m} \xi_i + \dfrac{1}{k} \sum\limits_{i=m+1}^{N} \xi_i \right) + \lambda \|w\|_0 \\ s.t \quad \xi_i \geq \sup\limits_{x^i \in \mathcal{U}_i} \left(1 - \delta_i(\langle w, x^i \rangle + b)\right), \xi_i \geq 0 \quad \forall i = 1, \ldots, N. \end{array} \tag{5}$$

2.3 Ellipsoidal Uncertainty Model

In this section, we consider a simple case when the input data uncertainty is described by ellipsoidal sets, called the ellipsoidal uncertainty model. This means that each input data x^i $(i = 1, \ldots, N)$ varies in an ellipsoid defined by

$$\mathcal{E}_i = \mathcal{E}_i(\overline{x}^i, P_i) = \{\overline{x}^i + P_i^{1/2} u : \|u\|_2 \leq 1\},$$

where $\overline{x}^i$ represents the centre, and the symmetric positive semidefinite matrix P_i represents the shape of the ellipsoid $\mathcal{E}_i$. The centre $\overline{x}^i$ is refereed as *nominal value* of x^i. Substituting $x^i = \overline{x}^i + P_i^{1/2} u$, $(\|u\|_2 \leq 1)$, we have

$$\begin{aligned} \sup_{x^i \in \mathcal{E}_i} \left(1 - \delta_i(\langle w, x^i \rangle + b)\right) &= 1 - \delta_i(\langle w, \overline{x}^i \rangle + b) + \sup_{\|u\|_2 \leq 1} \langle P_i^{1/2} w, u \rangle \\ &= 1 - \delta_i(\langle w, \overline{x}^i \rangle + b) + \|P_i^{1/2} w\|_2. \end{aligned}$$

Then, the robust feature selection problem takes the form

$$\min \left\{ (1-\lambda) \left(\frac{1}{m} \sum_{i=1}^{m} \xi_i + \frac{1}{k} \sum_{i=m+1}^{N} \xi_i \right) + \lambda \|w\|_0 : (w, b, \xi) \in K \right\}, \tag{6}$$

where $K = \left\{ (w, b, \xi) : \delta_i(\langle w, \overline{x}^i \rangle + b) \geq 1 - \xi_i + \|P_i^{1/2} w\|_2, \xi_i \geq 0, i = 1, \ldots, N \right\}$ is a closed convex set.

Below, we give some geometric interpretations for the ellipsoidal uncertainty model. The proofs are trivial, so we omit them.

Proposition 1. *The hyperplane* $\mathcal{H} = \{x \in \mathbb{R}^n : \langle w, x\rangle + b = 0\}$ *and the ellipsoid* $\mathbb{E} = \{x \in \mathbb{R}^n : x = x_0 + P^{1/2}u, \|u\|_2 \leq 1\}$ *have a common point if and only if*

$$|\langle w, x_0\rangle + b| \leq \|P^{1/2}w\|_2. \tag{7}$$

Proposition 2. *Suppose that* $(\hat{w}, \hat{b}, \hat{\xi})$ *is a solution to the problem* (6) *and* $\hat{w} \neq 0$. *Then we have, for any* $i = 1, \ldots, N$,

$$\hat{\xi}_i = \left[\sup_{x^i \in \mathcal{E}_i} \left\{1 - \delta_i(\langle \hat{w}, x^i\rangle + \hat{b})\right\}\right]_+ = \left[1 - \delta_i(\langle \hat{w}, \hat{x}^i\rangle + \hat{b})\right]_+, \tag{8}$$

where $\hat{x}^i$ *is determined by* $\hat{x}^i = \overline{x}^i - \delta_i \dfrac{P_i w}{\|P_i^{1/2} w\|_2} \in \mathcal{E}_i$.

It is easy to see that $\hat{x}^i$ in Proposition 2 is on the boundary of the ellipsoid $\mathcal{E}_i$. With the robust formulation, the constraints $\delta_i(\langle w, \overline{x}^i\rangle + b) - 1 \geq \|P_i^{1/2}w\|_2 - \xi_i$ and $\xi_i \geq 0$ mean that we try to find a hyperplane that separates not only $\overline{x}^i$ but also entirely corresponding uncertainty set. If separation constraints are violated, a part or entire uncertainty set is on wrong side and $\hat{x}$ is the "worst-case" – that is, $\hat{x}$ is a point in uncertainty set which is most severely misclassified. Therefore, the value of $\hat{\xi}^i$ will measure misclassification in worst-case. By solving the robust problem, we desire to reduce misclassification in worst-case.

Statistical Interpretation. Before the end of this section, we give another interpretation that is meaningful.

For each $i = 1, \ldots, N$, we assume that x^i is a random vector with mean $\mathbf{E}(x^i) = \overline{x}^i$ and the variance $\mathbf{Var}(x^i) = P_i$. Then, the quality $\zeta_i = 1 - \delta_i(\langle w, x^i\rangle + b)$ is also a random variable with mean $\mathbf{E}(\zeta_i) = 1 - \delta_i(\langle w, \overline{x}^i\rangle + b)$ and variance

$$\mathbf{Var}(\zeta_i) = \mathbf{Var}(\langle w, x^i\rangle) = w^T P_i w = \|P_i^{1/2}w\|_2^2.$$

For any $\rho > 0$, by Chebyshev inequality, we have

$$\mathbf{Pr}\left(\zeta_i > \mathbf{E}(\zeta_i) + \rho\sqrt{\mathbf{Var}(\zeta_i)}\right) \leq \mathbf{Pr}\left(|\zeta_i - \mathbf{E}(\zeta_i)| > \rho\sqrt{\mathbf{Var}(\zeta_i)}\right) \leq \frac{1}{\rho^2}.$$

Especially, if x^i is normal distribution, then so is ζ_i. Hence,

$$\mathbf{Pr}\left(\zeta_i > \mathbf{E}(\zeta_i) + \rho\sqrt{\mathbf{Var}(\zeta_i)}\right) = 1 - \Phi(\rho) \leq \exp(-\rho^2/2),$$

where Φ is the normal cumulative distribution function, that is

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} \exp\left(-\frac{s^2}{2}\right) ds.$$

Therefore, for large enough $\rho > 0$, the probability that $\zeta_i > \mathbf{E}(\zeta_i) + \rho\sqrt{\mathbf{Var}(\zeta_i)}$ is small. Let us choose a "safety parameter" $\rho > 0$ and ignore the "rare event"

$\zeta_i > \mathbf{E}(\zeta_i) + \rho\sqrt{\mathbf{Var}(\zeta_i)}$, the robust value of $\left[1 - \delta_i(\langle w, x^i\rangle + b)\right]_+ = (\zeta_i)_+$ in objective function of (3) can be taken at

$$\left[\mathbf{E}(\zeta_i) + \rho\sqrt{\mathbf{Var}(\zeta_i)}\right]_+ = \left[1 - \delta_i(\langle w, \overline{x}^i\rangle + b) + \rho\|P_i^{1/2}w\|_2\right]_+.$$

Then, a robust counterpart of (4) by this way is similar to (6).

3 DCA for Solving the Robust Feature Selection Problem (6)

3.1 DC Programming and DCA

For a convex function θ, the subdifferential of θ at $x_0 \in \text{dom}\theta := \{x \in \mathbb{R}^n : \theta(x_0) < +\infty\}$, denoted by $\partial\theta(x_0)$, is defined by

$$\partial\theta(x_0) := \{y \in \mathbb{R}^n : \theta(x) \geq \theta(x_0) + \langle x - x_0, y\rangle, \forall x \in \mathbb{R}^n\},$$

and the conjugate θ^* of θ is

$$\theta^*(y) := \sup\{\langle x, y\rangle - \theta(x) : x \in \mathbb{R}^n\}, \quad y \in \mathbb{R}^n.$$

A general DC program is that of the form:

$$\alpha = \inf\{F(x) := G(x) - H(x) \,|\, x \in \mathbb{R}^n\} \quad (P_{dc}),$$

where G, H are lower semi-continuous proper convex functions on $\mathbb{R}^n$. Such a function F is called a DC function, and $G - H$ a DC decomposition of F while G and H are the DC components of F. Note that, the closed convex constraint $x \in C$ can be incorporated in the objective function of (P_{dc}) by using the indicator function on C denoted by χ_C which is defined by $\chi_C(x) = 0$ if $x \in C$, and $+\infty$ otherwise.

A point x^* is called a *critical point* of $G - H$, or a generalized Karush-Kuhn-Tucker point (KKT) of (P_{dc}) if

$$\partial H(x^*) \cap \partial G(x^*) \neq \emptyset. \tag{9}$$

Based on local optimality conditions and duality in DC programming, the DCA consists in constructing two sequences $\{x^k\}$ and $\{y^k\}$ (candidates to be solutions of (P_{dc}) and its dual problem respectively). Each iteration k of DCA approximates the concave part $-H$ by its affine majorization (that corresponds to taking $y^k \in \partial H(x^k)$) and minimizes the resulting convex function (that is equivalent to determining $x^{k+1} \in \partial G^*(y^k)$).

Generic DCA scheme
Initialization: Let $x^0 \in \mathbb{R}^n$ be an initial guess, $0 \leftarrow k$.
Repeat
- Calculate $y^k \in \partial H(x^k)$

- Calculate $x^{k+1} \in \arg\min\{G(x) - \langle x, y^k\rangle : x \in \mathbb{R}^n\}$ (P_k)
- $k+1 \leftarrow k$

Until convergence of $\{x^k\}$.

Convergences properties of DCA and its theoretical basic can be found in [25,12]. It is worth mentioning that

- DCA is a descent method (*without linesearch*): the sequences $\{G(x^k) - H(x^k)\}$ and $\{H^*(y^k) - G^*(y^k)\}$ are decreasing.
- If $G(x^{k+1}) - H(x^{k+1}) = G(x^k) - H(x^k)$, then x^k is a critical point of $G - H$ and y^k is a critical point of $H^* - G^*$. In such a case, DCA terminates at k-th iteration.
- If the optimal value α of problem (P_{dc}) is finite and the infinite sequences $\{x^k\}$ and $\{y^k\}$ are bounded then every limit point x (resp. y) of the sequences $\{x^k\}$ (resp. $\{x^k\}$) is a critical point of $G - H$ (resp. $H^* - G^*$).
- DCA has a *linear convergence* for general DC programs, and has a finite convergence for polyhedral DC programs.

A deeper insight into DCA has been described in [12]. For instant it is crucial to note the main feature of DCA: DCA is constructed from DC components and their conjugates but not the DC function f itself which has infinitely many DC decompositions, and there are as many DCA as there are DC decompositions. Such decompositions play a critical role in determining the speed of convergence, stability, robustness, and globality of sought solutions. It is important to study various equivalent DC forms of a DC problem. This flexibility of DC programming and DCA is of particular interest from both a theoretical and an algorithmic point of view.

For a complete study of DC programming and DCA the reader is referred to [25,12,26] and the references therein. The solution of a nonconvex program (P_{dc}) by DCA must be composed of two stages: the search of an *appropriate* DC decomposition of f and that of a *good* initial point.

It should be noted that the convex concave procedure (CCCP) for constructing discrete time dynamical systems mentioned in [27] is a special case of DCA applied to smooth optimization. Likewise, the SLA (Successive Linear Approximation) algorithm developed in [2] is a version of DCA for concave minimization program.

In the last decade, a variety of works in Machine Learning based on DCA have been developed. The efficiency and the scalability of DCA have been proved in a lot of works (see e.g. [7,10,13,14,15,16,18,19,22] and the list of reference in [17]). These successes of DCA motivated us to investigate it for solving the robust feature selection in SVM problem under uncertain data.

3.2 Approximation of the ℓ_0-Norm

The ℓ_0 norm results in a combinatorial optimization problem, and hence is not practical for large scale problems. We consider a smooth approximation to the ℓ_0 norm proposed in [2].

For an $\alpha > 0$, let $\eta(x, \alpha)$ be the function defined by

$$\eta(x, \alpha) = \begin{cases} 1 - e^{-\alpha x} \text{ if } x \geq 0 \\ 1 - e^{\alpha x} \quad \text{if } x < 0. \end{cases} \tag{10}$$

Then, a good approximation of the zero norm $\|w\|_0$ is given by

$$\|w\|_0 \approx \sum_{i=1}^{n} \eta(w_i, \alpha). \tag{11}$$

In what follows, for a given α, we will use $\eta(x)$ instead of $\eta(x, \alpha)$. Using the approximation (11), we can formulate the robust feature selection problem (6) in the form

$$\min \left\{ F(w, b, \xi) := (1 - \lambda) \left(\frac{1}{m} \sum_{i=1}^{m} \xi_i + \frac{1}{k} \sum_{i=m+1}^{N} \xi_i \right) + \lambda \sum_{j=1}^{n} \eta(w_j) : (w, b, \xi) \in K \right\}. \tag{12}$$

3.3 A DC Formulation of Problem (12)

The approximation $\eta(x)$ can be represented as a difference of convex functions $\eta(x) = g(x) - h(x)$ given by

$$g(x) = \begin{cases} \alpha x \quad \text{if } x \geq 0 \\ -\alpha x \text{ if } x < 0 \end{cases} \quad \text{and} \quad h(x) = \begin{cases} \alpha x - 1 + e^{-\alpha x} \text{ if } x \geq 0 \\ -\alpha x - 1 + e^{\alpha x} \text{ if } x < 0 \end{cases}. \tag{13}$$

Consequently, the objective function of (12) can be expressed as

$$F(w, b, \xi) = G(w, b, \xi) - H(w, b, \xi),$$

where

$$G(w, b, \xi) := (1 - \lambda) \left(\frac{1}{m} \sum_{i=1}^{m} \xi_i + \frac{1}{k} \sum_{i=m+1}^{N} \xi_i \right) + \lambda \sum_{j=1}^{n} g(w_j),$$

and

$$H(w, b, \xi) := \lambda \sum_{j=1}^{n} h(w_j)$$

are clearly convex functions. Finally, the Robust Feature Selection DC programming problem (**RFSDC**) can be written as follows

$$\min\{G(X) - H(X) : X = (w, b, \xi) \in K\}. \tag{14}$$

3.4 DCA for Solving the RFSDC Problem (14)

According to the generic DCA scheme, at each iteration k, we have to compute a subgradient $Y^k = (\overline{w}^k, \overline{b}^k, \overline{\xi}^k)$ of H at $X^k = (w^k, b^k, \xi^k)$ and then solve the convex program of the form (P_k)

$$\min\{G(X) - \langle Y^k, X\rangle : X = (w, b, \xi) \in K\}. \tag{15}$$

It is easy to see that H is differentiable everywhere, so $Y^k = (\overline{w}^k, \overline{b}^k, \overline{\xi}^k) = \nabla H(X^k)$ is calculated by

$$\overline{b}^k = \frac{\partial H}{\partial b^k} = 0;\ \overline{\xi}_i^k = \frac{\partial H}{\partial \xi_i^k} = 0;\ i = 1, \ldots, N, \tag{16}$$

$$\overline{w}_j^k = \frac{\partial H}{\partial w_j^k} = \begin{cases} \lambda\alpha(1 - \exp(-\alpha w_j^k)) \text{ if } w_j^k \geq 0 \\ -\lambda\alpha(1 - \exp(\alpha w_j^k)) \text{ if } w_j^k < 0. \end{cases}, \quad i = 1, \ldots, n. \tag{17}$$

For solving the problem (15), we have

$$\min\{G(X) - \langle Y^k, X\rangle : X = (w, b, \xi) \in K\}$$

$$= \min\left\{(1-\lambda)\left(\frac{1}{m}\sum_{i=1}^{m}\xi_i + \frac{1}{k}\sum_{i=m+1}^{N}\xi_i\right) + \lambda\sum_{j=1}^{n}\alpha|w_j| - \langle\overline{w}^k, w\rangle : (w, b, \xi) \in K\right\}$$

$$\Leftrightarrow \min\left\{(1-\lambda)\left(\frac{1}{m}\sum_{i=1}^{m}\xi_i + \frac{1}{k}\sum_{i=m+1}^{N}\xi_i\right) + \lambda\alpha\sum_{j=1}^{n}t_j - \langle\overline{w}^k, w\rangle : (w, b, \xi, t) \in \Omega\right\}, \tag{18}$$

where Ω is a closed convex set defined by

$$\Omega := \left\{(w, b, \xi, t) \in \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^N \times \mathbb{R}^n : (w, b, \xi) \in K, |w_j| \leq t_j, j = 1, \ldots, n\right\}.$$

The problem (18) is an instance of Second Order Cone Programming (SOCP).

Our DCA applied to (14) can be described as follows.

Algorithm 1.

Initialization:

- Let ϵ be a tolerance sufficiently small, and set $k = 0$.
- Choose a starting point $X^0 = (w^0, b^0, \xi^0) \in \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^N$.
- Compute $F^0 = F(X^0)$.

Repeat

- Compute $Y^k = (\overline{w}^k, \overline{b}^k, \overline{\xi}^k) = \nabla H(X^k)$ via (16) and (17).
- Solve the convex problem (18) to obtain $X^{k+1} = (w^{k+1}, b^{k+1}, \xi^{k+1})$.
- Compute $F^{k+1} = F(X^{k+1})$.
- $k + 1 \leftarrow k$.

Until

$|F^{k+1} - F^k| < \epsilon$.

3.5 Robust Feature Selection Using ℓ_1-regularizer

We state here a formulation for the robust feature selection problem using ℓ_1-regularizer.

$$\min \left\{ (1-\lambda) \left(\frac{1}{m} \sum_{i=1}^{m} \xi_i + \frac{1}{k} \sum_{j=m+1}^{N} \xi_j \right) + \lambda \|w\|_1 : (w, b, \xi) \in K \right\}. \tag{19}$$

This formulation is slightly different from formulation studied in [4] when we put different weights on slack variables ξ_i. In the numerical experiments we will compare the solution of this formulation with the one computed by our approach.

4 Numerical Experiments

The numerical experiments aim to evaluate the performance of nominal solutions and robust solutions of the feature selection SVM problem under the impact of data uncertainty. We consider two approaches: our Algorithm 1, say DCA applied on (14), and a standard approach based on the ℓ_1-regularizer model (19). Nominal classifiers and robust classifiers with different values of noise level ρ (see definition below) were trained on training sets. The error rates (ordinary and worst case) were computed for the test set, in which the uncertainty of data of the test set is added by using shapes same as for the training set.

The algorithm has been coded in VC++ and implemented on a Intel Core i5 CPU 2×2.74 GHz, RAM 4GB. All convex problems (19) and (P_k) (18) are solved by the commercial software CPLEX 11.2.

4.1 Error Measures

Consider separating hyperplane $\mathcal{H}(w, b) = \{x \,|\, \langle w, x\rangle + b = 0\}$, and data $\{(x^i, \delta_i)\}_{i=1}^N$ or $\{\mathcal{E}(x^i, P_i), \delta_i\}_{i=1}^N$, where $\mathcal{E}_i = \mathcal{E}(x^i, P_i)$ are uncertainty sets.

Ordinary error. For a data point x^i, an ordinary error occurs when x^i is misclassified, i.e $\text{sign}\{\langle w, x^i\rangle + b\}$ differs from δ_i.

Worst case error. For a data point x^i, a worst case error occurs when x^i has an ordinary error or $\mathcal{H}(w, b)$ intersects uncertainty set $\mathcal{E}_i$.

4.2 Datasets

We use four real datasets from UCI repository [8] and two real microarray gene expression datasets. Three datasets from UCI involves two variants of the Wisconsin Prognostic Breast Cancer Database (WPBC 24 and 60), Spambase Data Set (SPA), and Internet Advertisements (ADV). Two gene expression datasets are Leukemia [9] and Lung Cancer (available at `http://www.chestsurg.org/publications/2002-microarray.aspx`).

We will add an uncertainty to continuous data. Particularly, for the dataset ADV, only first three attributes are real continuous (present size of an image) and the others is binary. And, we only add uncertainty to these attributes.

Table 1. Datasets used in experiments. The numbers in bracket present class distribution +1/-1.

Name	Number of features	Number of points in training set	Number of points in testing set
WPBC(24 mo)	32	133 (101/32)	65 (50/15)
WPBC(60 mo)	32	380 (142/238)	189 (70/119)
SPA	57	1725(1046/679)	576(348/228)
ADV	1558	2458 (344/2114)	821 (115/706)
Leukemina	7129	38 (27/11)	34 (20/14)
Lung Cancer	12533	32 (16/16)	149 (15/134)

The information about the datasets is summarized in Table 1. Below, we present the way to create uncertainty sets.

The Centre and Shapes of Uncertainty Sets. The centers of ellipsoids are equated with observed data points, say $\overline{x}^i \equiv x^i$. We set $P_i = P^+$ if the label $\delta_i = +1$, and $P_i = P^-$ if the label $\delta_i = -1$, where $P^{\pm} = \text{diag}(P_1^{\pm}, \ldots, P_n^{\pm})$ is determined by

$$P_j^+ = \frac{1}{m}\sum_{i=1}^{m}(x_j^i)^2 - \left(\frac{1}{m}\sum_{i=1}^{m}x_j^i\right)^2, \text{and } P_j^- = \frac{1}{k}\sum_{i=m+1}^{N}(x_j^i)^2 - \left(\frac{1}{k}\sum_{i=m+1}^{N}x_j^i\right)^2.$$

The Noise Level Parameter ρ. To investigate the effect of different amounts of data uncertainty, we use a noise level parameter $\rho \geq 0$ to scale uncertainty sets. That is, we replace P by $\rho^2 P$. By this way, we can control the degree of the perturbation. When $\rho = 0$, there is no pertubation in data points.

4.3 Experimental Setup

In experiments, we set $\alpha = 5$ and stop tolerance $\epsilon = 10^{-6}$ for DCA. The non-zero elements of w are determined according to whether $|w_j|$ exceeds a small threshold. In this experiment, we use the threshold 10^{-6}. The value of λ is chosen through a 10-fold cross validation procedure on training set from a set of candidates given by $\Lambda = \{0.001, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$.

4.4 Experimental Results

The computational results given by Algorithm 1 (ℓ_0(DCA)) and the l_1-regularizer (ℓ_1) are reported in tables 2, 3 and figure 1. We are interested in the efficiency (the sparsity and the classification error) as well as the rapidity of the algorithms.

Sparsity. When the noise level is increasing, the number of selected features of robust classifier has tendency to increase: on Leukemina dataset, increase averagely is 4 times, and on Lung-Cancer dataset increase averagely is 11 times.

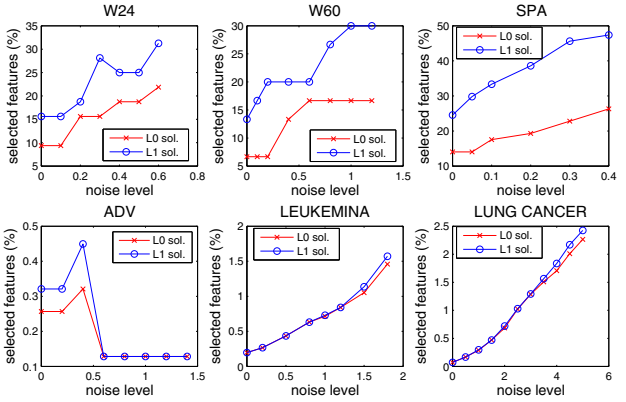


Fig. 1. The percentage of selected features versus noise level

ADV dataset is particular case since it has only three features considered as uncertain quantities. And robust model suppresses these features while nominal solution always uses them. These facts imply that the feature selection is not sensitive to the effect of noise.

In all cases, the classifiers obtained by Algorithm 1 (ℓ_0(DCA)) are sparser than those obtained by the regularizer ℓ_1 approach. For sparse datasets, such as Leukemina and Lung-Cancer, the two approaches give very similar results and the difference is only considerable when the noise level is high.

Classification Error. Except Leukemina and SPA datasets, nominal classifiers have less ordinary errors than robust classifiers. However, they are still comparable. Whereas the worst case error brings out the advantage of robust classifiers over nominal classifiers.

Table 2. The percentage of selected features and the training time in second. For robust solutions, the results are the average on all considered values of the noise level.

Dataset	percentage of selected features				running time (second)			
	Nominal sol.		Robust sol.		Nominal sol.		Robust sol.	
	ℓ_1	ℓ_0(DCA)	ℓ_1	ℓ_0(DCA)	ℓ_1	ℓ_0(DCA)	ℓ_1	ℓ_0(DCA)
WPBC (24mo)	15.62	9.37	23.93	16.65	0.03	0.09	0.23	0.72
WPBC (60mo)	13.33	6.67	23.33	13.33	0.07	0.19	1.36	4.31
SPA	24.56	14.03	38.94	20	0.61	4.57	7.86	57.24
ADV	0.32	0.25	0.2	0.17	16.23	19.49	39.61	68.18
Leukemina	0.19	0.19	0.8	0.77	5.28	10.29	12.75	49.33
Lung Cancer	0.07	0.07	0.88	0.85	7.6	14.9	23.33	79.15

Table 3. The average error (percentage) of the nominal solution and robust solution on all considered values of the noise level. For each dataset, the upper row represents the ordinary error and the under row represents the worst case error. The number in bracket is the standard deviation.

Dataset	Nominal solution		Robust solution	
	ℓ_1	ℓ_0(DCA)	ℓ_1	ℓ_0(DCA)
WPBC	15.4	16.9	17.9(1.4)	18.4(1.9)
(24 mo)	37.7	34.8	20.7(1.5)	20.7(2.1)
WPBC	8.5	11.6	10.1(1.7)	13.1(1)
(60 mo)	40.7	31.6	17.2(5.2)	20.9(3.5)
SPA	15.5	12.5	15.6(0.3)	11.4(0.8)
	39.1	41.6	22.4(5.6)	19.3(5.4)
ADV	6.5	5.7	6.7(0.07)	6.2(0.8)
	17.6	13.2	7.1(0.6)	6.3(0.6)
Leukemina	2.9	2.9	1.2(1.5)	1.2(1.5)
	17.6	17.6	11.6(7.4)	10.4(5.8)
Lung	1.3	1.3	1.5(0.7)	1.6(0.8)
Cancer	33.8	33.8	5.8(1.7)	5.2(1.5)

The classification errors of the nominal solutions given by the two approaches ℓ_0(DCA) and ℓ_1 are comparable on ordinary errors: the same results in the last two datasets (Leukemina, Lung cancer), and ℓ_0(DCA) is better than ℓ_1 in ADV and SPA datasets while ℓ_1 is better in WPBC datasets. As for worst case errors of nominal solutions, the two approaches give the same results in two datasets (Leukemina, Lung cancer), and ℓ_0(DCA) is better in three datasets (ADV, WPBCs), while ℓ_1 is better in SPA dataset.

Concerning the robust solutions, the worst case classification errors given by ℓ_0(DCA) are smaller than that performed by ℓ_1 approach in ADV, SPA, Leukemina, and Lung cancer datasets. The two approaches give the same result on WPBC(24mo) and ℓ_1 approach is slightly better on WPBC(60mo). Moreover, except for WPBC(24mo) the standard deviation of ℓ_0(DCA) is smaller than that of ℓ_1 approach. It means that ℓ_0(DCA) is less sensitivitive than ℓ_1 on the noise level.

Implementation and Training Time. The training time is given in Table 2. We observe that the training time to get nominal solution is less than that to robust solution. This is reasonable because robust formulations are SOCP programs, while nominal formulations are LP programs that require less time to the resolution. The training time of ℓ_1 approach is shorter than that of ℓ_0(DCA). This is not surprising because the former is a single convex problem, while the latter requires the resolution of some convex programs.

5 Conclusion

We have investigated a DC programming approach for the feature selection SVM problem under data uncertainty. Using robust optimization technique, we have

proposed a robust formulation that handle input uncertainty in ellipsoidal sets. Based on DC programming and DCA, we have developed an efficient algorithm to solve the resulting optimization problem. The computational results show that the proposed robust formulation is more resilient than the nominal formulation because the robust solutions are able to immunize against the effect of uncertainty. Comparative numerical results also show that the ℓ_0(DCA) approach is more robust than ℓ_1 approach in this robust feature selection SVM problem on both generalization and sparsity.

In this paper, we consider the ellipsoidal model for uncertain data in which features are interdependent but the uncertainty of each data point is independent. This model is reasonable in most practical data. However, other models of uncertainty should be studied in future, such as interval uncertainty models, where each feature is estimated independently, or uncertainty models on overall data $\{x^i\}_{i=1}^N$. These types of uncertainty are on data points x^i. In addition, the corruption on labels δ_i is also interesting.

References

1. Ben-Tal, A., El Ghaoui, L., Nemirovski, A.: Robust Optimization. Princeton University Press (2009)
2. Bradley, P.S., Magasarian, O.L., Street, W.N.: Feature Selection via mathematical Programming. INFORMS Journal on Computing 10(2), 209–217 (1998)
3. Bennett, K.P., Mangasarian, O.L.: Robust linear programming discrimination of two linearly inseparable sets. Optimization Methods and Software 1(1), 23–34 (1992)
4. Bhattacharyya, C., Grate, L.R., Jordan, M.I., El Ghaoui, L., Mian, I.S.: Robust sparse hyperplane classifier: application to uncertain molecular profiling data. Journal of Computational Biology 11(6), 1073–1089 (2004)
5. Bhattacharyya, C., Pannagadatta, K.S., Smola, A.J.: A second order cone programming formulation for classifying missing data. In: Advances in Neural Information Processing Systems, NIPS 17 (2004)
6. Bi, J., Zhang, T.: Support vector classification with input data uncertainty. Advances in Neural Information Processing Systems 17 (2004)
7. Collobert, R., Sinz, F., Weston, J., Bottou, L.: Large scale transductive SVMs. J. Machine Learn. 7, 1687–1712 (2006)
8. Frank, A., Asuncion, A.: UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine (2010), `http://archive.ics.uci.edu/ml`
9. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular Classifcation of Cancer: Class Discovery and Class Prediction by Gene Ex-pression Monitoring. Science 286, 531–537 (1999)
10. Krause, N., Singer, Y.: Leveraging the margin more carefully. In: International Conference on Machine Learning ICML (2004)
11. Le Thi, H.A., Pham Dinh, T.: Solving a class of linearly constrained indefinite quadratic problems by DC algorithms. Journal of Global Optimization 11(3), 253–285 (1997)

12. Le Thi, H.A., Pham Dinh, T.: The DC (difference of convex functions) Programming and DCA revisited with DC models of real world nonconvex optimization problems. Annals of Operations Research 133, 23–46 (2005)
13. Le Thi, H.A., Belghiti, T., Pham Dinh, T.: A new efficient algorithm based on DC programming and DCA for Clustering. Journal of Global Optimization 37, 593–608 (2006)
14. Le Thi, H.A., Le Hoai, M., Pham Dinh, T.: Optimization based DC programming and DCA for Hierarchical Clustering. European Journal of Operational Research 183, 1067–1085 (2007)
15. Le Thi, H.A., Le Hoai, M., Nguyen, N.V., Pham Dinh, T.: A DC Programming approach for Feature Selection in Support Vector Machines learning. Journal of Advances in Data Analysis and Classification 2(3), 259–278 (2008)
16. Thiao, M., Pham Dinh, T., Le Thi, H.A.: DC programming approach for a class of nonconvex programs involving l0 norm. In: Le Thi, H.A., Bouvry, P., Pham Dinh, T. (eds.) MCO 2008. CCIS, vol. 14, pp. 348–357. Springer, Heidelberg (2008)
17. Le Thi, H.A.: DC Programming and DCA., `http://lita.sciences.univ-metz.fr/~lethi/DCA.html`
18. Liu, Y., Shen, X., Doss, H.: Multicategory ψ-Learning and Support Vector Machine: Computational Tools. Journal of Computational and Graphical Statistics 14, 219–236 (2005)
19. Liu, Y., Shen, X.: Multicategory ψ-Learning. Journal of the American Statistical Association 101, 500–509 (2006)
20. Neumann, J., Schnörr, C., Steidl, G.: Combined SVM-based feature selection and classification. Machine Learning 61(1-3), 129–150 (2005)
21. Neumann, J., Schnörr, C., Steidl, G.: SVM-based Feature Selection by Direct Objective Minimisation. In: Rasmussen, C.E., Bülthoff, H.H., Schölkopf, B., Giese, M.A. (eds.) DAGM 2004. LNCS, vol. 3175, pp. 212–219. Springer, Heidelberg (2004)
22. Ronan, C., Fabian, S., Jason, W., Lé, B.: Trading Convexity for Scalability. In: Proceedings of the 23rd International Conference on Machine Learning, ICML 2006, pp. 201–208 (2006)
23. Shivaswamy, P.K., Bhattacharyya, C., Smola, A.J.: Second order cone programming approaches for handling missing and uncertain data. Journal of Machine Learning Research 7, 1238–1314 (2006)
24. Trafalis, T.B., Raghav, P., Kash, B.: Support Vector Machine Classification of Uncertain and Imbalanced Data using Robust Optimization. In: Proceedings of the 15th WSEAS International Conference on Computers (2011)
25. Pham Dinh, T., Le Thi, H.A.: Convex analysis approach to DC programming: Theory, algorithms and applications. Acta Math. Vietnamica 22(1), 289–357 (1997)
26. Pham Dinh, T., Le Thi, H.A.: DC optimization algorithms for solving the trust region subproblem. SIAM J. Opt. 8, 476–505 (1998)
27. Yuille, A.L., Rangarajan, A.: The Convex Concave Procedure. Neural Computation 15(4), 915–936 (2003)

A Hybrid Machine Learning Method and Its Application in Municipal Waste Prediction

Emadoddin Livani[1,*], Raymond Nguyen[2], Jörg Denzinger[3], Günther Ruhe[1,3], and Scott Banack[4]

[1] Department of Electrical and Computer Engineering, University of Calgary, Canada
elivani@ucalgary.ca
[2] David R. Cheriton School of Computer Science, University of Waterloo, Canada
raynguyen@gmail.com
[3] Department of Computer Science, University of Calgary, Canada
{denzinge,ruhe}@ucalgary.ca
[4] Waste and Recycling Services, City of Calgary, Canada
scott.banack@calgary.ca

Abstract. Prediction methods combining clustering and classification techniques have the potential of creating more accurate results than the individual techniques, particularly for large datasets. In this paper, a hybrid prediction method is proposed from combining weighted k-means clustering and linear regression. Weighted k-means is used to cluster the dataset. Then, linear regression is performed on each cluster to build the final predictors. The proposed method has been applied to the problem of municipal waste prediction and evaluated with a dataset including 63,000 records. The results showed that it outperforms the single application of linear regression and k-means clustering in terms of prediction accuracy and robustness. The prediction model is integrated into a decision support system for strategic and operational planning of waste and recycling services at the City of Calgary in Canada. The potential usage of the prediction model is to improve the resource utilization, like personnel and vehicles.

Keywords: Knowledge engineering, Machine learning, Prediction method, Sustainability and environment, Municipal waste prediction.

1 Introduction

With the increasing rate of information stored in databases, the development of efficient and effective tools for extracting the knowledge from these data has become an increasingly important task for researchers in the areas of databases, statistics, machine learning, and data visualization [1]. Usually the volume of such databases makes manual analysis very difficult. Data mining addresses this problem by extracting implicit, previously unknown, and potentially useful information from data

* Corresponding author, 2500 University Dr. NW, Calgary, AB, Canada, Tel: (1)4032105480.

P. Perner (Ed.): ICDM 2013, LNAI 7987, pp. 166–180, 2013.

using a set of processes performed automatically [2]. Research fields such as statistics and machine learning contributed greatly to the development of various data mining and knowledge discovery algorithms. The objectives of these algorithms include pattern recognition, prediction, association, and clustering [3,4].

One of the challenges that every machine learning algorithm is faced with is the scalability to large datasets. Hybrid data mining methods, which combine clustering and classification techniques, can improve the performance of single clustering or classification when running on large datasets [5]. Usually, they are composed of two learning stages, in which the first one is used for processing the data and the second one for making the final prediction [6,7]. Lenard et al. [6] consider two types of combination: clustering as the processing step (or outlier detection) and then classification for prediction; or in the opposite order, using classification to classify the training set and then cluster each set for future predictions. Choosing the right learner and combination depends on the type of the application and needs to be examined in practice.

Municipal waste management is an important area to apply predictive models [8-10]. However, few attempts have been made to reach a high accuracy level with a high volume of data. In this paper, we propose a hybrid prediction method that addresses the above gap by combining k-means clustering and linear regression methods. K-means clustering is used for creating clusters, each one containing the instances that are most similar to each other. The k-means clustering in this paper differs from the standard k-means since it applies a weighted similarity measure to calculate the distance between the data points. Linear regression is used to acquire the weights of the attributes for the similarity in k-means. After creating the clusters, linear regression is applied to create the final predictor for each cluster.

The proposed method has been evaluated through a case study with the Waste and Recycling Services (WRS) of the City of Calgary in Canada. The history data on the produced waste is a dataset of 63,000 records from the past few years. The results showed that the proposed prediction model outperforms the k-means clustering or linear regression in isolation.

The rest of the paper is organized as follows. Existing hybrid prediction methods are discussed in the next section. In Section 3, background and problem formulation are discussed. The proposed solution and methodology are explained in Section 4. Section 5 and 6 analyze the application of the prediction model in a case study. Finally, Section 7 provides a summary and an outlook at future research.

2 Related Work

Our work falls in the area of hybrid prediction, and hybrid data mining (or machine learning). Also, the current waste prediction models are briefly explained.

A hybrid prediction model for detecting Type-2 diabetic patients is proposed in [4]. Simple k-means clustering is used to validate chosen class labels of given data and then the C4.5 algorithm to build the final classifier model. The Pima Indians diabetes dataset from the University of California Irvine's machine learning repository was

used by the authors. They reached up to a maximum accuracy of 92%. The main difference between this work and ours is that this model is dedicated to classification based on nominal values while our model is to predict actual amounts.

In [5] s hybrid data mining technique is proposed for telecom churn prediction using neural networks. Self-organizing maps (SOM) and back-propagation artificial neural networks (ANN) are combined to filter out outliers and create the prediction model. The hybrid technique is shown to outperform a single neural network in terms of prediction accuracy. The difference of our work is that the both steps (k-means and linear regression) are used in creating the prediction model and the outlier detection is a separate step.

A hybrid prediction method is presented in [11] for prediction of physicochemical parameters. K-means clustering and principal component analysis (PCA) are integrated to reduce the number of network inputs in the next level, called multi-level perceptron (MLP) learning. The result of the MLP phase is the final prediction output. Accuracies ranging from 80% to 85% are reported in the paper. The attributes are not weighted in their approach, as opposed to ours.

A hybrid method based on k-means clustering and functional data analysis is proposed in [12] for short-term forecasting in a district-heating system. A two level clustering-based method for power prediction is presented in [13] for short-term prediction of power produced by a wind turbine at low wind speeds. In both methods, unweighted attributes are used while we use weighted attributes.

A strategy of household waste data analysis is proposed in [14]. A cluster analysis and a tree classifier constructed using the clustered data were presented. An analysis of the decision tree allowed translating the resulting tree in a set of production rules. After the interpretation of these rules, they were able to predict an environmental behavior, based on the information about waste generation and the questionnaire answers. Their approach is useful for classification (nominal values not numerical) while in this paper a predicting model for numerical values is proposed.

Prediction of municipal waste was studied in [8-10]. Using artificial neural network was studied in [15-17] to build a prediction model for solid waste production. Some common deficiencies in the previous works are the small size of the data, low accuracy of the prediction model, and lack of a systematic way for attribute selection. All of these problems are addressed by our proposed prediction model.

3 Background and Problem Statement

In this paper two machine learning methods, k-means clustering and linear regression, are used to create a hybrid model which encompasses them. K-means has been chosen because of its simplicity and efficiency in finding the different behaviors (clusters) in a dataset. Linear regression is widely used as a prediction technique as well. Naturally, there are other algorithms and techniques to consider, but simplicity and reliability make these algorithms a viable choice for our problem. In this section these two methods are briefly discussed and then the research questions which this paper means to answer will be stated.

3.1 K-means Clustering

K-means clustering is an unsupervised machine learning algorithm for partitioning a dataset into k separate clusters where instances (of the dataset) in each cluster are similar to each other [18]. The similarity is defined as the Euclidean distance between two instances. The k-means clustering is defined formally in Definition 1.

Definition 1. Given a set of observations $\{x_1, x_2, \ldots, x_n\}$, where each observation is a d-dimensional real-valued vector ($[a_1, \ldots, a_d]$), k-means clustering aims to partition the n observations into k clusters ($k < n$ and is given) $\mathbf{C} = \{C_1, C_2, \ldots, C_k\}$. The goal is to minimize the within-cluster sum of squares of distances to means, defined as Formula (1), where $\boldsymbol{\mu}_i$ is the mean of cluster C_i.

$$E = \sum_{i=1}^{k} \sum_{x_j \in C_i} (x_j - \mu_i)^2 \tag{1}$$

One of the most used clustering algorithms was first described by [19]. The algorithm starts with a given (random) assignment of the instances to the clusters and proceeds to assign an instance to the cluster whose mean is closest. The process is repeated until the assignments do not change. K-means is used for two purposes: clustering only and as a prediction model.

3.2 Linear Regression

Regression analysis is a well-known technique used for modeling and analyzing of data, using several variables to build a mathematical model that can explain the variation of a dependent variable. Linear regression was the first type of regression analysis to be studied rigorously and to be used extensively in practical applications [20,21]. Linear regression is formally defined in Definition 2.

Definition 2. Given a set of observations $\{x_1, x_2, \ldots, x_n\}$, where each observation is a d-dimensional real-valued vector ($[a_1, \ldots, a_d]$), linear regression finds the coefficients $[w_1 \ldots w_{d-1}]$ to calculate the value of the predicted attribute a_d from the predictive attributes $[a_1 \ldots a_{d-1}]$, as Formula (2).

$$a_d = w_1 a_1 + w_2 a_2 + w_3 a_3 + \cdots + w_{d-1} a_{d-1} \tag{2}$$

3.3 Problem Statement

The two prediction methods discussed previously in this section are well known in the area of machine learning and are often used in isolation for making predictions. However, there are some characteristics in both models which limit their applicability. In k-means clustering, after having all the instances clustered, when a new instance comes in, it will be assigned to a cluster, based on its distance to the cluster means. Then the target attribute (the attribute which is to be predicted) of the cluster mean

will be considered as the prediction output for the new instance. This approximation pattern occurs for all the new instances, causing the overall accuracy to drop. The question here becomes what would be a better value of the target attribute to be considered for the new instance, after determining the cluster? Another assumption of the simple k-means clustering is that all the attributes have equal weight in the distance function. As it will be shown further in the paper, the accuracy of the prediction increases with a weighted distance. Finally, the clustering in k-means is performed on the dataset without using any external knowledge. However, when there is domain knowledge on a problem, an additional partitioning of the dataset could improve the prediction accuracies by creating more compact clusters.

There is also a problem with linear regression because the instances are forced into being represented by a line. As a result, when the deviation in the dataset becomes larger, this approximation will not be so accurate. We introduce our hybrid prediction method to address the above problems by investigating the following research questions:

- Analyzing of the impact of partitioning using domain knowledge on the accuracy of the predictions
- Analyzing the impact of weighted attributes on the accuracy of the predictions
- Analyzing the impact of number of clusters on the accuracy of the predictions
- Comparing the accuracy of the hybrid method with k-means clustering and linear regression

Also, the challenges and issues of using the prediction model in a real world problem will be addressed and explained.

4 Hybrid Prediction Method

The proposed hybrid prediction method includes two main processes: building the prediction models, and making the predictions, which are explained in this section. Also, the cross-validation for the prediction method will be addressed.

4.1 Building the Prediction Models

Fig. 1 illustrates the process of building the prediction models. The process starts with the attribute selection and weighting, which is applied to the whole dataset using the linear regression method. This will result in a linear function, like Formula 2, which determines the value of the predicted attribute (or target attribute) a_d based on the predictive attributes $[a_1...a_{d-1}]$.

We consider the coefficient vector $[w_1 \ldots w_d]$, from Formula (2), as the weights related to the attributes $[a_1 \ldots a_d]$, respectively. The attributes whose coefficients are more than a predefined threshold will be chosen for building the prediction model. We use the Greedy technique to build the regression model, as it is more efficient than other alternatives such as M5 [21]. It creates a smaller sets of attributes

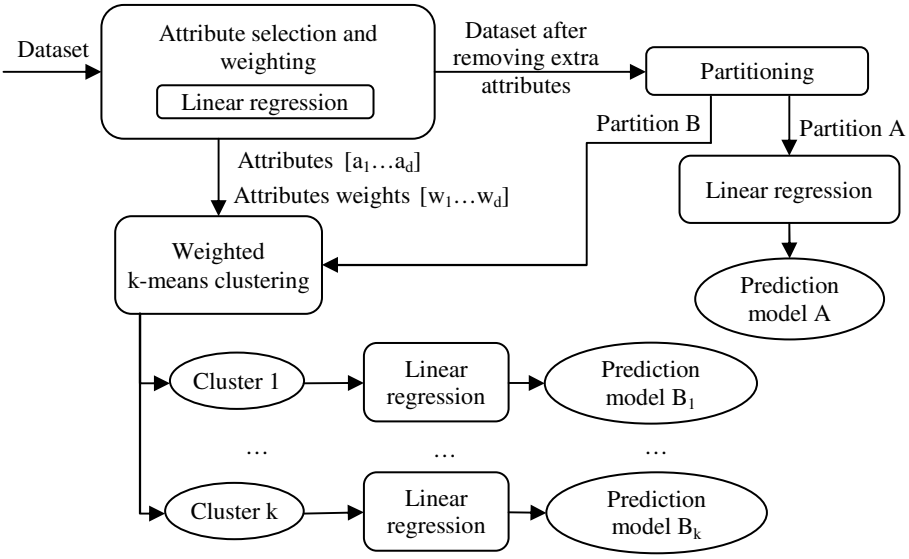


Fig. 1. Hybrid prediction: building the prediction models

compared to M5, while the accuracy of the predictions do not drop. This has been measured by creating two hybrid prediction models, one with Greedy and one with M5, and comparing their prediction accuracy.

After the unselected attributes have been removed from the dataset, the dataset is divided into two partitions, A and B, based on information from the domain expert. This knowledge is assumed to be not recognizable by the machine learning algorithm. The partitioning makes more compact clusters and produces a set of records which have better predictive abilities and are of higher reliability in terms of behavior. For an example of a human-induced partitioning, see [22] and Section 5. Partition A is fed into linear regression directly resulting in prediction model A.

Partition B contains the rest of the records which will be clustered in the next step by the k-means clustering. Weighted Euclidian distance is used as the similarity function, as in Formula (3). In this formula, x_i and x_j are two data points whose distance is to be measured, x_i^t (t = 1…d) is the value of the attribute t in the instance x_i, and w_t (t = 1…d) is the weight of the attribute t.

$$distance\left(x_i, x_j\right) = \sqrt{w_1\left(x_i^1 - x_j^1\right)^2 + w_2\left(x_i^2 - x_j^2\right)^2 + \ldots + w_d\left(x_i^d - x_j^d\right)^2} \qquad (3)$$

The results of this step are k clusters; each containing very similar records. Finally, the linear regression is performed on each cluster resulting in the prediction models $[B_1 \ldots B_k]$, where each B_i (i = 1 … k) is a linear formula as in Formula (2). Fig. 2 illustrates our hybrid method, where the dataset is divided to k (here 3) clusters and then linear regression is applied to each cluster. The lines are the final predictors.

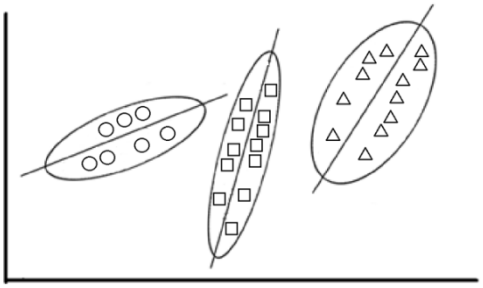


Fig. 2. Visualization of the hybrid prediction model

4.2 Making the Predictions

When a new record comes for prediction, it contains the value for all the attributes, except the target attribute. The following is the algorithm for making the prediction for each new instance:

1. Determine the partition of the new instance based on the domain knowledge
2. If it falls in Partition A, use Prediction Model A as the final predictor FP
3. Else if it falls in Partition B, do the following
   i. Calculate the distance of the new instance to each cluster mean $B_1 \ldots B_k$ using the similarity function of Formula (3)
   ii. Say B_i is the cluster whose mean is the closest to the new instance
   iii. Select the Prediction Model B_i as the final predictor FP
4. Use FP, which is a linear regression as in Formula (2), to calculate the value of the target attribute for the new instance

4.3 Cross-Validation

To cross-validate [23] the model, the dataset is randomly split into training and testing sets. The training set is used as the input data set in Fig. 1. The results are a set of clusters, each one having a unique predictor.

The records in the test set are entered into the prediction process, and the predicted value of the target attribute is compared to the actual value, to measure the prediction accuracy. For each instance x_i in the test set, the prediction error E_i is calculated as Formula (4), where V_i is the actual value of the target attribute of instance x_i and P_i (i = 1 ….N_t) is the predicted value. N_t is the number of instances in test set T.

$$E_i = |V_i - P_i|/V_i \tag{4}$$

The overall accuracy, A_{total}, of the prediction model is calculated as Formula (5).

$$A_{total} = 1 - (\frac{1}{N_t}) \sum_{x_i \in T} E_i \tag{5}$$

Another criterion was used to evaluate the compactness of the clusters: within-cluster sum of squared error (WCSS), as defined in Formula (1). This will be calculated for the training set ($WCSS_{training}$) and also for the whole dataset ($WCSS_{total}$) in the clustering step.

5 Case Study

The motivation to this research comes from a collaboration project with the City of Calgary's Waste and Recycling Services (WRS). The overall goal of the project is to develop a decision support system for strategic and operational planning of the services offered to the households. The services today include residential waste, recycling, and organics collection. There are costs and qualities associated with each type of service. The decisions made by the WRS managers are aimed to optimize the cost of the services and at the same time keep the quality of the services acceptable to the customers. Several attributes are considered in the decision making process, such as budget, technology, environment, waste amount, participation of people in the recycling programs etc.

An important cost factor in WRS is the operational cost of the collection process. In order to optimize the resources, the resource planner at WRS needs to pick a certain area of the city, based on the estimated workload, and assign it to a collection vehicle. Each area is called a 'beat'. There are 110 beats in the City of Calgary. The amount of waste is an important factor when estimating the workload of the beats. So, improved estimations on waste amounts, will lead to improved resource utilization, and saving in time and fuel.

The beats are re-designed when new neighborhoods are added to the city or improvements are made to the current designs. The prediction model would help designing more efficient beats which would ultimately reduce the operational cost.

5.1 Dataset and Attribute Selection and Weighting

The dataset contains records on residential waste with 101 attributes. The attributes represent two types of information for each record. First, information regarding the collection, which includes 21 attributes on the waste volume, date, travelled distance of the collection trucks etc. This information is updated on a daily basis. Second, information regarding the collection area (beat), including 80 attributes, which are mainly about the number of different types of housing, e.g. number of duplexes, number of apartments less than 4 units etc. This information is updated on a yearly basis. The dataset used in this study includes 63,000 records for a period of two and a half years.

The attribute selection, by using Weka [24], recognized 37 attributes as the most important from the original 101. A large portion of the selected attributes describe the type of housing within a beat. Examples are the number of town houses with 8 or more units, the number of mobile homes, and beat design unit. The latter is a measurement, used by the city to determine the size of the beats. The weights of the attributes vary from 0.25 to 38, with most falling in the range of 1 to 5. For a list of attributes, their definitions, and weights, see [25].

5.2 Partitioning Using Domain Knowledge

Using the domain knowledge of the WRS managers, we decided that a feature that could be used for partitioning is holidays. It is natural to assume that there will be varying amounts of garbage produced during special occasions. For instance, during the Christmas holidays, more garbage should accumulate in households from the packaging of the gifts. During a long weekend over the summer, there should be less garbage since many families take vacations and leave the house.

The partitioning was done by separating the data for the week after Christmas and New Year from the original dataset. The Christmas records included 788 records. The average waste amount for these records was 8325 kg/beat compared to 11452 kg/beat for the whole dataset. For the week following the New Year, 1565 records were filtered out with the average waste amount of 12408 kg/beat.

A linear regression model has been created for each of the Christmas and New Year clusters. 10-fold cross validation of the models showed an average accuracy of 79% and 86% for the Christmas and New Year clusters, respectively.

5.3 Creating the Clusters and the Final Predictors

The next step in the process is creating the clusters. The number of the clusters (k) is an input to k-mean clustering and normally is determined by examining different numbers. We study the impact of varying the number of clusters from 5 to 30, each time increasing by 5. In each configuration, the final predictors were created for each cluster using linear regression. So, six experiments were conducted and the accuracy and standard deviation of the predictions was measured for each configuration.

As shown in Fig. 3 and Fig. 4, the highest accuracy was achieved when using 25 clusters, reaching to 83%. It should be noted that 3-fold cross validation on a randomized dataset has been used to measure the accuracy of the hybrid method.

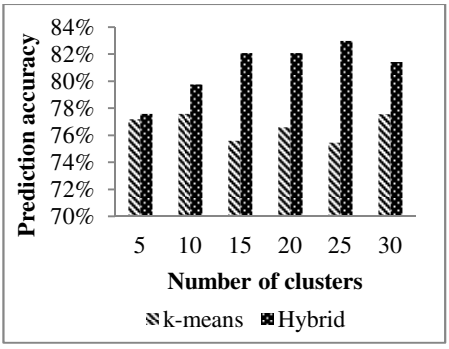


Fig. 3. Prediction accuracy of k-means clustering and Hybrid

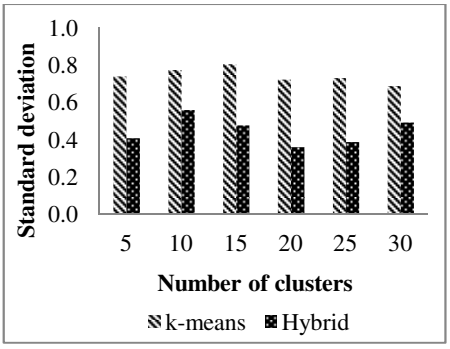


Fig. 4. Comparison of the standard deviation of the predictions

5.4 Comparison and Analysis

The comparison of the results obtained from the application of the hybrid method with the results from k-means clustering is shown in Fig. 3 and Fig. 4. It shows that in all cases our hybrid method makes more accurate predictions than a single k-means clustering. Also it has a lower standard deviation, which means that the predictions are more robust.

The accuracy of the linear regression method has been calculated using 10-fold cross validation [23] by randomizing the dataset into 10 folds. So, 10 runs have been performed, and in each one 9 folds were considered as training set and 1 fold as the test set. The results showed an average accuracy of 80.4% and standard deviation of 0.58 for the 10 runs. This shows that the accuracy of the hybrid method is higher that linear regression as well.

Another observation is made by comparing the linear regression with the Christmas and New Year records. While the prediction accuracy for the Christmas records (79%) is very close to the linear regression (80.4%), the New Year records are predicted more accurately (86%). This shows that the partitioning of the special records does improve the predictions for the New Year cluster and slightly for Christmas records.

Fig. 5 shows the comparison of weighted and unweighted clustering with respect to the prediction accuracy. In all cases the weighted clustering made more accurate predictions than the unweighted one. This shows that the complexity of adding the attribute weighting at the beginning of the method pays off with making more accurate predictions. Fig. 6 compares the within-cluster sum of squared errors, calculated as Formula (1), for the training set and the whole dataset, for the weighted k-means clustering. Comparing Fig. 5 and Fig. 6 shows that although the clusters are getting more compact, as getting smaller, the accuracy does not increase necessarily after a certain point (25 clusters). This is because of the instances being over-fitted in the clusters.

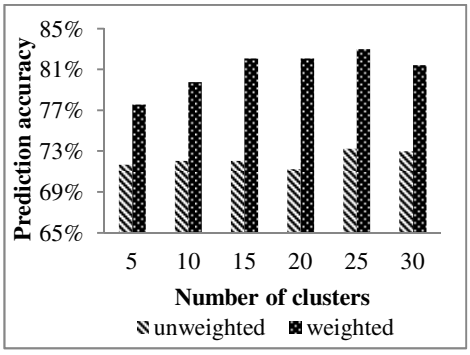


Fig. 5. Comparison of weighted and unweighted clustering

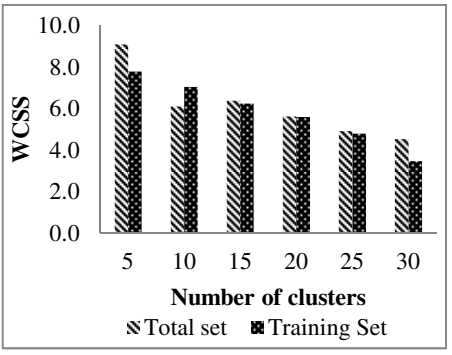


Fig. 6. Within-cluster sum of squared errors for the k-means clustering

In summary, the following answers are given to the research questions raised earlier in Section 3.3:

- Partitioning using domain knowledge increases the accuracy of the predictions, e.g. by 6% for the New Year records when using linear regression
- Using weighted attributes increases the accuracy of the predictions, e.g. by 10% when using 25 clusters
- Increasing the number of clusters improves the accuracy of the predictions until a certain point (here 25 clusters) and will reduce the accuracy afterward
- The hybrid method outperforms k-means clustering and linear regression in terms of accuracy (83% comparing to 77.6% and 80.4%) and standard deviation (0.39 comparing to 0.73 and 0.58) of the predictions

6 Discussion on Applicability

There are more factors which need to be considered, in addition to the accuracy of the predictions analyzed so far. The implementation of the prediction method and its performance, validity, and integration influence the applicability of a machine learning method. These factors are explained in this section.

Implementation and Performance. Implementation of the prediction method and its tool support plays an important role in the applicability of it. Automating the computations and providing a user friendly interface for the solution facilitate deploying the methodology. We used the Java libraries from the open source tool Weka [24] to build the prediction model. The code is easily transferable across different machines and can serve as a software service for prediction purposes.

The performance of our prediction method in terms of the computation time was acceptable to the industry. Building the prediction model from the dataset of 63,000 records takes 20 minutes on a PC machine with 4 GB RAM. This step happens once in a while depending on the problem domain. For example, in our case study the new data records are added every month to the dataset, when the model is updated and new clusters are formed. Then, the new projections are made for the future months, which depending on the prediction period take 1-3 minutes. Although there is not a graphical user interface for the prediction method, it can be easily done as the existing code runs as a standalone component.

Threats to Validity. Four types of validity threats are explained here: internal, construct, conclusion, and external.

There are some threats to the internal validity of the results, such as quality of the data and reliability of the utilized methods. There are always noises and incomplete data, which need to be removed before performing the computations. Also, when the data is integrated with other factors such as environment, or demographic factors, the time period of those factors need to be aligned. For example, in our case study, the designs of the beats change every few months. So, when integrating that information

with the waste records, the time period of each beat design must match with the timestamp on the waste records. In terms of the utilized methods, k-means clustering and linear regression, their underlying implementation impacts the prediction accuracies. We used Weka, which is a well-known tool in this context, to generate reliable results.

The construct validity of the prediction model depends on the measurements used for comparing and evaluating different techniques. Prediction accuracy, standard deviation, and clustering robustness are used in the literature for this purpose. Those measurements were used in this paper as well and seem to be enough for the evaluation purposes.

There is a threat to the validity of the stated conclusions. Several experiments have been performed to measure the correlation of the prediction accuracy with number of clusters. Other conclusions were made by comparing weighted and unweighted attributes, and also using isolated techniques. In all cases, the number of experiments is rather low (6-10), although the observations were always consistent among them. Further experiments with more data points (e.g. trying 100 different numbers of clusters) would provide more reliable conclusions.

Finally, the external validity of the prediction model is influenced by its reproducibility and applicability to other domains. The proposed prediction model and its steps are totally replicable exactly as stated in Section 4. The methodology does not depend on the underlying dataset or domain. So, though has not done yet, we believe that the methodology is transferable across various domains.

There are some factors that can limit the application of the hybrid prediction method, such as scale of the problem and insufficient information. Although the dataset used in the case study contained 63,000 records, evaluating with larger datasets (in the order of millions of records) would show its scalability better. Insufficient or inaccurate information can also limit the applicability of the prediction method by reducing the prediction accuracy.

The data attributes, their type, and storage method always change in industry. So the question is how the methodology, and even the software solution, accommodates those changes? This prompts for removing any dependency of the underlying methodology to the attributes and their characteristics.

Integration. Finally, how the machine learning method is integrated into the target application may be the most important factor in the applicability of it. Only making some predictions and providing numbers are not enough for the industry. The domain expert needs to know how he can make use of the prediction results, how the results are justified, and what can be learnt from them.

In our case study, several patterns have been extracted from the data, such as seasonal patterns e.g. production of more waste over the summer. Predicting this amount and, potentially new patterns, for the future helps the domain expert in the planning process.

The clustering is another example where new knowledge has been extracted. The clusters represent the beats with similar behavior in terms of generating waste. Although the domain expert does not need to know how they are generated,

representing the clusters and illustrating the differences among them helps him when making decision on the beats. For example, when a strategy is going to be applied to a certain part of the city, similar beats are expected to respond similarly.

Another potential benefit comes from the improved predictions of waste amounts. This would help the decision maker design more efficient beats, regarding the collection time and fuel consumption. For example, consider the following hypothetical scenario:

1. New beats are designed to improve the collection efficiency
2. The decision maker uses the hybrid prediction method to project the waste amount for the upcoming year
3. They find that in 20 beats, out of 110, the trucks will not be fully loaded on their second trip to the beats.
4. They re-design those beats and their neighbor beats in order to utilize the truck capacities more efficiently
5. They repeat steps 2-4 until an acceptable design is achieved
6. The city saves 5 extra trips by the vehicles per day because of this pro-active planning

Although the above scenario is hypothetical, currently research is undergoing to prove its usage, and evaluate that usage with real data. The proposed prediction method in this paper will be integrated into a decision support system for strategic and operational planning of waste and recycling services [26]. The system analyses strategic and operational decisions on the services and provides the estimated cost, quality, and risk associated with them. The prediction model will improve the reliability of the system by estimating the amount of waste and performing the computations on the projected amount. The details of this integration are beyond the scope of this paper and are currently under evaluation.

7 Conclusion

Hybrid prediction methods that integrate clustering and classification techniques have been applied to many prediction problems, especially for large datasets. In this paper a particular hybrid prediction method is proposed by integrating k-means clustering and linear regression. Firstly, using linear regression, a set of attributes is selected from the whole attribute set to reduce the computational complexity. Also, weights are assigned to the attributes by the Greedy method in linear regression. Then, using domain knowledge, some special records are separated from the main dataset forming their own partitions. Next, weighted k-means clustering is applied to the rest of the dataset to create clusters of similar records. Then, a linear regression model is created for each cluster and serves as the final predictor.

The proposed hybrid prediction method was applied to the problem of municipal waste prediction, through a collaboration project with the City of Calgary's Waste and Recycling Services (WRS). The history data of the household waste included 63,000 records and the need for an efficient prediction model deemed to be necessary. Domain knowledge has been extracted through meetings with WRS managers.

The experimental results showed that the proposed hybrid prediction method outperforms a single k-means clustering or linear regression model in terms of prediction accuracy. Also, using the weighted attributes and partitioning based on domain knowledge improves the accuracy of the predictions. The performance of the proposed solution and its dynamic design made it easier to use by the domain experts. This was part of the integration method considered to transfer the methodology to the target industry.

The prediction method will be integrated into a decision support system for waste and recycling services. The goal of the decision support system is to help the domain expert when evaluating the strategies on the services, and allocating the operational resources. Improved predictions will potentially increase the efficiency of collections by saving time and fuel, which will eventually reduce the operational cost. Evaluating this with real data is a part of our future work.

Machine learning methods have been studied for long time and applied to many domains. However, the actual usage of them in industry and their real impact are affected by factors such as performance, reliability, and openness to change. These challenges have been addressed in this research. Future research can be performed on applying the prediction method to other types of waste such as recyclables and organics. So, more accurate estimations can be made on the cost of these services and the potential income from them. Also, more research can be done on applying the hybrid prediction method, integrated into a decision support system, to other logistics domains such as postal services, snow removals, and delivery services.

Acknowledgement. We would like to thank the Natural Sciences and Engineering Research Council (NSERC Canada, CRD#386808-09) and The City of Calgary's Waste and Recycling Services (WRS) for their financial support of this research. In particular, we acknowledge the support of the WRS team when running the case study.

References

1. Han, J., Nishio, S., Kawano, H., Wang, W.: Generalization-based data mining in object-oriented databases using an object cube model. Data & Knowledge Engineering 25(1), 55–97 (1998)
2. Han, J., Kamber, M.: Data mining: Concepts and techniques. Morgan Kaufmann Publishers (2006)
3. Witten, H.I., Frank, E.: Data mining: Practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann Publishers (2005)
4. Patil, B.M., Joshi, R.C., Toshniwal, D.: Hybrid prediction model for Type-2 diabetic patients. Expert Systems with Applications 37(12), 8102–8108 (2010)
5. Tsai, C.-F., Lu, Y.-H.: Customer churn prediction by hybrid neural networks. Expert Systems with Applications 36(10), 12547–12553 (2009)
6. Lenard, M.J., Madey, G.R., Alam, P.: The design and validation of a hybrid information system for the auditor's going concern decision. Journal of Management Information Systems 14(4), 219–237 (1998)

7. Jain, A., Murty, M., Flyn, P.: Data clustering: A review. ACM Computing Surveys 31, 264–323 (1999)
8. Purcell, M., Magette, W.L.: Prediction of household and commercial BMW generation according to socio-economic and other factors for the Dublin region. Waste Management 29(4), 1237–1250 (2009)
9. Daskalopoulos, E., Badr, O., Probert, S.D.: Municipal solid waste: a prediction methodology for the generation rate and composition in the European Union countries and the United States of America. Resources, Conservation and Recycling 24(2), 155–166 (1998)
10. Dyson, B., Chang, N.B.: Forecasting municipal solid waste generation in a fast-growing urban region with system dynamics modeling. Waste Management 25(7), 669–679 (2005)
11. Grieu, S., et al.: Prediction of parameters characterizing the state of a pollution removal biologic process. Engineering Applications of Artificial Intelligence 18(5), 559–573 (2005)
12. Goia, A., May, C., Fusai, G.: Functional clustering and linear regression for peak load forecasting. International Journal of Forecasting 26(4), 700–711 (2010)
13. Kusiak, A., Li, W.: Short-term prediction of wind power with a clustering approach. Renewable Energy 35(10), 2362–2369 (2010)
14. Márquez, M.Y., Ojeda, S., Hidalgo, H.: Identification of behavior patterns in household solid waste generation in Mexicali's city: Study case. Resources, Conservation and Recycling 52(11), 1299–1306 (2008)
15. Jalili, M., Noori, R.: Prediction of municipal solid waste generation by use of artificial neural network: a case study of Mashhad. International Journal of Environmental Research 2, 13–22 (2008)
16. Noori, R., Abdoli, M.A., Farokhnia, A., Abbasi, M.: Results uncertainty of solid waste generation forecasting by hybrid of wavelet transform-ANFIS and wavelet transform-neural network. Expert Systems with Applications 36, 9991–9999 (2009)
17. Noori, R., Karbassi, A., Sabahi, M.: Evaluation of PCA and Gamma test techniques on ANN operation for weekly solid waste prediction. Journal of Environmental Management 91(3), 767–771 (2010)
18. Guojun, G., Chaoqu, M., Jianhong, W.: Data clustering theory algorithm and application, 1st edn. ASA-SIAM, Society for Industrial and Applied Mathematics (2007)
19. MacQueen, J.B.: Some Methods for classification and Analysis of Multivariate Observations. In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297. University of California Press (1967)
20. Tibshirani, R.: Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological) 58(1), 267–288 (1996)
21. Berk, R.A.: Regression Analysis: A Constructive Critique (Advanced Quantitative Techniques in the Social Sciences), 1st edn. Sage Publications, Inc. (2003)
22. Briand, L.C., Wieczorek, I.: Resource Estimation in Software Engineering. Encyclopedia of Software Engineering 2, 1160–1196 (2001)
23. Picard, R., Cook, D.: Cross-Validation of Regression Models. Journal of the American Statistical Association 79(387), 575–583 (1984)
24. Weka SVN Repository, `https://svn.scms.waikato.ac.nz/svn/weka/` (accessed January 2013)
25. Livani, E.: `http://www.ucalgary.ca/~elivani/HybridPrediction.xls` (accessed January 2013)
26. Livani, E., Paikari, E., Ruhe, G.: Decision Support System for Cost-benefit Analysis in Service Provision. In: Proc. ICEIS, vol. 2, pp. 198–203 (2011)

BiETopti-BiClustering Ensemble Using Optimization Techniques*

Geeta Aggarwal and Neelima Gupta

Department of Computer Science,
University of Delhi,
Delhi - 110007
India
ggupta@pgdav.du.ac.in, ngupta@cs.du.ac.in

Abstract. In this paper, we present an ensemble method for the biclustering problem that uses optimization techniques to generate consensus. Experiments have shown that the proposed method provides superior bi-clusters than the existing bi-clustering solutions most of the times. Bi-clustering problem has many applications including analysis of gene expression data.

Keywords: Bi-clustering, Cluster Ensemble, Optimization Techniques.

1 Introduction

To identify genes involved in a particular biological activity one needs to analyse large amount of gene expression data [1–5]. However, most of the genes responsible for one biological activity are responsive only to a small subset of samples rather than the entire set of conditions. Traditional clustering algorithms fair poorly in identifying such clusters of genes. Bi-clustering refers to simultaneous clustering of genes and samples, where-in bi-clusters may overlap on genes or samples or both. Thus bi-clustering the gene expression data where-in genes and samples may be responsible for more than one biological activity and hence may belong to more than one bi-cluster, turns out to be a better tool to identify such groups of genes.

Various approaches used to design algorithms for identifying bi-clusters lack robustness and stability with respect to the random initialization. Also, these algorithms aim to optimize different objective functions leading to different solutions. To an end user having no clue about which objective function is most suitable for the application, the choice of a particular algorithm becomes difficult. Ensemble methods aims to provide solutions which are more robust toward random seeds. In this paper, we present the bi-clustering ensemble problem as an optimization problem and show the performance of the approach on gene expression data.

* This project is supported by University of Delhi.

P. Perner (Ed.): ICDM 2013, LNAI 7987, pp. 181–192, 2013.

Combining bi-clustering solutions is a more challenging task than the combining unsupervised clusterings or supervised classifications, as the bi-clusters from two input schemes may involve different sets of samples. To be able to align bi-clusters one needs to factor in the similarity of the two bi-clusters on the conditions as well. In the absence of labeled training data, different labels are assigned to the bi-clusters by different schemes. To establish a correspondence between them one needs to solve a problem of k dimensional matching which is known to be NP-hard for $k \geq 3$. To make the problem more tractable one of the bi-clustering scheme is fixed as a reference and other bi-clusters are aligned with it. Hungarian method [6] can be used to compute the minimum weight bi-partite matching for label correspondence. However, accuracy of the established correspondence depends on the manner in which weights are assigned to the edges. [7] defined a probabilistic model to assign the edge-weights for clusters. We use a modified version of their model to suit the need of bi-clusters. After the alignment, we formalize the problem of generating the consensus as an optimization problem that simultaneously takes gene-wise as well as sample-wise perspective of the problem into consideration. This is the first time that optimization techniques have been used for bi-cluster ensemble wherein bi-clusters may overlap both on genes and conditions simultaneously and, a gene and a condition may belong to more than one bi-cluster.

We have used multiple runs of Iterative Signature Algorithm [1] to generate different input bi-clustering schemes. ISA works on the hypothesis that although the set of possible input seeds is huge, usually there are only a limited number of fixed points for a given set of thresholds. They run the algorithm for a large number of input seeds and reconstruct the modules from the recurring fixed points by fusing the solutions that were distinct but very similar, using a procedure that resembles agglomerative clustering.

Experimental studies were performed on the benchmark datasets of Prelic et al. [4] and the expression data of human Diffuse Large B-cell Lymphoma (DLBCL). BiET outperforms the input schemes most of the times for both the synthetic data sets. On DLBCL also, most of the bi-clusters show remarkable improvement over the input schemes. To study the statistical significance of the bi-clusters Average Rand Index (AvRI) was used on the synthetic data sets. In the absence of ground truth GO terms and motif analysis were used to compare the performance on the real datasets.

Remaining paper is organized as follows: the problem is defined in section 2. Section 3 discusses the related work. Our approach is presented in section 4. The experimental results are presented in section 5. The paper is concluded with the future work in section 6.

2 Problem Definition

Let G be a set of N genes and C be a set of d samples/conditions. Let E be an $N \times d$ expression matrix where each row represents the expression of a gene under d samples.

Definition 1. (Biclustering Solution) *A biclustering solution π defined over E is a triplet (k, X, Y):*

1. *$\{1 \dots k\}$ denote the set of bicluster labels.*
2. *$X : G \times \{1 \dots k\} \rightarrow \{0, 1\}$ is a function where $X(g, l)$ represents whether a gene g belongs to the bicluster labeled l such that $\sum_{l=1}^{k} X(g, l) \geq 1$.*
3. *$Y : C \times \{1 \dots k\} \rightarrow \{0, 1\}$ is a function where $Y(c, l)$ represents whether a condition c belongs to the bicluster labeled l such that $\sum_{l=1}^{k} Y(c, l) \geq 1$.*

For the ease of presentation we'll use x_{gl} to denote $X(g, l)$ and y_{cl} to denote $Y(c, l)$.

Definition 2. (Input Schemes) *is a collection $\Pi = \{\pi_1 \dots \pi_H\}$, where each of $\pi_i = (k_i, X^i, Y^i)$ is a biclustering solution.*

Definition 3. (Global Label Set) *is a set L of labels $\{1 \dots \sum_{i=1}^{H} k_i\}$.*

In general, different bi-clustering schemes may contain different number of bi-clusters. However, in this paper we have considered schemes with equal number of biclusters i.e. $k_i = k \ \forall \ i$. Since the bi-clusters may overlap both on genes and conditions, a gene (/ condition) may be assigned more than one label. Also, there may be a gene (/ condition) which does not belong to any bi-cluster, such a gene (/ condition) is assigned a special label 0. Hence, without loss of generality, we assume that in each scheme, each gene belongs to at least one bicluster and each sample belongs to at least one bicluster.

Definition 4. Bicluster Collection Representations
$\forall h \in L$, let $i = h/(k+1)+1$, $r = h \ mod \ (k+1)$, the gene-wise representation of the collection of input schemes is given by the vectors
$\lambda_h =< x^i_{g_1 r}, x^i_{g_2 r}, \dots x^i_{g_N r} >$

and the condition-wise representation is given by the vectors
$\mu_h =< y^i_{c_1 r}, y^i_{c_2 r}, \dots y^i_{c_d r} >$

The problem of bicluster ensemble is to derive a consensus that combines the H biclustering solutions and delivers a biclustering solution $\hat{\pi} = (k, X, Y)$ that achieves one or more of the following aims:

1. It improves the quality of the bi-clusters.
2. It is more robust and stable than its constituent schemes.

3 Related Work

Bagging and boosting [8–11] are standard techniques to obtain ensembles for the classification problem. Two approaches are largely used to design consensus functions in clustering. One that establishes label correspondence between the various partitions and then uses a consensus function; second that eliminates the

need of label correspondence and computes the consensus function directly. Most of the work [12–17] fall in the second category. [7, 18] fall in the first category. Several approaches have been used in literature to design consensus functions: majority voting, co-association, fusion using procedure similar to agglomerative clustering, graph partitioning, statistical and information theoretic methods. A detailed survey of consensus functions can be found in [19].The adaptive clustering ensemble technique proposed in [18] uses sampling techniques to generate individual partitions of the ensemble. Krumpleman and Ghosh [7] are the first ones to define a statistical measure (p-value) to capture the overlap between two clusters, wherein a clusterer may produce non-disjoint clusterings, to establish cluster correspondence before applying majority voting to design the consensus function for the ensemble. Optimization techniques [20, 21] have also been used to generate ensembles for clustering and projective clustering.

Wang et al. [22] and Gullo et al. [20] have proposed ensemble techniques for the related problems of co-clustering and projective clustering respectively. Though researchers sometimes claim that co-clustering, projective clustering and bi-clustering are all same but generally solutions for co-clustering do not allow clusters to overlap on objects and features whereas solutions for projective clustering allows overlapping of features but not of objects. Wang et al. presented an ensemble solution for co-clustering wherein they extract block-constant bi-clusters generalizing the grid-style partitions to allow different resolutions in different parts of the data matrix. A pair of bi-clusters may overlap on objects or on features but not on both at the same time. Gullo et al. presented an ensemble solution for projective clustering where in an object may belong to more than one bi-clusters but the total sum of the membership is one thereby meaning that if an object completely belongs to one bi-cluster it does not belong to any other. They project the clusters on one dimension in a fuzzy way. Thus bi-clustering is different from these problems/solutions wherein an object/feature may have a total membership more than one and a bi-cluster is defined by more than one feature. Also, bi-cluster may overlap both on objects and features simultaneously. Hanczar and Nadif [10] have proposed the use of bagging to improve the performance of bi-clustering schemes.

4 Our Approach

Our algorithm works in two phases. In phase-I we align two binary membership matrices using Hungarian method to compute the minimum weight bipartite matching. The p-measure defined by Krumpleman and Ghosh, to compute the weights on the edges, takes into account the non-disjointness of bi-clusters but it looks at the overlapping only amongst the genes. Modified p-measure to suit the definition of bi-clusters is used for the purpose. The consensus is generated using optimization in phase-II.

4.1 Phase I: Label Correspondence

In this phase we wish to align bi-clusters having maximum overlap with each other. In the absence of labeled data, one of the schemes is used as a reference and bi-clusters of other schemes are relabeled so that bi-clusters of the two schemes having maximum overlap are aligned with each other. This is the intuition underlying the probability based alignment function proposed in [7]. Borrowing the notation from them, we modify the definition of binary membership matrix to take into account the joint overlapping of bi-clusters on genes and conditions. A binary membership matrix M is a three dimensional matrix with genes, conditions and labels on the three dimensions. Hence each membership matrix is of size $N \cdot d \cdot k$ for an input scheme with k bi-clusters. Using the notation of a 3D matrix, m_{ijk} denotes whether the gene g_i and the condition $c_j$belong to the k^{th} bi-cluster or not.

Let M_{ref} and M_l be two binary membership matrices for the input schemes π_{ref} and π_l ($l = 1 \ldots H$). The objective is to align the third dimension of M_l with that of M_{ref} such that they have maximum match. Hamming distance appears to be the most reasonable choice. However, it does not account for the different densities of 1's in different columns. Thus Krumpleman and Ghosh suggested a probabilistic model (p-value) to capture the probability of an observed overlap. We use the following modified definition of the p-value:

Let BC_i and BC_j be the two bi-clusters, whose overlap is to be computed, of π_{ref} and π_l respectively . Let $N^{'}$ be the total number of (g,c) pairs. Let d_1 and d_2 be the number of (g,c) pairs in BC_i and BC_j respectively, i.e. $d_1 = (\sum_g x_{gi} \sum_c y_{ci})$ and $d_2 = (\sum_g x_{gj} \sum_c y_{cj})$. Let s be the number of (g,c) pairs common between BC_i and BC_j. This can be obtained by taking a scalar product of the corresponding columns (third dimension) of the two matrices.

The probability of the observed overlap being s is then given by

$$P_{ij}(s) = \frac{\binom{N^{'}}{d_1}\binom{d_1}{s}\binom{N^{'}-d_1}{d_2-s}}{\binom{N^{'}}{d_1}\binom{N^{'}}{d_2}} = \frac{\binom{d_1}{s}\binom{N^{'}-d_1}{d_2-s}}{\binom{N^{'}}{d_2}}$$

The above expression is the hyper-geometric distribution evaluated at s. Total probability of seeing at least s overlapping (g,c) pairs is then

$$p_{ij} = \sum_{t=s}^{min(d_1,d_2)} P_{ij}(t)$$

Clearly, higher is the overlap, lower is the p-value. We compute the pairwise p_{ij}-values for all the pairs of bi-clusters with one bi-cluster taken from π_{ref} and the other from π_l and, look for minimum weight bipartite matching with p_{ij} values as the edge weights to obtain the new labels ($X^{'}Y^{'}$) for the bi-clusters of

π_l. This is done by Hungarian method. The corresponding representation of the input collection after the alignment would be given by

$\lambda_h^{'} =< x_{g_1 r}^{i'}, x_{g_2 r}^{i'}, \ldots x_{g_N r}^{i'} >$
$\mu_h^{'} =< y_{c_1 r}^{i'}, y_{c_2 r}^{i'}, \ldots y_{c_d r}^{i'} >$

The algorithm for label correspondence is given in Algorithm 1.

Input: Membership matrices $M_1, M_2 \ldots M_H$ of H input schemes
Output: $(X^{1'}Y^{1'}), (X^{2'}Y^{2'}), \ldots, (X^{H'}Y^{H'})$: new labels of H input schemes
`/* without loss of generality assume` π_{ref} `is` π_1. `*/`
for $l = 2$ **to** H **do**
  for $i = 1$ **to** k **do**
    for $j = 1$ **to** k **do**
      calculate p_{ij} between the i^{th} and the j^{th} bi-clusters of π_{ref} and π_l bi-clustering schemes respectively;
    end
  end
  use Hungarian algorithm with p_{ij}-values as the edge weights and relabel the bi-clusters of π_l;
end

Algorithm 1. Label Correspondence

4.2 Phase II: Generating Consensus

In this section we present our main contribution. We formulate the problem of generating a consensus from the input solution schemes as an optimization problem. Let x_{gr} be an indicator variable that is 1 if the gene g gets label r and 0 other wise. For a fixed label r and a gene g, λ_{hg} denotes whether g belongs to the r^{th} aligned bicluster. Similarly, let y_{cr} be an indicator variable that is 1 if the condition c gets label r and 0 otherwise. For a fixed label r and a condition c, μ_{hc} denotes whether c belongs to the r^{th} aligned bicluster. Different values of h for which $h \ mod \ (k+1) = r$ holds represent aligned biclusters of different input schemes. The objective is to assign labels to genes and conditions so as to minimize the dissimilarity of the obtained bicluster from the corresponding aligned biclusters. First term of the objective function represents the dissimilarity over genes and the second term represents the dissimilarity over conditions.

Minimize
$\sum_{r=0}^{k} \sum_{h:hmod(k+1)=r} \sum_{g=g_1}^{g_N} |x_{gr} - \lambda_{hg}| + \sum_{r=0}^{k} \sum_{h:hmod(k+1)=r} \sum_{c=c_1}^{c_d} |y_{cr} - \mu_{hc}|$

subject to

$$\sum_{g=g_1}^{g_N} x_{gr} \geq 1 \forall 0 \leq r \leq k \quad (1)$$
$$\sum_{c=c_1}^{c_d} y_{cr} \geq 1 \forall 0 \leq r \leq k \quad (2)$$

$$\sum_{r=1}^{k} x_{gr} \geq 1 \, \forall g \in G \quad (3)$$
$$\sum_{r=1}^{k} y_{cr} \geq 1 \, \forall c \in C \quad (4)$$

The constraints (1) and (2) make sure that each bi-cluster has atleast 1 gene and 1 condition. The constraints (3) and (4) make sure that each gene and condition will belongs to at least one bi-cluster. The coefficient matrix of our optimization problem has a nice property of being unimodular. We used the non-linear programming solution in LINGO to solve the problem and the solution obtained was always integral.

5 Experimental Results

5.1 Datasets

We performed experimental studies on two benchmark synthetic datasets presented in Prelic et al. and the real dataset of DLBCL. The first synthetic dataset (DS1) was of size 110×110 consisting of 11 overlapping bi-clusters. To study the effect of noise, they created another synthetic dataset (DS2) of size 100×50 consisting of 10 non-overlapping bi-clusters. DLBCL dataset consists of 661 genes and 180 conditions.

5.2 Methodology

We implemented our algorithm BiETopti in Extended LINGO/Win32 13.0.2.16 on Intel Core i5-2430M CPU @2.40 Ghz with 4GB RAM using Windows 7 Home Basic Operating System. Input schemes were generated by running ISA on the expression data five times, each time with 100 gene seed vectors. The schemes were preprocessed to remove the bi-clusters with high overlap ($> 80\%$). This was done to remove the redundancy. Further genes and conditions not assigned to any bi-cluster by any of the input schemes are also removed. This was done to ensure the feasibility of the input. BiETopti code was then executed to get the final ensemble. Two sets of experiments were conducted on synthetic datasets of Prelic et al. In the first set, the thresholds (t_g, t_c) were fixed and the schemes were generated by running ISA with different gene seed vectors. In the second set of experiments, t_g was varied keeping both the gene seed vector and t_c fixed.

5.3 Performance Evaluation of Bi-clusters

Synthetic Datasets. Rand Index (RI) is a measure of similarity between two clusterings. Value of RI ranges between 0 and 1, with 0 indicating that the two data clusterings do not agree on any pair of points and 1 indicating that the data clusterings are exactly the same. We define AvRI, the measure of similarity between two bi-clustering schemes as an average Rand Index between the pairwise aligned bi-clusters of the two schemes. Let $AvRI_{impl}(j)$ denote AvRI between the input schemes π_j and the implanted bi-clusters π_{impl}.

Real Dataset. On real datasets, in the absence of ground truth, we cannot use AvRI to validate our bi-clusters. Hence, we use the domain knowledge to determine the biological significance of the bi-clusters. We validate our bi-clusters using functional annotation GO (Gene Ontology) and common patterns (motifs) in the promoter regions of the genes of a bi-cluster with the help of biological tools DAVID and RSAT available on line. p-value is used to determine the biological significance of the annotated GO term and the motifs. p-value represents the probability of seeing x or more genes from the input list of n genes annotated to a particular GO term, given the proportion of genes in the whole genome annotated to that GO Term is f out of m.

$$p - value = \sum_{j=x}^{n} \frac{\binom{f}{j}\binom{m-f}{n-j}}{\binom{m}{n}}$$

Smaller the $p - value$, more significant is the association of the particular GO term with the group of genes (i.e. it is less likely that the observed annotation of the particular GO term to a group of genes occurs by chance). There may be several GO terms with different $p - values$ associated with an input set of genes belonging to a bi-cluster. The best $p - value$ is used to compare the bi-clusters.

5.4 Results on Synthetic Data Sets

Table 1 compares the $AvRI_{impl}$ of the input schemes and that of the bi-clusters produced by BiETopti on DS1. The table shows the results for the first set of experiments i.e when the schemes are generated by varying seed vectors for fixed (t_g, t_c) values. We chose the scheme with maximum number of bi-clusters as a reference scheme.

Table 2 compares the $AvRI_{impl}$ of the input schemes and that of the bi-clusters produced by BiETopti on DS1 for the second set of experiments i.e when the schemes are generated by varying t_g for a fixed seed vector and t_c. The value of t_g was varied from $[-2.4, +2.0]$ in steps of 0.2. It was observed that for t_g values ranging from $[0.6, 1.6]$ schemes with bi-clusters identical to the implanted bi-clusters were obtained whereas schemes obtained for t_g varying from $[-2.4, -0.8]$ bi-clusters consisted essentially of all the genes and all the bi-clusters eventually reduced to a single bi-cluster after preprocessing. Thus we focused our study on t_g varying from $[-0.6, 0.4]$ and $[1.8, 2.0]$.

Both Tables 1- 2 show that BiETopti improves upon the performance of the individual input schemes.

Effect of noise: Noisy dataset (DS2) of Prelic et al. was used to study the impact of noise on the performance of BiETopti . Table 3 shows the results for the first set of experiments i.e when the schemes are generated by varying seed vectors for fixed (t_g, t_c) values and Table 4 gives the results for the second set of experiments i.e when the schemes are generated by varying t_g for a fixed seed vector and t_c. The tables show that BiETopti is able to improve the performance of the individual input schemes even in presence of noise.

Table 1. Comparison of $AvRI_{impl}$(input schemes) and that of BiETopti on DS1 when schemes are generated by varying seed vectors for fixed (t_g, t_c) values

Schemes → $t_g, t_c \downarrow$	$\pi 1$	$\pi 2$	$\pi 3$	$\pi 4$	$\pi 5$	**BiETopti**
-0.50 , 2	81.9	82	71.5	66.6	73.4	81.4
-0.40 , 2	77.7	77.9	77.6	63.2	70.7	78.2
-0.35 , 2	89.4	89.4	89.4	89.4	73.4	90.1
1 , 1	68.4	61.5	53.3	69.3	63.4	69.5
0 , 1	53.8	47.7	55.5	36.7	50.4	55.5

Table 2. Comparison of $AvRI_{impl}$(input schemes) and that of BiETopti on DS1 when schemes are generated by varying t_g for a fixed gene seed vector and fixed t_c

Schemes →	$\pi 1$	$\pi 2$	$\pi 3$	$\pi 4$	$\pi 5$	$\pi 6$	$\pi 7$	$\pi 8$	**BiETopti**
AvRI	80	83	48	80	65	56	45	25	80

Table 3. Comparison of $AvRI_{impl}$(input schemes) and that of BiETopti on DS2 when schemes are generated by varying seed vectors for fixed (t_g, t_c) values

Schemes → $t_g, t_c \downarrow$	$\pi 1$	$\pi 2$	$\pi 3$	$\pi 4$	$\pi 5$	**BiETopti**
.90, 1	87.8	76.8	86.8	65.9	85	88
1, .5	74.7	54.8	66.1	65.9	77.8	78.1
-.35, 2	50.9	50.8	50.9	50.9	50.5	50.9

Table 4. Comparison of $AvRI_{impl}$(input schemes) and that of BiETopti on DS2 when schemes are generated by varying t_g for a fixed gene seed vector and fixed t_c.

Schemes →	$\pi 1$	$\pi 2$	$\pi 3$	$\pi 4$	$\pi 5$	$\pi 6$	**BiETopti**
AvRI	89	87	74	92	87	79	89.2

Effect of Changing the reference scheme: We also studied the impact of changing the reference scheme on the performance. It was observed that there was no significant change in the accuracy of the bi-clusters on changing the reference scheme. As there was no effect of changing the reference scheme, we decided to choose the scheme with maximum number of bi-clusters as the reference scheme.

5.5 Results on Real Datasets

Experimental studies were performed on the expression dataset of DLBCL. We generated five input schemes by running ISA five times, each time with hundred different gene seed vectors. Sizes of the bi-clusters were kept to be comparable to eliminate the effect of size of bi-clusters on the $p - values$. The bi-clusters produced by BiETopti were evaluated using DAVID and RSAT tool. Table 5 shows the top bi-clusters of DLBCL obtained from BiETopti along with their

Table 5. Comparison of top bi-clusters of BiETopti with their aligned bi-clusters of the input schemes. '-' indicates that no bi-cluster of the scheme is aligned with the reference scheme.

DLBCL : $-\log p$ value of GO terms

$\pi 1$	$\pi 2$	$\pi 3$	$\pi 4$	$\pi 5$	**BiETopti**
13	12	16	17	16	16
6	6	6	5	4	6
5	7	13	7	7	7
16	16	11	19	6	19
5	7	7	8	8	9
2	2	17	16	16	16
2	3	2	3	3	3
2	6	9	5	5	6
5	3	8	22	6	8
5	5	5	2	3	5
8	8	8	-	6	8
-	2	2	14	2	13
5	4	3	-	5	5
16	-	22	2	5	20

DLBCL : $-\log p$ value of motifs

$\pi 1$	$\pi 2$	$\pi 3$	$\pi 4$	$\pi 5$	**BiETopti**
8	10	8	10	8	11
9	15	6	6	8	9
9	7	7	8	7	9
8	8	9	9	8	9
8	6	10	7	16	15
7	10	5	10	6	10
9	29	20	20	29	21
11	6	10	10	5	11
5	10	5	7	6	11
22	15	23	22	26	23
13	10	16	-	18	18
-	7	7	9	7	10
6	7	7	-	6	8
6	-	7	7	5	8

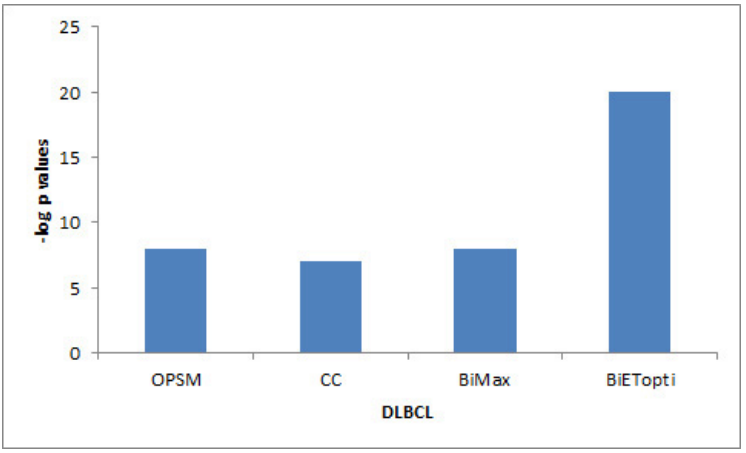


Fig. 1. Comparison of $-\log p$ values of BiETopti with that of OPSM, CC, BIMAX

aligned input bi-clusters which clearly shows that the quality of the bi-clusters obtained is better than maximum number of the input bi-clusters.

Comparison of BiETopti with other bi-clustering algorithms: Figure 1 shows the comparison of the bi-clusters produced by BiETopti with the bi-clusters produced by other bi-clustering algorithms like order-preserving sub matrix (OPSM) [23], Cheng and Church(CC) [2], BIMAX [4]. For these algorithms, bi-clusters were generated by executing these algorithms in BICAT tool. The figure shows that BiETopti outperforms all these algorithms.

6 Conclusion and Future Work

In this paper, we presented an ensemble method for the bi-clustering problem that uses optimization techniques to generate the consensus. In future, we would like to see how the algorithm performs on more data sets. We would also like to formulate the problem as a multi-objective problem and apply genetic algorithms like Particle Swarm Optimization and compare the performance.

References

1. Bergmann, S., Ihmels, J., Barkai, N.: Iterative signature algorithm for the analysis of large-scale gene expression data. In: Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics, Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel. Sven.Bergmann@weizmann.ac.il, vol. 67 (March 2003)
2. Cheng, Y., Church, G.M.: Biclustering of expression data. In: International Conference of Intelligent Systems Molecular Biology, pp. 93–103 (2000)
3. Gupta, N., Aggarwal, S.: Mib: Using mutual information for biclustering gene expression data. Elsevier Journal of Pattern Recognition 43, 2692–2697 (2010)
4. Prelic, A., Bleuler, S., Zimmermann, P., Wille, A., Buhlmann, P., Gruissem, W., Hennig, L., Thiele, L., Zitzler, E.: A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics 22, 1122–1129 (2006)
5. Tanay, A., Sharan, R., Shamir, R.: Discovering statistically significant biclusters in gene expression data. In: Proceedings of ISMB 2002, pp. 136–144 (2002)
6. Srinivasan, G.: Operations Research: Principles and Applications. Prentice-Hall of India (2002)
7. Krumpelman, C., Ghosh, J.: Matching and visualization of multiple overlapping clusterings of microarray data. In: CIBCB 2007, pp. 121–126 (2007)
8. Dudoit, S., Fridlyand, J.: Bagging to improve the accuracy of a clustering procedure. Bioinformatics 19, 1090–1099 (2003)
9. Fischer, B., Buhmann, J.M.: Bagging for path-based clustering. In: IEEE Trans. Pattern Anal. Mach. Intell., vol. 25, pp. 1411–1415. IEEE Computer Society, Washington, DC (2003)
10. Hanczar, B., Nadif, M.: Using the bagging approach for biclustering of gene expression data, vol. 74, pp. 1595–1605. Elsevier Science Publishers B.V, Amsterdam (2011)
11. Moreau, J.V., Jain, A.K.: The bootstrap approach to clustering. In: Proc. of the NATO Advanced Study Institute on Pattern Recognition Theory and Applications, pp. 63–71. Springer, London (1987)

12. Fred, A.: Finding consistent clusters in data partitions. In: Kittler, J., Roli, F. (eds.) MCS 2001. LNCS, vol. 2096, pp. 309–318. Springer, Heidelberg (2001)
13. Fred, A.L.N.: Data Clustering Using Evidence Accumulation. In: Proc. of the 16th Int'l Conference on Pattern Recognition, pp. 276–280 (2002)
14. Hu, X., Yoo, I.: Cluster ensemble and its applications in gene expression analysis. In: Proceedings of the Second Conference on Asia-Pacific Bioinformatics, APBC 2004, vol. 29, pp. 297–302. Australian Computer Society, Inc., Darlinghurst (2004)
15. Karypis, G., Kumar, V.: A fast and high quality multilevel scheme for partitioning irregular graphs. SIAM J. Sci. Comput. 20, 359–392 (1998)
16. Strehl, A., Ghosh, J.: Cluster ensembles – A knowledge reuse framework for combining multiple partitions. Journal on Machine Learning Research (JMLR) 3, 583–617 (2002)
17. Topchy, A., Jain, A.K., Punch, W.: Combining multiple weak clusterings. In: ICDM, pp. 331–338 (2003)
18. Topchy, A.P., Bidgoli, B.M., Jain, A.K., Punch, W.F.: Adaptive clustering ensembles. In: ICPR, pp. 272–275 (2004)
19. Ghaemi, R., Sulaiman, N., Ibrahim, H., Mustapha, N.: A survey: Clustering ensembles techniques. In: World Academy of Science, Engineering and Technology, vol. 38 (2009)
20. Gullo, F., Domeniconi, C., Tagarelli, A.: Projective clustering ensembles. In: Proceedings of the 2009 Ninth IEEE International Conference on Data Mining, ICDM 2009, pp. 794–799. IEEE Computer Society, Washington, DC (2009)
21. Singh, V., Mukherjee, L., Peng, J., Xu, J.: Ensemble clustering using semidefinite programming with applications. Mach. Learn. 79(1-2), 177–200 (2010)
22. Wang, P., Laskey, K.B., Domeniconi, C., Jordan, M.: Nonparametric bayesian coclustering ensembles. In: Proceedings of the Eleventh SIAM International Conference on Data Mining, SDM 2011, April 28-30, pp. 331–342. SIAM / Omnipress, Mesa (2011)
23. Ben-Dor, A., Chor, B., Karp, R.M., Yakhini, Z.: Discovering local structure in gene expression data: The order-preserving submatrix problem. Journal of Computational Biology 10(3/4), 373–384 (2003)

Multiple Buying Behavior as an Indicator of Brand Loyalty: An Association Rule Application

Diren Bulut[1], Umman Tuğba Gursoy[1], and Kemal Kurtulus[2]

[1] School of Business, Istanbul University, Istanbul, Turkey
{Dbulut,tugbasim}@istanbul.edu.tr
[2] Faculty of Economics & Administrative Sciences, Zirve University
profdrkurtulus@gmail.com

Abstract. Brands are working hard to build a brand equity which hope to lead the companies to have more loyal customers. Loyal customers are more cost efficient and have the intention to make multiple buying. The aim of this paper to track multiple buying behavior among customers with high brand loyalty. In order to see the relations between products chosen and preferences, data mining technique was used. Associations between products and future buying intentions were examined. High degrees of associations between products are presented. The future intentions were parallel to loyalty and satisfaction levels.

Keywords: Marketing, Data Mining, Associations Rules, Brand Loyalty.

1 Introduction

Brands aim to have loyal customers, not just because it is cheaper to satisfy the existing customer but also it requires less marketing activity to create awareness, high brand image and trust with these customers. Even though it is desired to have loyal customers, brand loyalty has different dimensions to be evaluated.

In the most general approach, brand loyalty could be grouped under four different dimensions: cognitive, affective, conative and behavioral [1]. Continues buying and or rebuying intentions are usually asked to customers to understand the loyalty level. Brand switching intentions are also examined to understand the behavioral dimension [2].

Applying association rules in order to see together, multiple (continues, increased and repetitive) buying behavior in a consumer group is a simple, yet a clever way to understand the consumer buying intentions. The same customers' buying behavior could be measured and product link could be mapped.

2 Brand Loyalty

Loyalty could be defined as the attachment to a certain thing, person or idea [3]. Brand loyalty on the other hand generally defines as consumers' willingness to continue their relationship with the brand [4].

P. Perner (Ed.): ICDM 2013, LNAI 7987, pp. 193–204, 2013.

Within the business perspective brand loyalty is the long-term, profitable process which reduces the corporate marketing cost [5]. Brands would like to continue their relationship with their customers and lower their costs of creating awareness and positive image. To be able to create this long relationship with the customers, it is crucial to build brand equity for the customers and give them reason to continue their relationship with the brand [6, 7].

Aspects of the brand loyalty may not be very easy to measure. Loyalty to a brand may not always have a cognitive level or a rational reason for the customer. Some brands offer an emotional level of attachment to their customers [1, 8].

Even in most of the practitioners and academicians would not measure all of them, four different levels of brand loyalty could be a base to define the approaches. Cognitive stage is usually link to positive experience with the brand. Affective level comes with brand/product satisfaction after accumulation of positive experiences. Conative stage is more personal and can be defined as the attitude development stage with the input data. Action stage is the reaction stage with the attitude. If the consumer combines positive attitudes, it may lead to multiple buying behavior like; rebuying, increased amount of buying, repetitive buying and buying and trying the new products of the brand [1, 9, 10].

Action level is usually not very easy to measure. Most of the times observation or database research is necessary to really understand the buying behavior. In this process researchers usually try to measure “buying intentions”. Rebuying [9, 11,12], increased buying or amount of money spend [11], repetitive purchase [10], or searching for the brand/product when not available [9, 12], brand switching intentions [2] or buying a new product of the brand [2, 12] are usually ask as intention to be able to understand the attitude.

Therefore multiple/repetitive buying behavior and/or intention are usually used to measure brand loyalty. When a data set of consumers available to measure the repetitive and multiple buying behavior, it may give the company an idea about their loyal customer profile and their needs. For this purpose it is possible to use many different analyses as well as relations rule in data mining.

3 Data Mining

Across a wide variety of fields, data are being collected and accumulated at a dramatic pace. There is an urgent need for a new generation of computational techniques and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of data. These techniques and tools are the subject of the emerging field of knowledge discovery in databases.(KDD).

KDD refers to the overall process of discovering useful knowledge from data while data mining refers to a particular step in this process. Data mining is the application of specific algorithms for extracting patterns from data. [13]

Data mining is the process of searching and analyzing data in order to find implicit, but potentially useful, information. It involves selecting, exploring and modeling large amounts of data to uncover previously unknown patterns, and ultimately comprehensible information, from large databases.

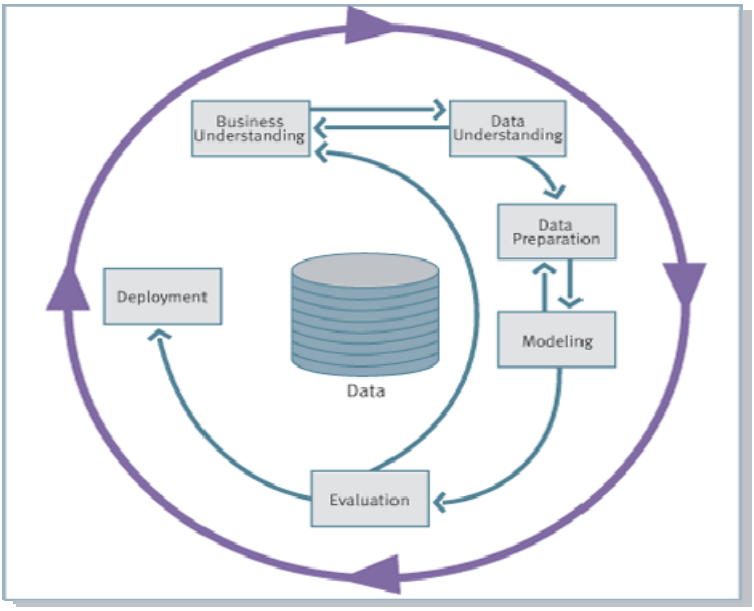
Data mining uses a broad family of computational methods that include statistical analysis, decision trees, neural networks, rule induction and refinement, and graphic visualization. Although, data mining tools have been available for a long time, the advances in computer hardware and software, particularly exploratory tools like data visualization and neural networks, have made data mining more attractive and practical[14].

4 The Data Mining Process

The Cross Industry Standard Process for Data Mining (CRISP-DM) project has developed an industry- and tool-neutral Data Mining process model. This project defined and validated a data mining process that is applicable in diverse industry sectors. This will make large data mining projects faster, cheaper, more reliable and more manageable. The CRISP-DM process consists of six stages. (www.crisp-dm.org, 12.07.2006)

Data mining process focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a **problem definition**.

The **data understanding** phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.



The **data preparation** phase covers all activities to construct the final dataset from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modeling tools.

In **modeling** phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed.

At **evaluation** stage in the project you have built a model (or models) that appears to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives.

Deployment: Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it.

5 Association Rules Analysis

Data mining techniques are specific implementations of algorithms used in data mining operations. Association Rules are described briefly below [15] .

Progress in bar-code technology has made it possible for retail organizations to collect and store massive amounts of sales data, referred to as the basket data. A record in such data typically consists of the transaction date and the items bought in the transaction. Successful organizations view such databases as important pieces of the marketing infrastructure [16].

Mining of association rules is a fundamental data mining task. Its objective is to find all co-occurrence relationships, called associations, among data sets. The classic application of association rule mining is the market basket data analysis [17] .

Market Basket Analysis process analyzes customer buying habits by finding associations between the different items that customers place in their shopping baskets. The discovery of such associations can help retailers develop marketing strategies by gaining insight into which items are frequently purchased together by customers [18].

Such valuable information can be used to support a variety of business-related applications such as marketing promotions, inventory management, and customer relationship management [19].

Market basket data consists of a set of binary attributes called items. Each data instance represents a customer transaction, and each item of that transaction can take on the value “true or false”, indicating whether or not the corresponding customer bought that item in that transaction.

An association rule is an implication expression of the form $X \longrightarrow Y$, where X and Y are disjoint item sets. Each association rule is usually evaluated by a support and a confidence measure. They represent the strength of the rule [20].

The support of an association rule is the ratio of the number of transactions (instances) having the value true for all items in both the set X and the set Y divided by the total number transactions.

Support (X ⟶ Y)=P(X U Y) / N

The confidence of an association rule is the ratio of the number of transactions having the value true for all items in both the set X and the set Y divided by the number of transactions having the value true for all items in the set X.

Confidence (X ⟶ Y)= P(X U Y)/ P(X)

The association rule discovery task consists of extracting from the data being mined all rules with support and confidence greater than or equal to user-specified thresholds called minimum support.

6 Application

Data Set

The research was conducted to an electronic device company's customers, who declare that they have at least one product from this brand. These consumers were reached through internet (company's website, consumer data base, internet blogs and chat rooms related with the technology and the brand) and total number of 1359 consumer's had participated to the online survey. After eliminating the missing data, non-owners of the brand and under 18 customers we have 1178 participants.

The electronic/technology device company has different group of products which most of them are easily related with the umbrella brand name. For the ethical reasons the name of the products or the company will not be declared. The product group names and numbering will be used (e.g. tablet computer1,2, or smartphone 1,2,3,4).

Table (244 fields, 1.178 records)

File Edit Generate

	Q1_1	Q1_2	Q1_3	Q1_4	Q1_5	Q1_6	Q1_7	Q1_8	Q1_9	Q1_...	Q1_...	Q1_...	Q1_...	Q1_...	Q1_...	Q1_...	Q1_...	Q1_...	Q1_19	Q1_20	Q1
1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	
2	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	
5	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
6	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	
7	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	
8	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0	0	0	
9	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	
10	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	

The demographic profile of the participants, the owned products, the first owned product, related software they are using and their brand loyalty, brand satisfaction levels and their future purchase intentions were asked. Also the reasons are asked for choosing this brand.

The majority of the participants are male (78.8%) and the average age is 29.6 (st.dev. 9.961). 49.6% of the participants are married and the education mode is University graduates with 56.8%. The participants are asked about their family sizes (including themselves) and the majority of the group declared that they have the

family sizes of 3-4 (30.9%-31.8%). The mode group of income is 1000-1500€ (approximately) with 13.2%. Even though the occupation groups vary, designer/artist attracts attention with the highest level with 22%.

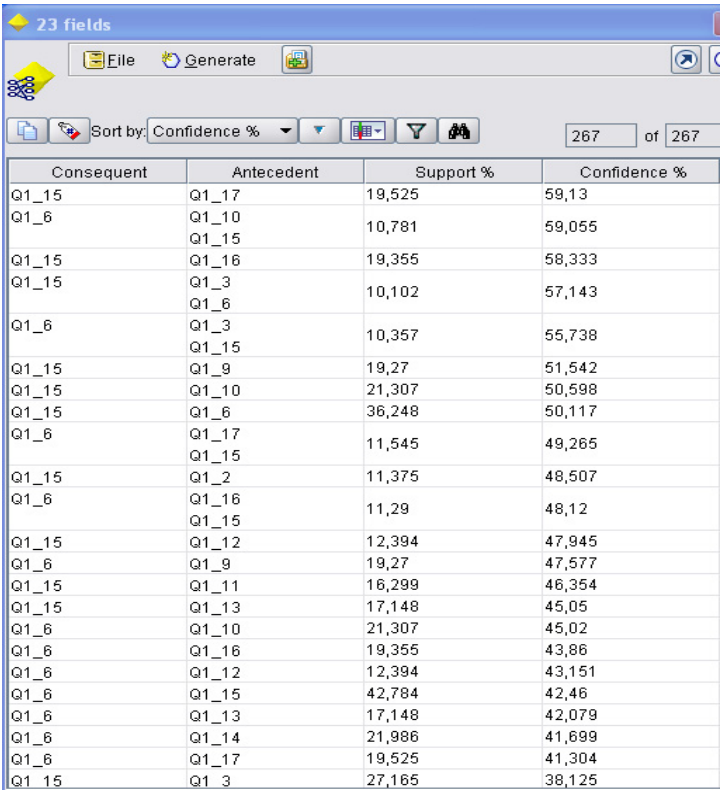
Brand satisfaction and loyalty levels were measured and compared with the average. Both brand loyalty and brand satisfaction scores are significantly high compared to the scale average (sig.=0.000/ 90.1 0ver 95 total score.). For the future buying intentions participants are asked to name products but also given the chance to choose the alternative to say that "I am not going to buy any other product from this company" and "I am not planning to buy any other product till the company develops a new one". Only 7 participants answered that they are not planning to buy any other product from this company, while 145 declares that they are waiting for a new product to be developed by the company.

To be able to see the multiple buying behavior and continues trends of buying companies products association rules were applied to products own, first product owned, future buying plans and software used.

Association Rules Analysis

IBM Modeler program and Apriori algorithm was used to analyze the data. Rule sets can be seen in Table .

Table – Rule Sets



Consequent	Antecedent	Support %	Confidence %
Q1_15	Q1_17	19,525	59,13
Q1_6	Q1_10 Q1_15	10,781	59,055
Q1_15	Q1_16	19,355	58,333
Q1_15	Q1_3 Q1_6	10,102	57,143
Q1_6	Q1_3 Q1_15	10,357	55,738
Q1_15	Q1_9	19,27	51,542
Q1_15	Q1_10	21,307	50,598
Q1_15	Q1_6	36,248	50,117
Q1_6	Q1_17 Q1_15	11,545	49,265
Q1_15	Q1_2	11,375	48,507
Q1_6	Q1_16 Q1_15	11,29	48,12
Q1_15	Q1_12	12,394	47,945
Q1_6	Q1_9	19,27	47,577
Q1_15	Q1_11	16,299	46,354
Q1_15	Q1_13	17,148	45,05
Q1_6	Q1_10	21,307	45,02
Q1_6	Q1_16	19,355	43,86
Q1_6	Q1_12	12,394	43,151
Q1_6	Q1_15	42,784	42,46
Q1_6	Q1_13	17,148	42,079
Q1_6	Q1_14	21,986	41,699
Q1_6	Q1_17	19,525	41,304
Q1_15	Q1_3	27,165	38,125

- First rule set: 59.13% of the customers who buy Tablet Computer 2, also buy Smartphone 4. Support ratio is 19.525%.
- 59.055% of the customers who buy Mp3 player/Data storage 2 and Smartphone 4, also buy Laptop 2. Support ratio is 10.781%.
- 58.333% of the customers who buy Tablet Computer1, also buy Smartphone 4. Support ratio is 19.355%.
- 57.143% of the customers who buy Computer3 and Laptop 2, also buy Smartphone 4. Support ratio is 10.102%.
- 55.738% of the customers who buy Computer3 and Smartphone 4, also buy Laptop 2. Support ratio is 10.357%.
- 51.542% of the customers who buy Mp3 player/Data storage 1, also buy Smartphone 4. Support ratio is 19.27%.
- 50.598% of the customers who buy Mp3 player/Data storage 2, also buy Smartphone 4. Support ratio is 21.307%.
- 50.117% of the customers who buy Laptop 2, also buy Smartphone 4. Support ratio is 36.248%.

There is a very high level of preference Smart Phone 4 with other product groups like; tablet computer, laptop and Mp3 player Data storage.

Which Software do you use with the products?

13 fields

File Generate

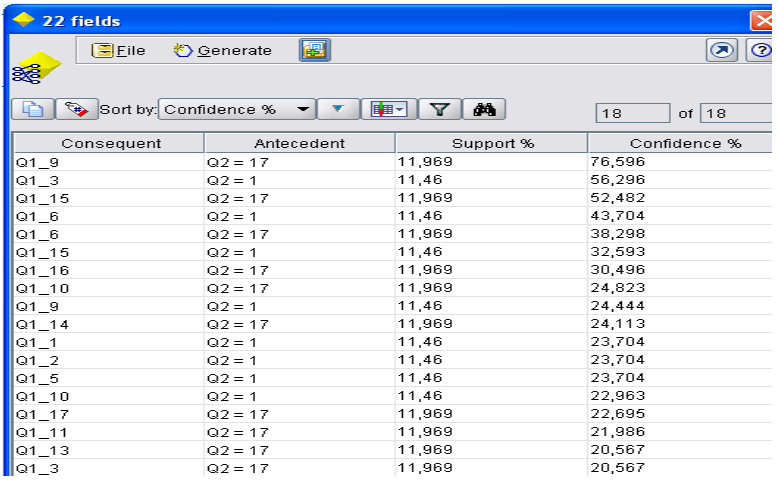
Sort by: Confidence % 24 of 24

Consequent	Antecedent	Support %	Confidence %
Q6_1	Q1_6	36,248	96,487
Q6_1	Q1_3	27,165	95,625
Q6_2	Q1_3	27,165	95,625
Q6_2	Q1_6	36,248	95,316
Q6_2	Q1_15	42,784	94,841
Q6_4	Q1_6	36,248	83,138
Q6_4	Q1_15	42,784	82,738
Q6_4	Q1_3	27,165	81,562
Q6_1	Q1_15	42,784	73,016
Q6_12	Q1_15	42,784	58,73
Q1_15	Q1_6	36,248	50,117
Q6_6	Q1_6	36,248	48,478
Q6_5	Q1_6	36,248	45,199
Q6_12	Q1_6	36,248	43,794
Q6_6	Q1_3	27,165	43,125
Q6_3	Q1_15	42,784	42,857
Q1_6	Q1_15	42,784	42,46
Q6_5	Q1_3	27,165	40,938

- First rule set: 96.487% of the customers who buy Laptop 2, also buy Basic operating system. Support ratio is 36.248%.
- 95.625% of the customers who buy Computer 3, also buy Basic operating system. Support ratio is 27.165%.
- 95.625% of the customers who buy Computer 3, also buy Music pro. Support ratio is 27.165%.
- 95.316% of the customers who buy Laptop 2, also buy Music pro. Support ratio is 36.248%.
- 94.841% of the customers who buy Smartphone 4, also buy internet. Support ratio is 42.784%.
- 94.841% of the customers who buy Smartphone 4. also buy internet. Support ratio is 42.784%.
- 83.138% of the customers who buy Laptop 2, also buy Music pro. Support ratio is 36.248%.
- 82.738% of the customers who buy Smartphone 4, also buy Music pro. Support ratio is 36.248%.
- 81.562% of the customers who buy Laptop 3, also buy internet. Support ratio is 27.165%.
- 73.016% of the customers who buy Smartphone 4, also buy basic operating system. Support ratio is 42.784%.
- 58.73% of the customers who buy Smartphone 4, also buy smartphone applications. Support ratio is 42.784%.

The basic operation system is free and specific for the brand with the technology required to use it. For more professional versions consumers need to buy the program. Internet and music program are also free and downloadable. Some smart phone applications are free and some could be purchased online. Most of the professional programs which consumers need to pay for are not listed in high association.

What is the first Product you have?

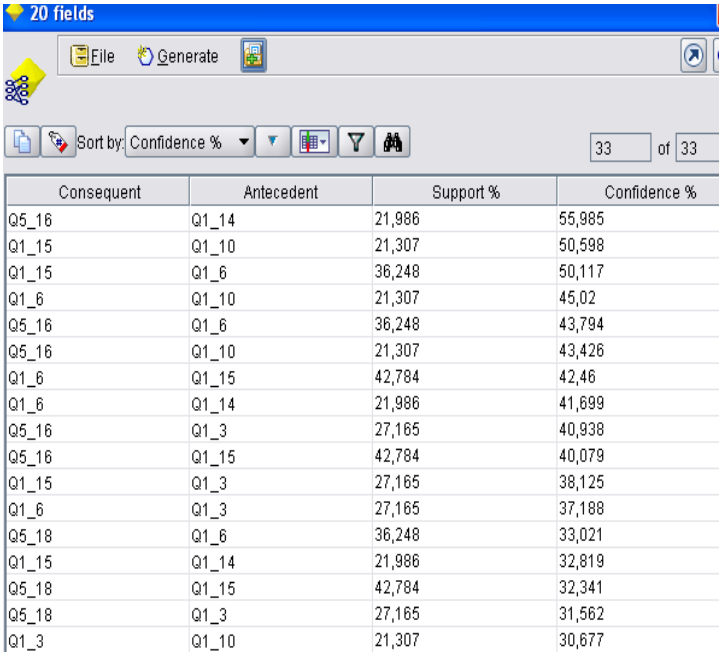


Consequent	Antecedent	Support %	Confidence %
Q1_9	Q2 = 17	11,969	76,596
Q1_3	Q2 = 1	11,46	56,296
Q1_15	Q2 = 17	11,969	52,482
Q1_6	Q2 = 1	11,46	43,704
Q1_6	Q2 = 17	11,969	38,298
Q1_15	Q2 = 1	11,46	32,593
Q1_16	Q2 = 17	11,969	30,496
Q1_10	Q2 = 17	11,969	24,823
Q1_9	Q2 = 1	11,46	24,444
Q1_14	Q2 = 17	11,969	24,113
Q1_1	Q2 = 1	11,46	23,704
Q1_2	Q2 = 1	11,46	23,704
Q1_5	Q2 = 1	11,46	23,704
Q1_10	Q2 = 1	11,46	22,963
Q1_17	Q2 = 17	11,969	22,695
Q1_11	Q2 = 17	11,969	21,986
Q1_13	Q2 = 17	11,969	20,567
Q1_3	Q2 = 17	11,969	20,567

- 52.482% of the customers who have Mp3 player 1, also have Smartphone 4. Support ratio is 11.969%.
- 38.298% of the customers who have Mp3 player 1first, also have Laptop 2. Support ratio is 11.969%.
- 30.496% of the customers who have Mp3 player 1first, also have Tablet Computer 1. Support ratio is 11.969%.

The first products are usually Mp3 players, which is very understandable considering the product development stage of the company. Computer, mp3 player and laptops are followed by smart phones and tablet computers. There has been advances in older classes but this association relations fits with the continues and loyal consumer profile, who has been a fan of the company products over decades.

What is the first product you intend to buy?

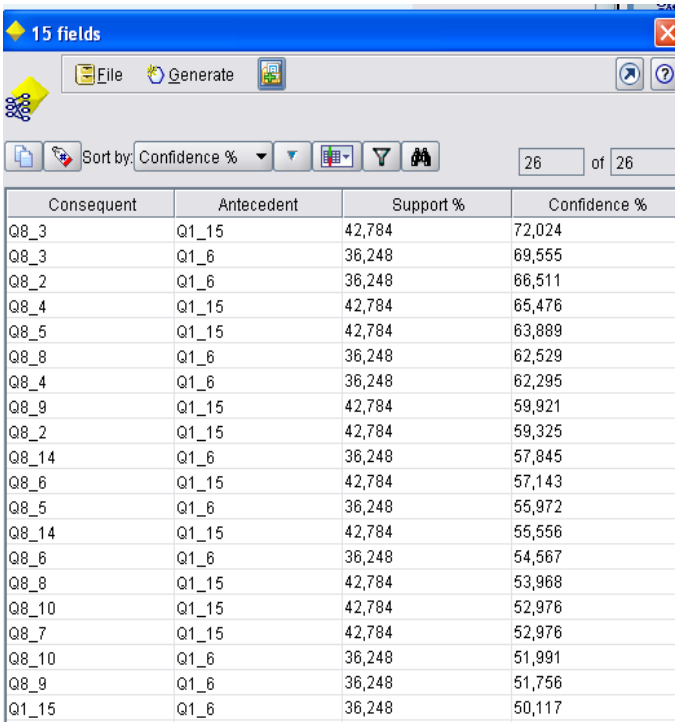


Consequent	Antecedent	Support %	Confidence %
Q5_16	Q1_14	21,986	55,985
Q1_15	Q1_10	21,307	50,598
Q1_15	Q1_6	36,248	50,117
Q1_6	Q1_10	21,307	45,02
Q5_16	Q1_6	36,248	43,794
Q5_16	Q1_10	21,307	43,426
Q1_6	Q1_15	42,784	42,46
Q1_6	Q1_14	21,986	41,699
Q5_16	Q1_3	27,165	40,938
Q5_16	Q1_15	42,784	40,079
Q1_15	Q1_3	27,165	38,125
Q1_6	Q1_3	27,165	37,188
Q5_18	Q1_6	36,248	33,021
Q1_15	Q1_14	21,986	32,819
Q5_18	Q1_15	42,784	32,341
Q5_18	Q1_3	27,165	31,562
Q1_3	Q1_10	21,307	30,677

- 55.985% of the customers -who have Smartphone 3- are intend to buy Smartphone 5.
- 50.985% of the customers -who have Mp3 player- are intend to buy Smartphone 4.
- 50.171% of the customers -who have Laptop 2- are intend to buy Smartphone 4.

The Smartphone users, mp3 player owners and laptop owners are all interested in the Smartphone technology. It would be a good assumption that the Smart Phone is one of the most popular products of the company.

Why do you choose these products?



Consequent	Antecedent	Support %	Confidence %
Q8_3	Q1_15	42,784	72,024
Q8_3	Q1_6	36,248	69,555
Q8_2	Q1_6	36,248	66,511
Q8_4	Q1_15	42,784	65,476
Q8_5	Q1_15	42,784	63,889
Q8_8	Q1_6	36,248	62,529
Q8_4	Q1_6	36,248	62,295
Q8_9	Q1_15	42,784	59,921
Q8_2	Q1_15	42,784	59,325
Q8_14	Q1_6	36,248	57,845
Q8_6	Q1_15	42,784	57,143
Q8_5	Q1_6	36,248	55,972
Q8_14	Q1_15	42,784	55,556
Q8_6	Q1_6	36,248	54,567
Q8_8	Q1_15	42,784	53,968
Q8_10	Q1_15	42,784	52,976
Q8_7	Q1_15	42,784	52,976
Q8_10	Q1_6	36,248	51,991
Q8_9	Q1_6	36,248	51,756
Q1_15	Q1_6	36,248	50,117

- 72.024% of the customers buy Smartphone 4 because of its design.
- 69.555% of the customers buy Laptop 2 because of its design.
- 66.511% of the customers buy Laptop 2 because it is appropriate for their business.
- 65.476% of the customers buy Smartphone 4 because it is easy to use its software.

When we aim to understand why the consumers choose their products from this company, for laptop and smartphone “design” appears to be the most important reason. “Appropriate for the business” is one of the other most important reasons for buying Laptop2 with 66.511% association. Smartphones are associated with easy to use software.

7 Conclusions

Brand loyalty is always desirable for a more profitable business. To be able to gain brand loyalty marketers are working hard to create an attachment. For predicting the future of the brand and plan the marketing efforts it is also important to measure the brand loyalty.

One of the most preferred level of brand loyalty measurement is multiple (continues, increased and repetitive) buying behavior. Even it might be difficult to measure the action, researchers still can use the intention measurements. When the researcher could reach to an actual sales data or customer product lists, it is more possible to measure the action.

With this research, it is aimed to see the consumer's multiple buying behavior in a very high brand satisfaction and brand loyalty group. (Total score of 90.1 over 95).

Smartphone 4 is one of the most popular products. The costumer who own it usually own Tablet computer 1 or 2 and an Mp3 player. Another group combination is Computer 3 owners also have laptop2 and smartphone 4. The association percentages are considerably high.

For the consumers demanding for high technological products, it is very important to have complementary products such as the software. It is asked which software the customers are using, the main operating system software and internet program and music program in considerably high ratios. The important part is all of these programs are free, either coming with the device or open source to be downloaded. Some other professional product for movie making, or music recording are not really high. This brand is usually associated with designers and artist which is also a big proportion of the sample, but these programs are not in high demand.

The first product of the large portion of the sample is one of the models of the Mp3 player/data storage devices which has been long on the market and developed in time. The following or combined products are usually the more recent products like tablet computers, smartphone or laptop 2. This also shows old and loyal customers starting with the original computer products are no longer the majority and new generation product users are a big portion of the group. Mp3 player/data storage device users usually have the tendency to buy and/or try different products of the brand.

Design of the product is one of the most important differentiation point of the brand. The reason for choosing this brand is highly related with the design. For product groups such as technologic devices, phones etc. usually the most important reasons tend to be more rational, related with performance, quality or quality/ price ratio. But design appears to be highly important for Smartphone 4 and Laptop 2 owners. "Appropriate for the business" is one of the other most important reasons for owning laptop 2. Brand could focus on design and art oriented target group to increase the sales of this product. Smartphones are associated with easy to use software. This might be one of the advantages that brand could promote on.

The study was conducted only with the owners of the brand products. So as a starting point they all have at least one product. This is required to evaluate the satisfaction and the brand loyalty levels, as it wouldn't be healthy to evaluate a non-owner's brand loyalty. For the future studies, same or similar brand could be evaluated among non-users also. Sales data base could lead to a time line of the purchase and could indicate the starting point more efficiently. Different cultures and markets are recommended for future studies.

References

1. Quester, P., Lim, A.L.: Product Involvement/Brand Loyalty: Is There a Link? Journal of Product & Brand Management 12(1), 22–38 (2003)
2. Bennet, R., Rundle-Thiele, S.: A Comparison of Attitudinal Loyalty Measurement Approaches. Brand Management 9(3), 193–209 (2002)
3. Kotler, P., Armstrong, G.: Principles of Marketing, 13th edn. Pearson (2009)
4. Kapferer, J.N.: The New Strategies Brand Management. Kogan Page Ltd. (2008)
5. Griffin, J.: The Internet's Expending Role in Building Customer Loyalty. Direct Marketing 59(7), 50–53 (1996)
6. Keller, K.L.: Strategic Brand Management: Building, Measuring and Managing Brand Equity. Pearson International, Prentice Hall (2008)
7. Reichheld, F.F.: Loyalty-Based Management. Harvard Business Review (1993)
8. Vincent, L.: Legendary Brands: Unleashing the Power of Storytelling to Create a Winning Marketing Strategy. Dearbon Trade Publishing (Kaplan Professional) (2002)
9. Oliver, R.L.: Wherence Consumer Loyalty. Journal of Marketing 63, 33–44 (1999)
10. Jacoby, J., Kyner, D.B.: Brand Loyalty vs. Repeat Purchasing Behavior. Journal of Marketing Research 10, 1–9 (1973)
11. Bagozzi, R.P., Dholakia, U.M.: Antecedents and Purchase Consequences of Customer Participation in Small Group Brand Communities. International Journal of Research in Marketing 23, 45–61 (2006)
12. Algesheimer, R., Dholakia, U.M., Herrmann, A.: The Social Influence of Brand Community: Evidence from European Car Clubs. Journal of Marketing 69, 19–34 (2005)
13. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: Knowledge Discovery and Data Mining: Towards a Unifying Framework. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-1996), Portland, Oregon (1996)
14. Shaw, M., Subramaniam, C., Tan, G.W., Welge, M.E.: Knowledge Management and Data Mining for Marketing. Decision Support Systems 31(1), 127–137 (2001)
15. Moss, L.T.: Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support Applications. Addison-Wesley (2003)
16. Berry, M.J.A., Linoff, G.S.: Mastering Data Mining, p. 7. Wiley (2000)
17. Breault, J.L., Goodall, C.R., Fos, P.J.: Data Mining a Diabetic Data Warehouse. Artificial Intelligence in Medicine 26, 37–54 (2002)
18. Bounsaythip, C., Rinta-Runsala, E.: Overview of Data Mining for Customer Behaviour Modeling, VTT Information Technology Research Report TTEI, p.18 (2001)
19. Frand, J.: Data Mining: What is Data Mining? (July 12, 2006), `http://www.anderson.ucla.edu`
20. Fule, P.: Exploratory Medical Knowledge Discovery: Experiences and Issues. ACM SIGKDD Explorations Newsletter 5(1), 94–99 (2003)

Matching Semi-structured Documents Using Similarity of Regions through Fuzzy Rule-Based System

Alireza Ensan and Yevgen Biletskiy

University of New Brunswick
Fredericton, New Brunswick, Canada
{alireza.ensan,biletski}@unb.ca

Abstract. The present work briefly describes a novel approach for categorizing semi-structure documents by using fuzzy rule-based system. We propose fuzzy logic representation for semi-structured documents and then by proposing new metric, categorize documents into different classes. The idea behind of our approach is to divide web pages into different semantic sections and by using fuzzy logic system extract features and weight harvested terms to represent semi-structure documents. A set of metrics are also used to measure similarity between documents based on the weight of each region in the text. A clustering algorithm is also explained that categorized documents into several categories. This idea is inspired as a subfield of the area of *Matchmaking* that tries to match document creators and users in order to find the best similarities between them and connect them for further collaborations.

Keywords: Data mining, Information extraction, Document categorization, Fuzzy system.

1 Introduction

Tremendous amounts of information are freely accessible to everyone through digital networks especially the Internet. The access and finding interests among this massive amount of information is not possible unless they are suitably categorized and organized. One of the most important tasks for text document organization is feature extraction from text. It refers to the task of extracting some features, in the form of single words or phrases, which can represent semantics of documents. The reason for the feature extraction is to reduce the dimension of documents, which are processed, and also remove noisy data that can affect the result of text processing. In order to classify web pages in HTML format, we can utilize the characteristics of web pages which are represented in HTML language. HTML markup can provide the ability of using some important parts of documents with the purpose of feature selection. We take advantage of fuzzy logic to take in to account human expert' decision that guides us to weight and select terms form HTML pages. We also discuss a K-Nearest-Neighbor (KNN) categorization method that uses a new metric for evaluating similarity based on particular weights achieved from the procedure of feature selection.

P. Perner (Ed.): ICDM 2013, LNAI 7987, pp. 205–217, 2013.

In our work, we take benefits from the fuzzy logic representation by means of heuristic combinations of criteria [1] and categorization for web page representation with four contributions:

1. The method used in [1] assumes that three different sections (introduction, body, and conclusion) of each web page come out based on a stochastic assumption. For example, introduction of each text usually appears in the first 10% lines of document. This assumption is not accurate and cannot be the basis of the position-based text analyzing because of different structure of HTML documents with other text formats. In order to overcome this shortcome, we propose a novel approach to divide web pages in different semantic sections. Regarding the amount of relevance of each term to the each section, its importance for the document is evaluated.
2. The second contribution is the defining new fuzzy rules and a sub fuzzy system with its corresponding rule, which is able to compute the weight assigned to each selected term. This new knowledge base system is designed to utilize semantic sections in the document.
3. Contrasting the proposed approach used in [1] that notices to term frequency in one document, we bring the idea of term frequency in a set of document to filter the output weights of harvested terms. We also rank outputs based on a ranking algorithm
4. We propose a clustering algorithm that uses weighted terms in different sections of each web document to compute semantic distance among documents.

In the rest of this paper, in section 2 we review state of art in the area of term weighting and feature selection and similarity measurement. Then in section 3, we explain our proposed approach. Finally, we conclude this paper with the future word and conclusion.

2 Background

A substantial amount of research has been done in the area of document organization [2]. In every text processing task, a method for feature extraction or term selecting is a challenging problem. There are many attempts which have been done to discuss different approaches of feature selection and term weighting in text [3]. The most popular approach for the feature extraction is Vector Space Model (VSM) [4] that aims to represent documents based on the *bag of words* as a vector. There are many extensions for VSM such as Latent Semantic Analysis or Probabilistic Latent Semantic Analysis that unlike other VSM methods generate a set of concepts related to both documents and terms with. The main idea of these methods is that closer similar terms are more possible to occur in the same document [5]. Several researchers have categorized the area of feature extraction from different point of views. It can be divided in two major categories: a) supervised and b) unsupervised [6]. Both supervised and unsupervised feature extraction methods make use of each term as a dimension of a space model defined to represent a particular document in a particular corpus. For example, TF-IDF (Term Frequency- Inverse Document Frequency) is one the most common unsupervised ways to weight terms in documents

[6]. There are also some supervised approaches that utilize trained dataset to reduce dimensionality of model's space such as chi-squared, information gain, mutual information, and odds ratio [6]. With the intention of web pages categorization, hierarchal structure of HTML documents helps us to rank sections in documents based on their importance. HTML markup can provide the ability of using some important parts of documents with the purpose of feature selection [7]. A human rule-based system can help to consider the importance of each section in a document in the procedure of feature selection and term weighting. Fuzzy rule-based system can employ human expert's knowledge to build a rule-based system that determines semantically important terms based on their positions in the text. Extended Fuzzy Combination of Criteria (EFCC) [1] is a fuzzy rule-based system that determines the importance of each harvested term based on its position in the web page. Some fuzzy rules are also employed to perform inference in the knowledge base. We aim to sophisticate the idea of fuzzy logic representation with semantic sections extraction. Moreover, we define a new fuzzy rule-based system to support our knowledge base. After performing the idea of document representation, we discuss document similarity measuring and categorizing issue. Semantic similarity approaches mainly focus on the semantic distances between different objects in order to find proper matches between those objects. Many attempts have been performed to measure the similarity between documents based on Vector Space Model (VSM) such as cosine similarity, and Dice coefficient [8]. Our proposed method uses the idea of vector of terms; In addition, our proposal employs an extended version of a similarity measuring method basing on cosine similarity metric. It employs a set of weights for each semantic section in the text which achieved from the output of fuzzy rule-based system.

3 Fuzzy Rule-Base System for Document Representation

In order to categorize semi-structure document (aka web page), in the first place we explain fuzzy rule-base system for document representation, and then in the next section, we discuss how this representation assists us to categorize documents.

We first divide the document to different semantic section based on the decision of human expert. Then, an information extraction method is performed to put information in certain semantic sections. After that, a fuzzy rule system models the importance of positions in web pages (e.g., title, keywords, tags, etc...). Finally, fuzzy system defines importance level for each semantic sections based on the preference of human expert. For instance, a domain expert recognizes that for a job finding website, the job experience is the most descriptive section of each job application. The algorithm is based on fuzzy logic representation of document comprises of subsequent steps as follows.

Step 1: Semantic sections extraction. We perform the inspiration of Information extraction from HTML document and its re-formatting into an XML document which is semantically more organized. We exploit the idea of position-based information extraction from CODE method discussed in [9]. A semantic annotation process parses a semi-structure document to create a Data Object Model (DOM) tree [9]. This tree would be converted to structured documents (XML) with certain sections in next steps. The human expert generates a template that assists the system to extract information, which is an XML template file that contains tags without any data.

Step 2: Fuzzy system design. In this step we design a fuzzy system for implementing inference on fuzzy rules. Each fuzzy system includes fuzzy variables, inference rules, output fuzzy set, and defuzzification process [10]. Below we explain each part of our fuzzy system. In the first place we define linguistic variables as inputs for the fuzzy rule system as follows.

Keywords: In most of web documents, some keywords or tags are provided to represent the meaning of documents; these keywords are also used for search engines. Certainly, these words are the most descriptive words in each web page. This section usually determines some HTML tags such as <Keywords> …</Keywords>, <tag>…</tag>, and etc. Figure 1 illustrates the membership function for this linguistic variable.

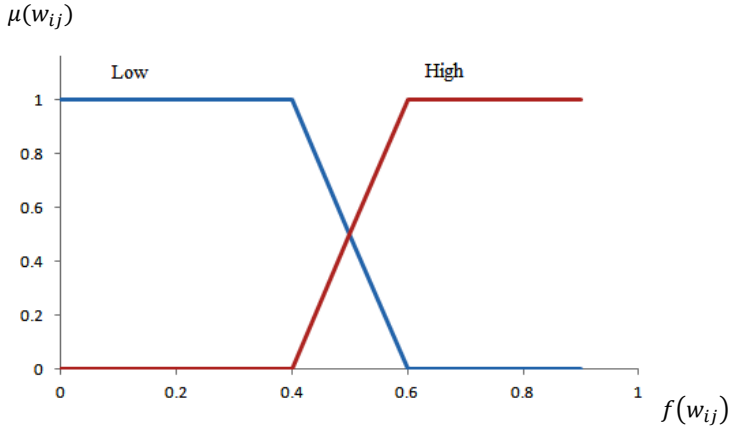


Fig. 1. Membership function for the linguistic variable Keyword

In Fig.1, $\mu(w_{ij})$ is the value of membership function for the word w_{ij} that is the *ith* word in the *jth* document. Moreover, $f(w_{ij})$ is normalized by the occurrence of a term in the section of keywords, and defined as:

$$f(w_{ij}) = \frac{Count_{Keyweord}\,(w_{ij})}{\max_{k}^{count\{w_{kj}\}}}, \tag{1}$$

where, $Count_{Keyweord}\,(w_{ij})$ is the number of occurrences of a word in the keywords section, and $\max_{k}^{count\{w_{kj}\}}$ is the maximum number of occurrence of a keyword in this section.

Title: Another important part of each document, which can be very descriptive, is title. The words that occur between title tags are important and descriptive. The membership function can be defined as in Fig. 2.

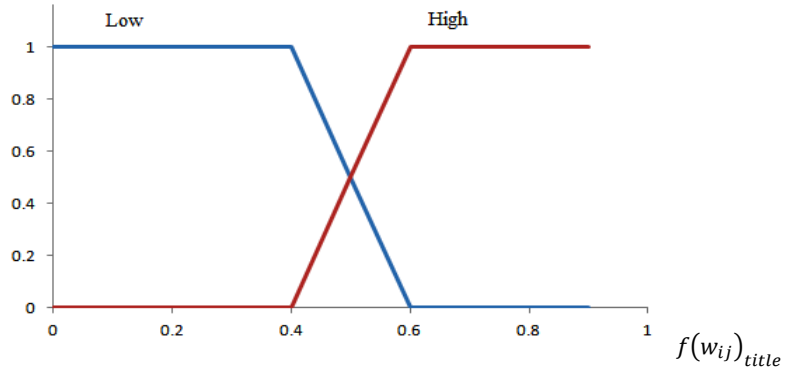


Fig. 2. Membership function for the linguistic variable Title

In this membership function:

$$f(w_{ij})_{title} = \frac{Count_{title}(w_{ij})}{\max_{k}^{count\{w_{kj}\}}}, \quad (2)$$

where $Count_{title}(w_{ij})$ denotes to the number of occurrences of a word in the title sections, and $\max_{k}^{count\{w_{kj}\}}$ is the maximum number of occurrencies of a word in this section.

Term Frequency: This linguistic variable defines low or high frequency of a word in whole of the text. Similar to the previous linguistic variables, we normalize the number of occurrences of a word in whole of text by the following formula (and as in Fig. 3):

$$f(w_{ij})_{document} = \frac{Count_{document}(w_{ij})}{\max_{k}^{count\{w_{kj}\}}}, \quad (3)$$

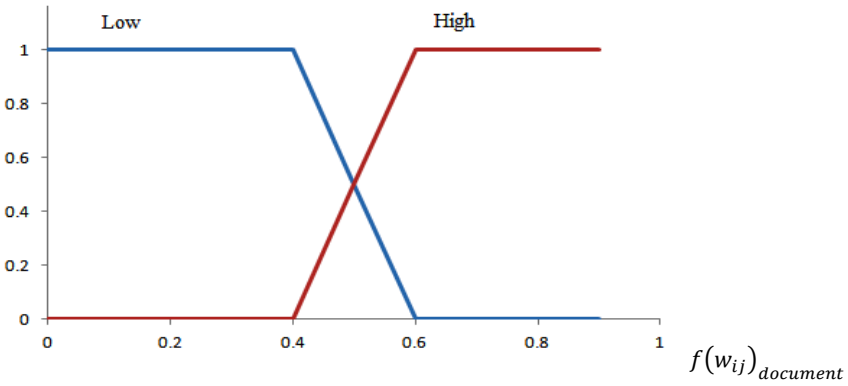


Fig. 3. Membership function for the linguistic variable Term frequency

Semantic Tags: We have explained that we extract information from an HTML document and put information between semantic tags. Each section is named a *semantic section* and we characterize another membership function that illustrates the dependency of each word in each semantic section (Fig. 4).

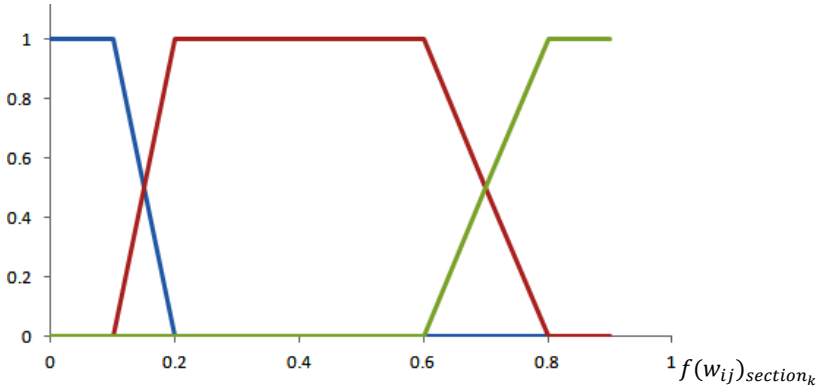


Fig. 4. Membership function for the linguistic variable Semantic tag

In each semantic section, we define a normalization factor similar which is to previous linguistic variables. Here:

$$f(w_{ij})_{section_k}=\frac{Count_{section_k}(w_{ij})}{max_m^{count\{w_{mj}\}}}, \quad (4)$$

where, $Count_{section_k}(w_{ij})$ is the occurrence number of a word in the k-th section of document j and $max_m^{count\{w_{mj}\}}$ is the maximum number of occurrence of a word in the semantic section k.

Step 3: Creation of Sub-Fuzzy System. In order to build a knowledge base with the capability of inference on predefined rules, we generate a sub fuzzy system that aggregates all semantic sections variables in one fuzzy variable. To achieve this goal, we define a variable which is named Semantic Section Preference. This variable can take the value of Preferred and Not preferred. These values can be assigned by human expert who recognizes that which semantic section is preferred or not preferred.

Then we define another fuzzy variable that represents the importance of each word belonging to semantic sections regarding its frequency and the human expert's preference. The membership function for this variable which is named Importance level is presented in Fig. 5. In this definition, $0 \leq \Gamma(w_{ij}) \leq 0$, is a coefficient that determine the importance of each word regarding its relatedness to a particular semantic section. This variable is considered as the antecedent of rules in the sub-fuzzy system. A set of IF-Then rules are defined for this sub system which is listed in table 1.

Table 1. IF-THEN rules for the sub-fuzzy system

IF-THEN RULES		
IF		THEN
Semantic Section Preference	Term dependency to Semantic tags	Importance Level
Preferred	High or Medium	High
Preferred	Low	Low
Not Preferred	High	High
Not Preferred	Low or Medium	Low

A defuzzification process is based on the Center of M ass (COM) algorithm [10] that returns the amount of importance level after inference process. This fuzzy variable is set as the input of the main fuzzy system.

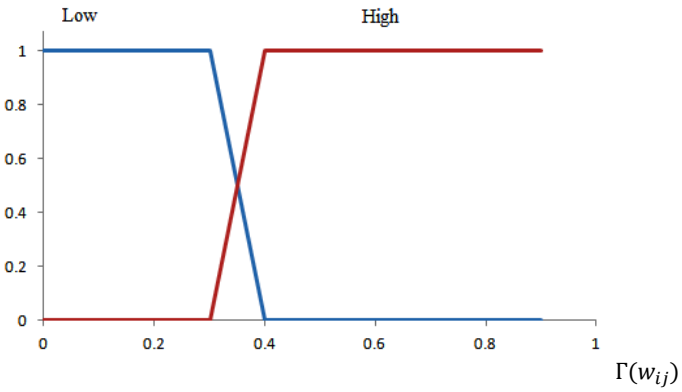


Fig. 5. Membership function for the linguistic variable Importance Level

Step 4: Building knowledge base for Fuzzy System. e define a fuzzy variable that is considered as the antecedent of IF-THEN rules for the main fuzzy rule system. This linguistic variable is called Relevance that illustrates the relevance of each word to documents. Relevance's membership function is shown in Fig. 6.

The amount of relevance of a word to the whole of a text is defined by the TF-IDF formula. This variable is determined by 5 different fuzzy values including: Not Relevant (NR), Low Relevant (LR), Medium Relevant (MR), High Relevant (HR), and Very Relevant (VR). Now we define a knowledgebase that determines importance of each word that belongs to each semantic section which is shown in Table 2. The output of the inference in this fuzzy rule system is used as input for the main fuzzy system. For the main fuzzy system, we define fuzzy variables that represent relevance of each word in the document. The idea behind defining these

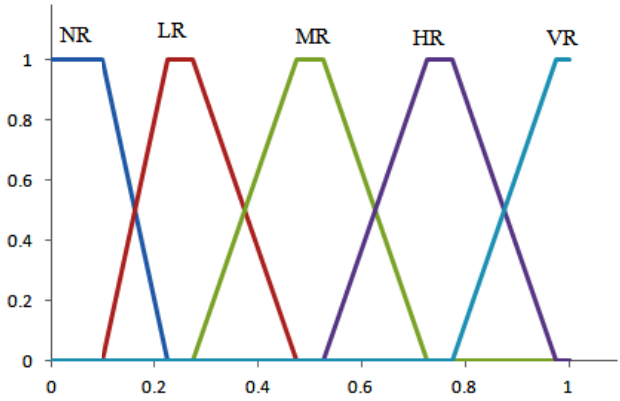


Fig. 6. Membership function for the linguistic variable Relevance

rules is, firstly, when a term occurs in the keyword section, it means that it is very relevant to the semantic of the document. Second, when one term appears in the title, based on it is important in the text; it can be considered as relevant or non-relevant term in the text. Finally, when all of these rules cannot determine the importance of a term, the frequency of the term in the document gets important for evaluating the *Relevance* of the term to the document.

Table 2. IF-THEN rules for the main fuzzy system

IF-THEN Rules				
IF				THEN
Title	Keyword	Text Frequency	Importance Level	Relevance
	High			VR
	Low			VR
High			High	VR
High			Low	MR
Low			High	MR
Low			Low	LR
		High	High	HR
		High	Low	MR
		Low	High	MR
		Low	Low	NR

Step 5: Defuzzification. For defuzzification process, we use Center of Mass (COM) algorithm that returns a crisp value which is a weight assigned to a particular word in the document. The outputs of fuzzy rules are aggregated in one fuzzy set by maximum aggregation as it is explained in [10].

Step 6: Term Weighting and Selection. Now we have one fuzzy output that by computing the center area under the final curve, we can obtain to the weight of a particular word. The output of the defuzzification process is $\Psi(w_{ij})$ that is the

relevance of a word in the document. In order to avoid consideration of terms that occur many times in a document and appear in many documents of a document corpus, such as some verbs, articles, etc., we bring the idea of corpus-based term weighting TF-IDF to filter the result as:

$$\text{weight}\left(w_{ij}\right) = \Psi\left(w_{ij}\right) \times \text{tf} - \text{idrf}(w_{ij}) \tag{5}$$

Here, we want to rank our weighted terms in documents. These terms are weighted based on their importance and semantic position in the document and also decision of domain expert to prioritize some semantic sections rather others. We use an algorithm as follows:

1. First order terms with regard to their weights;
2. Select terms with higher rank in the first category of relevance to a document(e.g., very relevant);
3. If the desired number (α) of terms are not achieved, go to the next category of relevance to seek more terms;
4. Put all selected terms in a vector V_D with the length α.

This algorithm selects most related and also most important terms in the document in order to further processing tasks.

4 Document Categorization

In this section, we first proposed semantic similarity measurement metric which uses weighted regions in the text. Then we use this metric to categorize documents in a collection. Our new metric takes in to account the result of defuzzification from the previous section to compute the similarity. It extends the idea of cosine similarity [8] and considers fuzzy variables to compute similarity between documents. We represent each document with a set of weighted words in a vector V_D which are attained from the section 3.12.4. In the first place, we measure the similarity between two sections of documents:

$$Sim_{section_i}(D_1, D_2) = W_i \times \frac{2 \times \sum_{j=1}^{b} \min\left(count_{ij1}, count_{ij2}\right)}{\sum_{j=1}^{b}(count_{ij1} + count_{ij2})}$$

where, b is the number of similar words between documents D_1 and D_2, that are identified in the *ith* section (including *semantic sections, title*, and *keywords*) of each document. In addition, $count_{ij1}$and $count_{ij2}$ denote to the frequency of each word in the section i of documentsD_1 and D_2. Moreover, W_i is the weight of *ith* section of each document which is computed as follows:

$$W_i = \sum_{k=1}^{\alpha} Weight(w_{ik}), \tag{7}$$

where $w_{ik} \in V_D$ and $w_{ik} \in Section\ i$.

In this formula, α is the length of Vector V_D which the number of weighted words for the document representation. Weight for a section i is aggregation of weights of all words that belong to V_D and belong to the section i in the document.

The similarity between whole of two documents is computed as the aggregation of similarity between different sections as subsequent formula:

$$Sim\ (D_1, D_2) = \sum_{i=1}^{\beta} Sim_{section_i}(D_1, D_2), \tag{8}$$

where β is the number of existing sections in each document that denote to the number of *semantic sections*, *title* section, and *keywords* section in each document.

In order to categorize documents in different classes, we first use KNN algorithm [11] to classify documents in different classes using the semantic metric proposed in the formula 8. Then we search among all V_D vectors in one category; such term that has the most weight is selected as the name of the category. The algorithm follows these steps:

1. Implement KNN algorithm to classify documents. The formula 8 is used to evaluate semantic distance between documents. When the amount of similarity is increased in this formula, semantic distance between those documents decreases;
2. Continue until K cluster of documents are achieved
3. Generate a collection of all terms belong to all documents in a cluster
4. Select term with the highest weight as the representative name for that cluster.

This algorithm classifies documents in different classes with proposed name for each class.

5 Evaluation

A collection of 500 research papers available on the internet were gathered in order to generate sample proposals, though the proof of effectiveness of this approach is done heuristically. These research papers are selected from four different categories in the context of Computer Science including: *Natural Language Processing (NLP)*, *Software Product Line*, *Semantic Web*, and *Software Testing*. The documents have been divided into sections (or regions) and converted to the HTML. This allows building the DOM, so perform the first step of information extraction. The DOM allowed to automatically converting the document into XML, where XML tags represent the document's sections (or regions).

The evaluation of the Fuzzy system is performed when this system works as a component of another system. As a result, the idea for evaluation was to build a text classifier and use the Fuzzy-based system for term weighting and term selection; then compare the performances of the text classifier using traditional term selection and using the Fuzzy-based system. The question is how to select an appropriate text classifier that can reflect the effects of the Fuzzy system. In the most important survey for the text classifiers, Sebastiani and his colleagues divided text classifiers in two major categories including: a) Parametric classifiers, where training data is used to

approximate the parameters of a text classifier such as Naïve Bayes, which is a probabilistic approach for text classification, and b) Non-parametric classifiers, which can employ VSM for document representation such as K- Nearest- Neighbor (KNN) [7].

Since, the proposed Fuzzy system uses VSM for document representation, Non-parametric classifiers are able to classify the performance of a Fuzzy system. For the current work, the KNN classifier, one of the most known classifiers, is used for evaluating the document in the Fuzzy system. The basic idea behind the KNN is to first label a set of training data with different categories; afterward for each new document, *K*, which is a constant; documents which have the most similarity (by using cosine similarity metric [13]) with the query documents are selected. The majority category among *K* documents is assigned as the category of the query document. By using the training set, a classifier is built and the test set is used to evaluate the performance of the classifier. Text classifiers are usually evaluated by the following criteria:

$$Precision = \frac{\sum_{i=1}^{m} TP_i}{\sum_{i=1}^{m} TP_i + \sum_{i=1}^{m} FP_i}, \tag{9}$$

$$Recall = \frac{\sum_{i=1}^{m} Tp_i}{\sum_{i=1}^{m} Tp_i + \sum_{i=1}^{m} FN_i}, \tag{10}$$

$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall}, \tag{11}$$

where, in a given test set, TP_i (true positive under the category of c_i) is the number of correct results. In other words, a correct result refers to the document if the test set is correctly categorized by the classifier. FP_i (False positive under the category of c_i) denotes the number of documents in the test set that the classifier has wrongly predicted in the category of c_i. Also, FN_i (false negative under the category ofc_i) is the number of missing true results, and m is the number of existing categories in the training set. TN_i (true negative under the category ofc_i) is the number of correctly identified wrong results.Precision demonstrates how many false positive documents are considered by the text classifier. Moreover, Recall represents the sensitivity of a text classifier and shows how many true results are missed by the classifier. Furthermore, F-measure demonstrates the performance of a text classifier, and is widely used in order to evaluate text classifiers.

In the present work, 80 percent of documents are allocated for the training set and 20 percent are considered the test set. In order to build the KNN classifier, Weka, which is a popular tool to implement machine learning algorithms, is used [14]. Working with Weka and the way that data is used to evaluate the classifier are explained in detail in Appendix A. Before the assessment of the Fuzzy system, the *Summary* section in the proposals is assumed as the *Preferred* Section. It denotes the

fact that the human expert believes the most informative part of the document is the *Summary* section. The Fuzzy system considers the importance of the *Summary* section and also takes into account other important terms in the whole document with regard to the Fuzzy rules. However, there is no rule for traditional term selection methods to recognize the importance of different sections in the document. In fact, traditional approaches do not distinguish between different parts of a document; they just process the whole document as a set of words.

The performances of the KNN classifier using the Fuzzy-based term selection are compared with traditional term selection methods such as DF and TF-IDF; and the result is illustrated in Fig. 7. The Y axis shows the F-measure (the performance) of the KNN classifier and the X axis shows the number of documents which are processed for the evaluation study. The result shows that the Fuzzy system outperforms traditional term selection approaches, when there are some preferences on the different parts of the information. It can be observed form Fig. 7 that when the number of documents is increased, the performance (F-measure) increases. The reason for this behavior is that there are more examples available for the classifier to decide in the classification of documents. The result is achieved by considering 400 terms (features) with the highest rank for the VSM model of each document.

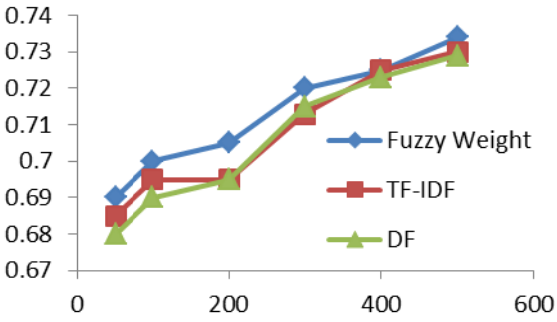


Fig. 7. F-Measure of 5-NN classifier with different documents for 400 features

6 Conclusion

In this work, we proposed a new method to represent and classify semi-structured documents. We have designed a fuzzy rule-based system as guidance for document representation. We also use this fuzzy system to measure the similarity between documents and categorize them in a set of classes. The next step of our proposed approach is to evaluate: 1) fuzzy logic –based term weighting system and 2) document similarity metric. In order to evaluate the first one, we should measure any performance criteria of a system that use this method of document representation. Basically, it cannot be evaluated solely and its performance should be compared as a component of a bigger system. We have evaluated the F-Measure of the result and

compare it with traditional approaches of document representation. In order to evaluate the second part, we can use the human correlation metric which is proposed in [12]. We aim to use these metrics and evaluations methods to illustrate how effective our approach is.

References

[1] Garc´ıa-Plaza, P., Fresno, V., Mart´ınez, R.: Web page clustering using a fuzzy logic based representation and self-organizing maps. In: Proceedings of the WI-IAT, pp. 851–854 (2008)

[2] Aggarwal, C.C., Zhai, C.X.: A Survey of Text Classification Algorithms. Mining Text Data, pp. 163–222. Springer, US (2012)

[3] Forman, G.: Feature Selection for Text Classification. In: Liu, H., Motoda, H. (eds.) Computational Methods of Feature Selection, pp. 257–276. CRC Press/Taylor and Francis Group (2008)

[4] Salton, G., Wong, A., Yang, C.S.: A Vector Space Model for Automatic Indexing. Communications of the ACM 18(11), 613–620 (1975)

[5] Thomas, H.: Probabilistic latent semantic analysis. In: Uncertainity in Artificial Intelligence (1999)

[6] Lan, M., Lu, Y.: Supervised and traditional term weighting methods for automatic text categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(4), 721–735 (2009)

[7] Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys 34, 1–47 (2002)

[8] Lin, D.: An Information-Theoretic Definition of Similarity. In: Proc. Int'l Conf. Machine Learning, ICM (1998)

[9] Biletskiy, Y., Brown, J.A., Ranganathan, G.R.: Information extraction from syllabi for academic e-Advising. Expert Systems with Applications 36(3), 4508–4516 (2009)

[10] Jang, R., Mizutani, E.: Neuro-Fuzzy and Soft Computing. Prentice Hall, Englewood Cliffs (1997)

[11] Mitchel, T.M.: Machine Learning. Mc Graw Hill (1996)

[12] Lee, M., Pincombe, B., Welsh, W.: An Empirical Evaluation of Models of Text Document Similarity. In: Proceedings of the 27th Annual Conference of the Cognitive Science Society, pp. 1254–1259 (2005)

[13] Huang, A.: Similarity measures for Text Document Clustering. In: Proceedings of New Zealand Computer Science Research Student Conference, July 3, 2012, pp. 49–56. Weka (2008)

[14] Weka digital library (2010), `http://www.cs.waikato.ac.nz/ml/weka/` (retrieved July 3, 2012)

Data Mining Application for Cyber Credit-Card Fraud Detection System

John Akhilomen

j.akhilomen1@unimail.derby.ac.uk

Abstract. Since the evolution of the internet, many small and large companies have moved their businesses to the internet to provide services to customers worldwide. Cyber credit-card fraud or no card present fraud is increasingly rampant in the recent years for the reason that the credit-card i s majorly used to request payments by these companies on the internet. Therefore the need to ensure secured transactions for credit-card owners when consuming their credit cards to make electronic payments for goods and services provided on the internet is a criterion. Data mining has popularly gained recognition in combating cyber credit-card fraud because of its effective artificial intelligence (AI) techniques and algorithms that can be implemented to detect or predict fraud through Knowledge Discovery from unusual patterns derived from gathered data. In this study, a system's model for cyber credit card fraud detection is discussed and designed. This system implements the supervised anomaly detection algorithm of Data mining to detect fraud in a real time transaction on the internet, and thereby classifying the transaction as legitimate, suspicious fraud and illegitimate transaction. The anomaly detection algorithm is designed on the Neural Networks which implements the working principal of the human brain (as we humans learns from past experience and then make our present day decisions on what we have learned from our past experience). To understand how cyber credit card fraud are being committed, in this study the different types of cyber fraudsters that commit cyber credit card fraud and the techniques used by these cyber fraudsters to commit fraud on the internet is discussed.

Keywords: Cyber credit card fraud, cyber credit-card fraudsters, black-hat hackers, neural networks, data mining.

1 Introduction

Imagine a scenario at the end of the month where you as a credit-card owner received your credit-card statement; you noticed on your credit-card statement that a purchase was made on your credit-card for a blackberry phone you never bought nor made an order for. You called your credit card company to explain to them that you never made this purchase but you were told that you did made that order since it was recorded on their system the purchase made with your legitimate information. Then they went ahead to tell you that from their logged file, you actually made that purchase for a blackberry on "www.ebay.com" and the monetary transaction was

P. Perner (Ed.): ICDM 2013, LNAI 7987, pp. 218–228, 2013.

successfully made to EBay. Now afterwards your credit-card company decided to investigate further and they called the internet company EBay. EBay checking their logged file for the transaction told your credit-card company that the blackberry phone was delivered to a shipping address in Turkey while you the actual credit-card owner lived in the USA. Obviously in a case like this, you are a victim of an internet credit-card fraud or no card present fraud. When your credit card or credit card information is stolen and used to make unauthorized purchases on e-commercial systems on the internet, you become a victim of internet credit card fraud or no card present fraud. This is nothing new and there is nothing unusual about this because identity theft and credit-card fraud are present-day happenings affecting many people and involving substantial monetary losses. Fraud is a million dollar business and it's increasing every year. The PwC global economic crime survey of 2011 suggests that 34% of companies worldwide have reported being victim of fraud in the past year and increasing from 30% as reported in the year 2009[9]. Fraud is as old as humanity itself and can take an unlimited variety of different forms. However, in recent years, the development of new technologies like the internet has provided further ways in which fraudsters can commit fraud. Fraud is a very skilled crime; therefore a special method of intelligent data analysis to detect and prevent it is necessary[11]. These methods exist in the areas of Knowledge Discovery in Database, Data Mining, Machine Learning and Statistics. They offer applicable and successful solutions in different areas of fraud crime. The aim of this study is actually focused on modeling an applicable system for detecting fraud in a real-time transaction on the internet. This model implements the anomaly detection algorithm of Data Mining, using Neural Networks to learn patterns used by a particular credit-card owner and then match the patterns learned with the pattern of the current transaction to detect anomalies.

2 What Is Cyber Credit-Card Fraud or No Card Present Fraud?

Recent and current scholars investigating credit-card fraud have divided credit-card fraud into two types; the online credit card fraud (or no card present fraud) and the offline credit card fraud (card present fraud)[1]. The online credit-card fraud (in this paper is cyber credit card fraud) is committed with no presence of a credit-card but instead, the use of a credit-card information to make electronic purchase for goods and services on the internet. The offline credit-card fraud is committed with the presence of a credit-card which in most cases have been stolen or counterfeited and thereby used at a local store or a physical location for the purchase or some goods or services. However, to define cyber credit-card fraud, it is a scenario where the credit-card information of a credit-card owner has been stolen, or in some cases valid credit-card information has been uniquely generated (just like credit-card companies or issuers do) and thereby used for electronic payment on the internet or via the telephone. In most cases, no I.T or computer skill may be required to commit online credit-card fraud because of the different techniques in which credit-card information can be stolen by cyber fraudsters.

3 Who Are the Cyber Credit-Card Fraudsters?

3.1 Credit-Card Information Buyers

They are fraudsters with little or no professional computer skills (e.g. Computer Programming, Networking, etc.) who buy hacked (or stolen) credit-card information on an illegal “credit-card sales” website. They buy this credit-card information with the intention of making electronic payment for some good and services on the internet.

3.2 Black Hat Hackers

Recent research on Hackers in terms of Computer Security defined a "black hat hacker" (also known as a cracker) as a hacker who violates computer security with malicious intent or for personal gain[8]. They choose their targets using a two-pronged process known as the "pre-hacking stage"; Targeting, Research and Information Gathering, and Finishing the Attack. These types of hackers are highly skilled in Computer Programming and Computer Networking and with such skills can intrude a network of computers. The main purpose of their act of intrusion or hacking is to steal personal or private information (such as credit-card information, bank-account information, etc.) for their own personal gain (for instance creating a “credit-card sales” website where other cyber credit-card fraudsters with little or no computer skills can buy credit-card information).

3.3 Physical Credit-Card Stealer

They are the type of fraudsters who physically steal credit-cards and write out the information on them. They physically steal these plastic credit-cards (maybe by pick-pocketing in a crowded place) and write out the credit-card’s information with the intention of using this credit-card information to make electronic payment for some good and services on the internet.

4 Different Techniques for Credit-Card Information Theft by Cyber Credit-Card Fraudsters:

In other to detect cyber credit-card fraud activities on the internet, a study conducted on how credit-card information is stolen is a good approach. Listed below are studied different techniques which are used for credit-card fraud information theft.

4.1 Credit-Card Fraud Generator Software

These are software written to generate valid credit-card numbers and expiry dates. Some of these software are capable in generating valid credit-card numbers like credit-card companies or issuers because it uses the mathematical Luhn algorithm that credit-card companies or issuers use in generating credit-card numbers to their credit-card consumers or users. In other cases, this software is written by black-hat hackers with hacked credit-card information stored on a database file from which the software

can display valid credit-card information to other type of cyber credit-card fraudsters who have bought the software to use. This technique is some cases used by black-hat hackers to sell their hacked credit-card information to other online credit-card fraudsters with little or no computer skills.

4.2 Key-Logger and Sniffers

Black-hat hackers with professional Programming or computer skills are able to infect a computer by installing and automatically running sniffers or key-logger computer programs to log all keyboard inputs made into the computer on a file with the intention of retrieving personal information (like credit-card information, etc.). These black-hat hackers or fraudsters are able to infect users' computers by sending spam emails to computer-users requesting them to download free software or games, or sometimes they create some porn-sites so that when these computer-users browse these porn sites or download those free software or games, the sniffers or key-loggers are automatically downloaded, installed and ran on the users' computers. While the sniffer or key-logger is running under the users' computer, they sniff and log all the keyboard-input made by the user over a connected network. Therefore, any computer-user can unknowingly share their private information (credit-card information, etc.) through viral-infecting software such as these. In some cases, no Programming or computer skill is required to sniff a computer-user's key-board input because this software are also being shared or sold to other cyber credit-card fraudsters with little or no computer skills.

4.3 Spyware, Site-Cloning and False Merchant Sites

They are software created by black-hat hackers, installed and ran on users' computer to keep track of all their website activities. From knowing the website activities of the victimized computer-user on the internet, electronic or banking websites regularly visited by the computer-user can be cloned and sent to the user for usage with the intention of retrieving personal or private information (like bank log-in's). In the case of false merchant sites, fake websites can be created to advertise and sell cheap products to computer-users, and thereby asking for payment via credit-card. If a credit-card payment is made on any of these fake merchant sites, the user's credit-card information is therefore stolen.

4.4 CC/CVV2 Shopping Websites

Cyber credit-card fraudsters with no professional computer skills can buy hacked credit-card information on these websites to use for fraudulent electronic payment for some goods and services on the internet.

4.5 Physical Stolen Credit-Card Information

Fraudsters can physically steal the credit-card of a user to write out the credit-card information and then use for fraudulent electronic payment on the internet.

5 Methodology

5.1 Implementing Data Mining Techniques for Credit Card Fraud Detection System

Data mining is popularly used to effectively detect fraud because of its efficiency in discovering or recognizing unusual or unknown patterns in a collected dataset. Data mining is simply a technology that allows the discovery of knowledge in a dataset. In other words, with Data mining knowledge can be discovered in a dataset. Data is collected from different sources into a dataset and then with Data mining, we can discover patterns in the way all data in the dataset relates with another and then make predictions based on the patterns discovered. Data mining takes a dataset as an input and produces models or patterns as output. One of the popular effective Data mining techniques used in data security is the Neural Networks. The concept of the Neural Networks is designed on the functionality of the human brain. From kindergarten until college, we are developed from an infantry stage of life unto the adult stage through different experiences or a set of data through how we're schooled. And we use this past experience or training we have acquired to make present day decisions. This is the Neural Networks. The Neural Networks makes predictions and classifications from what it has learned. The Anomaly detection algorithm is an implementation of the Neural Networks. Anomaly detection (sometimes called deviation detection) is an algorithm implemented to detect patterns in a given data set that do not conform to an established normal behavior[10]. The patterns thus detected are called anomalies and often translate to critical and actionable information in several application domains. The Anomaly detection is categorized into three; Unsupervised anomaly, Semi-supervised and Supervised anomaly detection. Unsupervised anomaly detection techniques detect anomalies in an unlabeled test data set under the assumption that the majority of the instances in the data set are normal by looking for instances that seem to fit least to the remainder of the data set. Supervised anomaly detection techniques require a data set that has been labeled as "normal" and "abnormal" and involves training a classifier (the key difference to many other statistical classification problems is the inherent unbalanced nature of outlier detection). Semi-supervised anomaly detection techniques construct a model representing normal behavior from a given normal training data set, and then testing the likelihood of a test instance to be generated by the learnt model[10]. As seen in the diagram on fig. 1, this data mining application uses Supervised Anomaly detection to detect credit card fraud in a transaction and thereby classifies a transaction as Ok, suspicious fraud or illegitimate transaction.

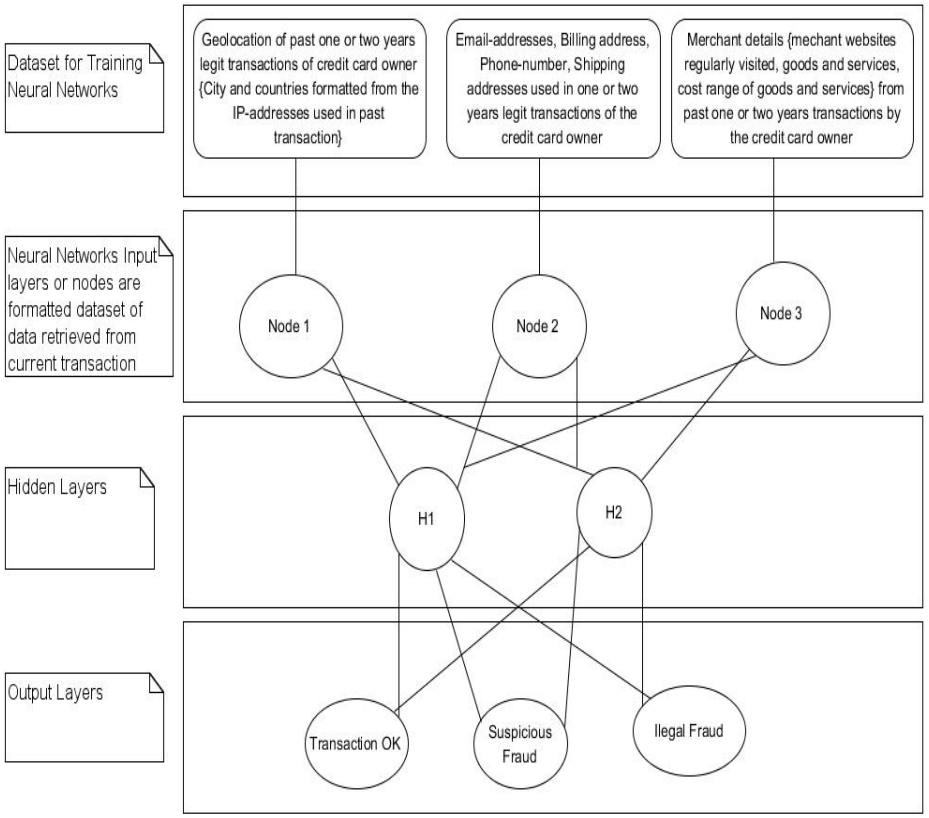


Fig. 1. The Learning and Classification of Neural Networks in the system

5.2 Credit Card Fraud Detection Model

This Data mining application applies the anomaly detection algorithm to detect cyber credit card fraud in an online credit-card transaction implementing Pattern recognition with Neural Networks. Anomaly detection algorithm is a technique used in Data mining applications to detect specific patterns or relations within the data provided for Fraud detection process. There is a fixed pattern to how credit-card owners consume their credit-card on the internet. This fixed pattern can be drawn from legitimate regular activities of the credit-card owner for the past one or two years on its credit-card; the regular merchant websites the credit-card owner regularly makes electronic payment for goods and services, the geographical location where past legitimate transactions have been made, the geographical location where goods have been shipped to by the credit-card owner, the email-address and phone number regularly used by the credit card owner for notification. Using the Neural Network technology, the computer-program or software can be trained with this fixed pattern to use it as knowledge in classifying a real-time transaction as fraudulent or legitimate transaction. In this Data mining application for credit-card fraud detection, the anomaly detection algorithm is implemented for cyber credit-card fraud detection

process. Once the data to be analyzed is selected, the anomaly detection algorithms will be applied to perform a data mining process for matching the behavior of the current transaction if it differs in behavior with the owner's past transactions on its credit-card. If the behavioral pattern in the current transaction differs with the learned pattern of the original credit-card owner, the system will continue to match the pattern of the current transaction if it's similar with past cyber credit-card fraud transactions. If the system returns false (of mismatch patterns between the current transaction and past fraud transactions) then the system classifies the transaction as suspicious fraud but if true, then the system will classify the transaction as illegal fraud transaction.

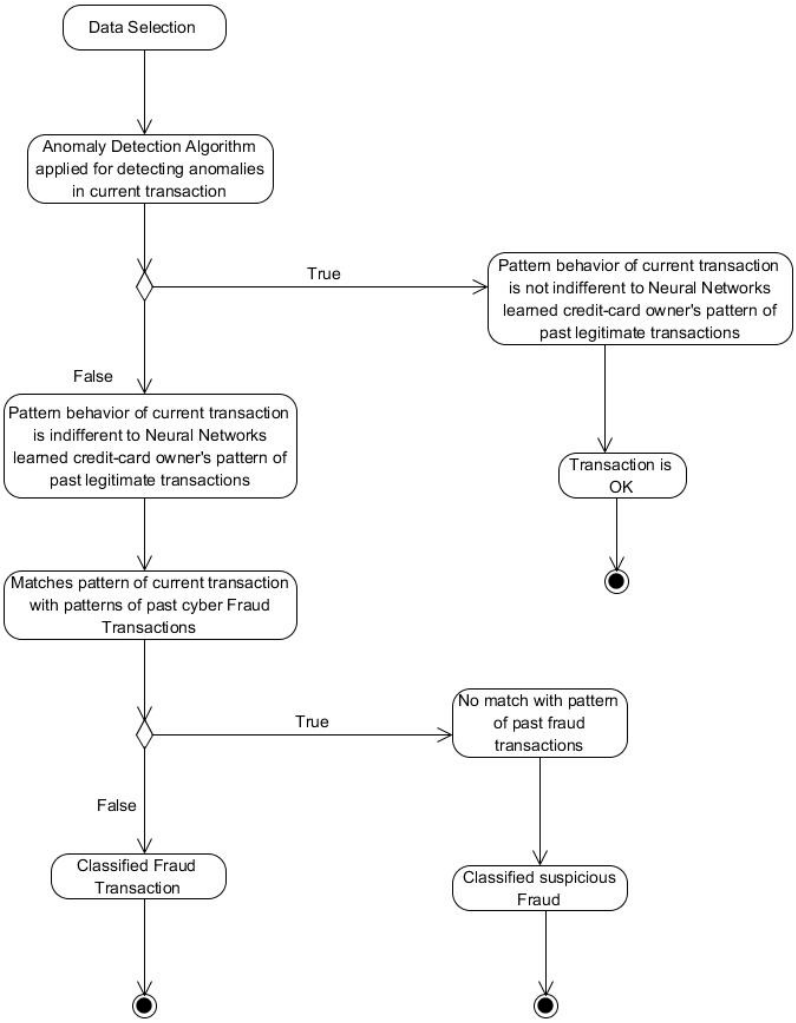


Fig. 2. Shows the system's model for credit card fraud detection process in a transaction

5.3 Pattern Recognition to Train Neural Networks :

Geolocation of Real-Time Transaction. The geolocation technology provides the absolute geographical location of an internet-connected computer by its IP address. An IP address is a unique network identifier issued by an Internet Service Provider to a computer-user every time they are logged on to the Internet[12]. This Data mining application is trained with IP-addresses (City and Country location being formatted from the IP-addresses) of internet-connected-computers the credit-card owner has used in the past one or two years legitimate transaction on its credit-card. This is a good mechanism to train Neural Networks for cyber credit-card fraud detection because in training Neural Networks with the City and Country locations formatted from IP-addresses where the credit-card owner has regularly made legitimate transactions from for the past one or two years, Neural Networks can know if the internet-connected-computer of the current transaction behaves in pattern like the internet-connected computers the credit-card owner has regularly made his past one or two years legitimate credit-card transactions. While this is a very good anti-fraud mechanism and useful for tracking fraudsters, the IP addresses can also be changed using Proxy servers. Anonymous proxy servers allow Internet users to hide their actual IP address and run their computers behind a fake IP address of their desired region[13]. The main purpose using a proxy server is to remain anonymous or to avoid being detected. Fraudsters hide themselves behind anonymous proxy servers to commit credit-card fraud on the internet. This Data mining application automatically flags for suspicious fraud if a proxy-server is detected in a transaction.

Email Address and Phone Number. When a credit-card is issued to an individual by a credit-card issuer or company, an email address or phone number from the individual is registered with the credit card so that the individual can receive notification via telephone or email of any transaction that's been made on their credit-card. For this reason, fraudsters do use different email-addresses and phone numbers when committing cyber fraud on credit cards. Although, It is important to take note that the cyber fraudsters do not only use email-addresses registered with free domains (like Yahoo, Google or Hotmail), but also they do pay to get registered email-addresses with non-free domains. Therefore, in this data mining application, Neural Networks will be trained with the email addresses and phone number the credit-card owner has used in past one or two years internet credit-card transactions.

Shipping Address. Although it is not uncommon for people sending gifts to others to request different shipping address. It is very difficult to retrieve goods or apprehend fraudsters once the goods have left the country of residence of the original credit-card owner. Fraudsters will possibly not send goods to the legitimate cardholder's billing-address. But it is possible that credit-card owners will send goods to legitimate shipping address different to their billing address. Therefore, in this data mining application, Neural Networks will be trained with Shipping addresses and oversea orders used by the credit-card owner in past one or two years transactions.

Merchants' Websites, Regular Good and Services Purchased in Past Credit Cardholder's Transactions. Neural Networks will be trained with the merchant websites the credit-card owner has regularly visited and the type of goods and services they have regularly purchased on its credit-card for the past one or two years. Neural Networks will be trained with the cost range of goods and services purchased in the past one or two years transactions of the credit cardholder's credit card.

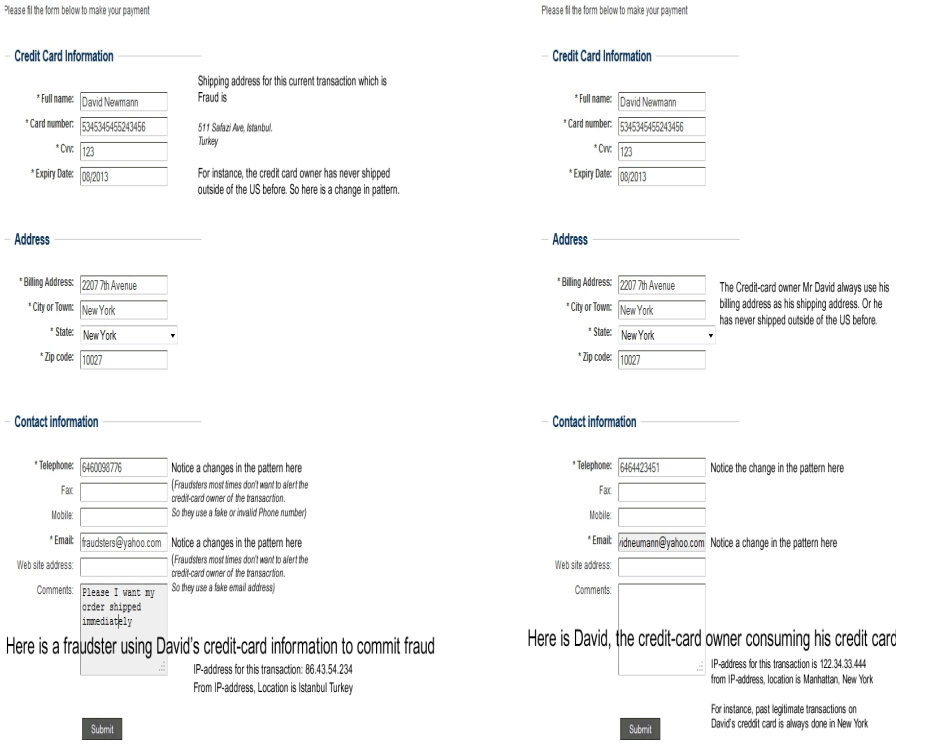


Fig. 3. Shows a form used to purchase order by a cyber credit-card fraudster and an actual credit card owner

6 Conclusion

Efficient and well organized credit-card fraud detection system is a crucial requirement for credit card issuing companies to run their business well. This is because of the increase in the credit card fraud activities in the present day happenings. But the internet related credit card fraud and as sometimes known as no-card present fraud seems to be rampant increasing in the recent years more than the card-is-present fraud. It is easier and safer to commit online credit fraud because of its anonymousity and the fact that in most cases, no engineering or computing skill maybe required. As seen in this paper, there are fraudsters who just buy credit-card

information from other fraudsters and then commit fraud. For this reason, the internet credit card fraud is increasing over the years and has made credit card issuing companies lost millions of dollars. Credit card fraud detection has drawn quite a lot of interest from research community and a number of techniques have been proposed to counter fraud. In this paper, a data mining application has been modeled as a subsystem which can be used with software systems and applications in financial institutions to detect credit-fraud in a transaction on the internet. This Data mining application accepts input formatted on a pattern on which a transaction is being executed and matches it with the credit-card holder's patterns of its credit-card online consumptions it's been trained with to classify a real-time transaction as legit, suspicious fraud or illegitimate transaction. The data mining application modeled in this paper uses the anomaly detection algorithm of the Neural Networks to detect fraud in a real-time transactions and it not prone to errors because of its classification of Transactions (legitimate, Suspicious Fraud and illegitimate). In the case of the suspicious fraud classification, the financial institution using the system can investigate further by calling the credit-card owner regarding the suspicious fraudulent transaction.

7 Future Work

More studies would be conducted on patterns upon which credit card owners consume their credit-cards on the internet after which the new gathered information from the undertaken studies is updated on the system's model for training Neural Networks. This system would be developed and then tested using real world transaction data. On this paper, a system's model which implements the Classification and Prediction technique of Data Mining is discussed and designed for cyber credit-card fraud detection. However, there is need for the development and testing of the system from the discussed model described on this paper.

References

1. Al-Khatib, A.M.: Electronic payment fraud detection techniques. World of Computer Science and Information Technology Journal 2(4), 137–141 (2012)
2. Ogwueleka, F.N.: Data mining application in credit-card Fraud detection system. Journal of Engineering Science and Technology 6(3), 311–322 (2011)
3. Yashpal, S., Chauhan, S.: Neural networks in data mining. Journal of Theoretical and Applied Information Technology 5(6), 37–42 (2005-2009)
4. Khyati, C., Bhawna, M.: Exploration of data mining techniques in fraud detection: credit-card. International Journal of Electronics and Computer Science Engineering I(3), 1765–1771
5. Dhecpa, V., Dhanapal, R.: Analysis of credit-card fraud detection methods. International Journal of Recent Trends in Engineering 2(3), 126–128 (2009)
6. Khyati, C., Jyoti, Y., Bhawna, M.: A review of fraud detection techniques: credit-card. International Journal of Computer Applications 45(I), 39–44 (2012)

7. Sam, M., Karl, T., Bram, V.: Credit-card Fraud Detection Using Bayesian and Neural Networks, http://www.personeel.unimaas.nl/k-tuylslpublicationslpaperslmaenf02.pdf (accessed December 12, 2012)
8. Hacker (computer security) : Wikipedia, the free encyclopedia, http://en.wikipedia.org/wiki/Hacker_(computer_security) (accessed December 12, 2012)
9. Cybercrime: protecting against the growing threat Global Economic Crime Survey – PWC Global Economic, http://www.pwc.com/en_GX/gx/economic-crime-survey/assets/GECS_GLOBAL_REPORT.pdf (accessed December 12, 2012)
10. Anomaly Detection: Wikipedia, the free encyclopedia, http://en.wikipedia.org/wiki/Anomaly_detection (accessed December 12, 2012)
11. Data Analysis Techniques for Fraud Detection, http://en.wikipedia.org/wiki/Data_Analysis_Techniques_for_Fraud_Detection (accessed December 12, 2012)
12. Preventing Credit Card Abuse: Anti-Fraud Strategies, http://www.lawzilla.com/content/fed-bus-12301.shtml?&lang=en_us&output=json&session-id=3cd3dad0fc218a1ad59460ff032578fd (accessed December 12, 2012)
13. Precautions for internet traders to prevent fraudulent credit card, http://www.technade.com/2007/02/precautions-for-internet-traders-to_25.html?&lang=en_us&output=json&session-id=3cd3dad0fc218a12578fd (accessed December 12, 2012)

Feature Representation for Customer Attrition Risk Prediction in Retail Banking

Yanbo J. Wang[1,2], Gang Di[3], Junxuan Yu[2,4], Juan Lei[5], and Frans Coenen[6]

[1] Information Management Center, China Minsheng Banking Corp., Ltd.
No. 2, Fuxingmennei Avenue, Xicheng District, Beijing, 100031, China
[2] Institute of Finance and Banking, Chinese Academy of Social Sciences
No. 5, Jianguomennei Dajie, Beijing, 100732, China
wangyanbo@cmbc.com.cn
[3] Department of Science and Technology, The People's Bank of China
No. 32, Chengfang Jie, Xicheng District, Beijing, 100800, China
dg@pbc.gov.cn
[4] Department of Financial Markets, Longjiang Bank
No. 436, Youyi Lu, Harbin, Heilongjiang, 150018, China
tinyfour99@126.com
[5] Department of Retail Banking, China Minsheng Banking Corp., Ltd.
No. 2, Fuxingmennei Avenue, Xicheng District, Beijing, 100031, China
leijuan2@cmbc.com.cn
[6] Department of Computer Science, University of Liverpool
Ashton Building, Ashton Street, Liverpool, L69 3BX, UK
Coenen@liverpool.ac.uk

Abstract. Nowadays, customer attrition is increasingly serious in commercial banks, particularly with respect to *middle-* and *high-valued* customers in retail banking. To combat this attrition it is incumbent for banks to develop a prediction mechanism so as to identify customers who might be at risk of attrition. This prediction mechanism can be considered to be a classifier. In particular, the problem of predicting risk of customer attrition can be prototyped as a *binary* classification task in data mining. In this paper we identify a set of features, for customer "attrition *vs.* non-attrition" classification, based on the RFM (Recency, Frequency and Monetary) model. The reported evaluation indicates that proposed set of features produces a much more effective classifier than that generated using previously suggested features.

Keywords: Classification, Customer Attrition Risk, Feature Representation, Retail Banking, RFM Model.

1 Introduction

With increased competition within the banking industry, customer attrition/churn is of increasingly serious concern in commercial banking, particularly with respect to *middle-* and *high-valued* customers in retail banking. More and more commercial banks are turning to Customer Relationship Management (CRM) so as to retain their existing customers. In [8], the author clearly states that "*retaining customers is more*

P. Perner (Ed.): ICDM 2013, LNAI 7987, pp. 229–238, 2013.

profitable than building new relationships". In [5], the authors even attempt to clarify that attracting a new customer is about five times more costly than retaining an existing customer. Hence, "*the retention of existing customers has become a priority for businesses to survive and prosper*" [8]. Consequently, accurately identifying those customers who might be at risk of attrition has become an essential problem to be solved. For commercial banks in general, a prediction mechanism that can be used to classify whether an existing customer will churn (or not) in the near future (in the next business/observation period, e.g. a month or a quarter) is desirable. This prediction mechanism can be considered to be a classifier. In particular, the problem of predicting risk of customer attrition can be prototyped as a *binary* classification task in the context of data mining.

A number of techniques exist to support binary classification such as Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT), Rule Generation (RG), and so on. With respect to the prediction of customer attrition risk for retail banking previous work has suggested a set of 22 features that can be used for the desired prediction. In this paper we propose an alternative set of 7 features, for customer "attrition *vs.* non-attrition" classification, based on the RFM (Recency, Frequency and Monetary) model. To support our study, a set of experiments was run, using four *real* customer datasets (2 balanced *plus* 2 unbalanced) within the retail banking sector. Using a variety of classification techniques, the results presented in this paper demonstrate that our proposed 7 feature representation outperforms the previously proposed 22 feature representation.

The rest of this paper is organised as follows. The following section describes the data mining classification task and the related work in feature representation. Section 3 presents our proposed (RFM based) feature representation approach for banking customer "attrition *vs.* non-attrition" classification. Experimental results, based on the collected customer data from a *real* retail banking environment are shown in Section 4. Finally, conclusions and direction for future work are given at the end of this paper.

2 Classification and Feature Representation

2.1 Classification in Data Mining

Data Mining, also referred to as "Data Science", is "*the science of extracting useful information from large data sets or databases*" [3]. The domain of data mining is positioned at the intersection of data management, statistics, machine learning, pattern recognition, and other related areas. Classification is "*one of the most important and widely used data mining techniques, especially in the area of marketing and Customer Relationship Management (CRM)*" [15].

The aim of the classification process is to build a knowledge model that allows the value associated with one variable (the target attribute) to be predicted from the given values of the other variables (the descriptive attributes/features). In classification, the target attribute values are typically categorical values (such as "*Colour: red, yellow, green, blue*", "*Level: high, middle, low*", etc.), while the descriptive attribute/feature values can be categorical or numerical (for example "*Age: 25, 28, 41*", "*Price: 9.99, 15.00, 1380.60*", etc.).

In our study of retail banking customer attrition risk prediction, since the target attribute binary valued ("attrition" or "non-attrition"), we prototype our data mining task to be a *binary* classification task. The binary classification, also referred to as 2-class classification, "*learns from both positive and negative data samples, and assigns either a predefined category* (*class-label*) *or the complement of this category to each 'unseen' instance*" [12].

In previous work, a number of techniques/approaches have been proposed for the purpose of binary classification study, these include: Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT) and Rule Generation (RG). Each is briefly considered below:

- **SVM.** The SVM approach [1] aims to find a hypothesis $\hat{h}$ which minimizes the true error defined as the probability that $\hat{h}$ produces an erroneous result. SVM makes use of linear functions of the form: $f(x) = w^T x + b$, where w is the weight vector, x is the input vector, and $w^T x$ is the inner product between w and x. The main concept of SVM is to select a hyper-plane that separates the positive and negative examples while maximizing the smallest margin. Standard SVM techniques produce *binary* classifiers as opposed to multi-classifiers. Two common approaches to support the application of SVM techniques to the multi-class problem are "One Against All" (OAA) and "One Against One" (OAO).
- **LR.** The LR approach was first introduced in the 1970s; "*it became available in statistical packages in the early 1980s*" [9]. LR is a type of regression analysis used for predicting the outcome of categorical dependent variables (i.e. "yes" *vs.* "no", or "high" *vs.* "low", etc.), based on independent variables (descriptive features). This technique attempts to model the probability of a "class/¬class" outcome using a linear function of the descriptive features, and then applying the log-odds of "class" (the *logit* of the probability) to fit the regression.
- **DT.** In the case of the DT approach this was first introduced in the mid 1980s [10]. The approach solves the classification problem using a "greedy" strategy. C4.5 [11] is the best known DT classification algorithm which operates by recursively splitting the training data-instances on the data-attribute that produces the maximum information gain to generate the tree. This tree is then pruned according to an error estimate. Finally the result from the training, a tree classifier, is used to classify "unseen" data-instances.
- **RG.** The RG approach aims to solve the classification problem by generating a set of "*IF–THEN*" patterns/rules, whereby each rule is expressed in the form of "*feature*(*s*) $\Rightarrow$ *category*". The generated set of (human readable and understandable) rules represents the (built) classifier and presents to the end users why and how the classification predictions have been made. A well known RG algorithm is RIPPER [2] which operates using a rule-induction process.

2.2 Target Attribute Representation

The target attribute of our classification task is whether an existing customer will churn/attrite (or not) in the near future. Based on the suggestions given by domain

experts (professional financial/customer managers from retail banking sector), our study simply follows the idea presented in [13] and defines customer attrition as the situation where "*the reduction of a customer asset quarter-average-balance is higher than 70% during the last three months*".

2.3 Descriptive Feature Representation

In previous studies on retail banking customer attrition prediction [6] [7], 22 descriptive attributes/features were used to predict whether existing customers were likely to churn or not. These 22 attributes were grouped into the following five categories as follows:

- **Demographical information**, where typical features are "customer age" and "customer gender";
- **Account/Transaction information**, for example "number of years since account opening";
- **Financial information**, which include "customer deposits' month-average-balance for each of the last six months" and "the customer deposits' balance divided by the customer assets' balance at the end of each of the last six months";
- **Product information**, where typical features are the "number of product holdings", "whether a foreign-exchange customer or not", "whether a net-banking customer or not", "whether a tripartite depository customer or not", and "whether the customer is using account information messenger"; and
- **Loan information**, where typical features are "whether a loan customer or not" and "whether holding a loan currently".

With respect to the work described in this paper these 22 descriptive features were adopted as a benchmark. In addition, as will become clear in the following sub-section, seven further features were derived.

3 Proposed Feature Representation Approach

In this section, we introduce our proposed feature representation approach for retail banking customer "attrition *vs.* non-attrition" classification using the RFM model. The RFM model is first introduced in Sub-section 3.1 and the representation in Sub-section 3.2.

3.1 The RFM Model

The well-known RFM (Recency, Frequency and Monetary) model was first introduced by Hughes in 1994 [4]. In a marketing context, this model considers the following three concepts:

- **Recency**, how long ago a customer last made a purchase;
- **Frequency**, how many purchases the customer makes in a given time period; and
- **Monetary**, the total amount of money spent by the customer in a given time period.

In [14], the authors summarise the RFM model as "*a behavior-based model used to analyze the behavior of a customer and then make predictions based on the behavior in the database*".

3.2 The Proposed Approach

In our banking context, we only consider the Monetary element of the RFM model and note that it can be substituted for by the customer's deposit balance or the customer's asset balance. For the time being we also chose to ignore the Frequency concept. We also considered that the Recency concept can be presented in the context of the concept of Monetary element, thus the month-average-balance value in the last month, the last two months, the last three months, and so on.

3.2.1 Derived Descriptive Features

The seven new descriptive features derived, augmenting those presented above, may be categorized as follows:

- **Deposits' balance.** Two descriptive features: "the ratio between the deposits' average-balance in the most recent month and the average-balance over the last three most recent months" and "the ratio between the deposits' average-balance in the most recent month and the average-balance over the last six most recent months". These two features serve to capture customer behaviour with respect to the changing/moving of deposit balances over the past six months.
- **Assets' balance.** Five descriptive features generated using the same philosophy as for the Deposits' balance category, but considering "the ratio between the assets' average-balance in the most recent month (recent most two months, three months, four months and five months) and this average-balance in the most recent six months". These five features serve to demonstrate customer behaviour with respect to changing/moving asset balances in the past six months. The main reason for having five features instead of two features in this category is that some of the financial planning products (counted as being part of customer assets) are short-term products, i.e. 30 days or 60 days.

3.2.2 Mixed Descriptive Features

As already noted, in our retail banking application, the newly derived 7 features were mixed with the 22 previously identified features. The assumption was that an improved result would be produced.

4 Experimental Results

In this section, we present two groups of evaluations for our proposed feature representation approach, one that considers (class) balanced data and one that considers unbalanced data, using the customer data obtained from a *real* retail

banking environment. All the evaluations were conducted using the WEKA data mining workbench [1] [16] [17] [18]. The experiments were run on a 3.00 GHz Pentium(R) Dual-Core CPU with 1.96 GB of RAM running under the x86 Windows Operating System. For the experiments we compared the classification effectiveness with respect to the Benchmark 22 features used in [6] [7], and the 7 features derived using the RFM model and a combination of the two (RFM mixed).

4.1 Description of Data

From a *real* commercial bank's EDW (Enterprise Data Warehouse), we collected (and anonymised) the retail banking VIP customer (attrition) data over a nine month period. The class-labels (target attribute values) of this dataset are "attrition" *vs.* "non-attrition", in other words the churn status of each customer between the seventh and ninth month. The 29 extracted features (descriptive attributes) may be grouped into the six categories identified above ("demographical", "account/transaction", "financial", "product", "loan", and "RFM" information).

After the data cleansing process where data-instances with missing and/or noisy values were eliminated, we:

- **Creating two balanced datasets:** by randomly selecting 2000 "attrition" plus 2000 "non-attrition" data-instances from the cleaned data collection;
- **Creating two unbalanced datasets:** by randomly selecting 1200 "*attrition*" plus 2800 "*non-attrition*" data-instances from the cleaned data collection.

The reason we included unbalanced datasets in our experiments was that the number of "attrition" customers is always less than the number of "non-attrition" customers, and usually we say that about 15% to 35% of customers churn each year in retail banking.

4.2 Description of Classification Approaches

In WEKA version 3.6.4, the implementation of SVM is SMO; the implementation of LR is called Logistic; the implementation of C4.5/C5.0 DT classification is J48; and the RIPPER RG classification is JRip. In our experiments, we ran these implemented methods using our prepared data and the WEKA platform. Note that for both J48 and JRip the parameter "minNumObj" was chosen to be 20.

4.3 Description of Results

4.3.1 Using Balanced Data

In Table 1 (see as follows), we show the comparison between our proposed 7 feature representation approach and the benchmark (with 22 descriptive features) by using two balanced datasets. Table 1 (a) demonstrates the comparison using the overall accuracy of classification; Table 1 (b) shows the comparison based on the recall of attrition; and Table 1 (c) uses the precision of attrition.

[1] The well-known WEKA software, a Data Mining and Machine Learning Software in Java, may be obtained from `http://www.cs.waikato.ac.nz/~ml/weka/`

Table 1. Comparison of Feature Representation Approaches based on Four Different Classifiers (using 2 Balanced Datasets)

1 (a)	Dataset A			Dataset B		
	Bench-mark (22)	RFM-Derived (7)	RFM-Mixed (29)	Bench-mark (22)	RFM-Derived (7)	RFM-Mixed (29)
SMO	62.00	79.94	81.00	61.88	79.70	80.74
Logistic	67.20	79.54	80.22	66.80	80.06	80.18
J48	74.00	83.48	84.46	75.08	84.44	85.62
JRIP	74.24	82.96	84.94	75.02	84.40	85.84
Average	*69.36*	*81.48*	***82.66***	*69.70*	*82.15*	***83.10***
# Bests	*0*	*0*	***4***	*0*	*0*	***4***

1 (b)	Dataset A			Dataset B		
	Bench-mark (22)	RFM-Derived (7)	RFM-Mixed (29)	Bench-mark (22)	RFM-Derived (7)	RFM-Mixed (29)
SMO	63.40	71.20	73.40	60.60	70.60	72.90
Logistic	63.70	72.90	75.30	63.30	73.10	75.30
J48	71.70	82.20	80.90	71.80	80.90	82.60
JRIP	70.00	80.90	83.00	70.20	81.80	82.60
Average	*67.20*	*76.80*	***78.15***	*66.48*	*76.60*	***78.35***
# Bests	*0*	*1*	***3***	*0*	*0*	***4***

1 (c)	Dataset A			Dataset B		
	Bench-mark (22)	RFM-Derived (7)	RFM-Mixed (29)	Bench-mark (22)	RFM-Derived (7)	RFM-Mixed (29)
SMO	61.70	86.20	86.60	62.20	86.30	86.40
Logistic	68.50	84.00	83.50	68.10	84.90	83.40
J48	75.10	84.40	87.10	76.80	87.10	87.90
JRIP	76.50	84.40	86.30	77.70	86.30	88.30
Average	*70.45*	*84.75*	***85.88***	*71.20*	*86.15*	***86.50***
# Bests	*0*	*1*	***3***	*0*	*1*	***3***

From Table 1 (a, b and c), it can be observed that when using the 7 RFM-derived features instead of the 22 benchmark features the overall accuracy, recall of attrition and the precision of attrition of our customer "attrition *vs.* non-attrition" classification study were all increased significantly; when using the 29 mixed features rather than the 7 derived features, the performance of classification reached even better results with respect to the average overall accuracy, average recall of attrition, average precision of attrition, and the number of best cases.

4.3.2 Using Unbalanced Data

In Table 2 the results obtained from the experiments to compare the operation of our proposed classification process when applied to unbalanced datasets is presented. Table 2 (a) demonstrates the comparison using the overall accuracy of classification; Table 2 (b) shows the comparison based on the recall of attrition; and Table 2 (c) uses the precision of attrition.

Table 2. Comparison of Feature Representation Approaches based on Four Different Classifiers (using 2 Unbalanced Datasets)

2 (a)	Dataset C			Dataset D		
	Bench-mark (22)	RFM-Derived (7)	RFM-Mixed (29)	Bench-mark (22)	RFM-Derived (7)	RFM-Mixed (29)
SMO	69.90	83.40	84.18	70.04	84.54	85.22
Logistic	75.02	82.60	83.50	74.18	83.88	84.42
J48	78.92	86.78	87.96	79.04	86.80	88.44
JRIP	79.74	86.86	87.96	79.82	86.36	88.50
Average	*75.90*	*84.91*	***85.90***	*75.77*	*85.40*	***86.65***
# Bests	*0*	*0*	***4***	*0*	*0*	***4***

2 (b)	Dataset C			Dataset D		
	Bench-mark (22)	RFM-Derived (7)	RFM-Mixed (29)	Bench-mark (22)	RFM-Derived (7)	RFM-Mixed (29)
SMO	2.40	67.00	68.70	12.90	66.50	68.00
Logistic	25.30	61.10	63.70	24.30	60.60	63.30
J48	50.00	73.80	77.40	50.10	72.60	75.90
JRIP	51.20	74.50	75.70	54.90	72.90	75.50
Average	*32.23*	*69.10*	***71.38***	*35.55*	*68.15*	***70.68***
# Bests	*0*	*0*	***4***	*0*	*0*	***4***

2 (c)	Dataset C			Dataset D		
	Bench-mark (22)	RFM-Derived (7)	RFM-Mixed (29)	Bench-mark (22)	RFM-Derived (7)	RFM-Mixed (29)
SMO	46.80	75.00	76.20	50.30	78.70	79.70
Logistic	74.80	76.20	77.30	70.10	80.90	80.60
J48	71.20	80.50	81.50	71.50	81.40	84.00
JRIP	73.20	80.30	82.70	71.30	79.90	84.50
Average	*66.50*	*78.00*	***79.43***	*65.80*	*80.23*	***82.20***
# Bests	*0*	*0*	***4***	*0*	*1*	***3***

From Table 2 (a, b and c), it can be observed that when applying the original 22 features (as the benchmark), the recall of attrition was unacceptably low. When substituting the 7 RFM-derived descriptive features for the 22 benchmark features, the overall accuracy, recall of attrition and the precision of attrition of our customer "attrition *vs.* non-attrition" classification study were all increased significantly; when using the 29 mixed features rather than the 7 derived features, even better results were produced with respect to the average overall accuracy, average recall of attrition, average precision of attrition, and the number of best cases.

5 Conclusions

Customer attrition, especially for *middle-* and *high-valued* customers in retail banking, has become a more and more serious concern for commercial banks. To combat this attrition it is incumbent on banks to develop a prediction mechanism (such as a *binary* classifier) so as to identify customers who might be at risk of attrition. Based on this background, in this paper the features that are best suited to the identification of churn were considered. In earlier studies the use of 22 features was

proposed. In this paper an alternative set of 7 features is considered, derived using the RFM model. The classification effectiveness using the previously proposed 22 "benchmark" features, the proposed seven RFM features and a mixture of the two was compared using four different data sets (two balanced and two unbalanced) generated from a collection of *real* retail banking customer records. With respect to the reported comparison four different classifiers were considered. Regardless of which classification algorithm is adopted the results clearly demonstrate that our proposed seven feature representation significantly improves on the classification performance (with respect to the overall accuracy, recall of attrition and the precision of attrition) using the previously proposed 22 features. Even better results were produced when both sets of features were combined. Further research is suggested to produce an improved feature representation approach for our retail banking customer classification application.

Acknowledgement. The authors would like to thank Dr. Jiongyu Li and Yuzhi Guo from the China Minsheng Banking Corp., Ltd., Prof. Maoqing Zhou and Qian Gao from the Institute of Fiance and Banking at the Chinese Academy of Social Sciences, Prof. Yu Chen from the World New Economics Research Institute, and Greaman Wu from Beijing Bimu-cloud software Co.LTD for their support with respect to the work described here.

References

1. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A Training Algorithm for Optimal Margin Classifiers. In: Proceedings of the 5th ACM Annual Workshop on Computational Learning Theory, Pittsburgh, PA, USA, pp. 144–152 (1992)
2. Cohen, W.W.: Fast Effective Rule Induction. In: Proceedings of the 12th International Conference on Machine Learning, Tahoe City, CA, pp. 115–123 (1995)
3. Hand, D., Mannila, H., Smyth, P.: Principles of Data Mining. The MIT Press, Cambridge (2001)
4. Hughes, A.M.: Strategic Database Marketing. Probus Publishing, Chicago (1994)
5. Kandampully, J., Duddy, R.: Relationship Marketing: A Concept beyond Primary Relationship. Marketing Intelligence and Planning 17(7), 315–323 (1999)
6. Lei, J., Di, G., Coenen, F., Wang, Y.J.: A Hybrid LR/DT Classification Approach for Customer Attrition Risk Prediction in Retail Banking. In: Poster and Industry Proceedings of the 12th Industrial Conference on Data Mining, Berlin, Germany, pp. 95–100 (2012)
7. Li, F., Lei, J., Tian, Y., Punyapatthanakul, S., Wang, Y.J.: Model Selection Strategy for Customer Attrition Risk Prediction in Retail Banking. In: Proceedings of the 9th Australasian Conference on Data Mining, Ballarat, Austrilia, pp. 105–110 (2011)
8. Luck, D.: The Importance of Data within Contemporary CRM. In: Rahman, H. (ed.) The Book Data Mining Applications for Empowering Knowledge Societies, pp. 96–109. IGI Global, Hershey (2009)
9. Peng, C.-Y.J., So, T.-S.H.: Logistic Regression Analysis and Reporting: A Primer. Understanding Statistics 1(1), 31–70 (2002)
10. Quinlan, J.R.: Induction of Decision Trees. Machine Learning 1(1), 81–106 (1986)

11. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Francisco (1993)
12. Wang, W., Wang, Y.J., Xin, Q., Bañares-Alcántara, R., Coenen, F., Cui, Z.: A comparative study of associative classifiers in mesenchymal stem cell differentiation analysis. In: Kumar, A.V.S. (ed.) The Book Knowledge Discovery Practices and Emerging Applications of Data Mining: Trends and New Domains, pp. 223–243. IGI Global, Hershey (2011)
13. Wang, Y.J., Lei, J., Li, F.: A Case Study of Data Mining in Retail Banking: Predicting Attrition Risk for VIP Customers. In: Poster and Industry Proceedings of the 11th Industrial Conference on Data Mining, New York, USA, pp. 78–79 (2011)
14. Wei, J.-T., Lin, S.-Y., Wu, H.-H.: A Review of the Application of RFM Model. African Journal of Business Management 4(19), 4199–4206 (2010)
15. Welcker, L., Koch, S., Dellmann, F.: Improving Classifier Performance by Knowledge-Driven Data Preparation. In: Proceedings of the 12th Industrial Conference on Data Mining, Berlin, Germany, pp. 151–165 (2012)
16. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann Publishers, San Francisco (2000)
17. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann Publishers, San Francisco (2005)
18. Witten, I.H., Frank, E., Hall, M.A.: Data Mining: Practical Machine Learning Tools and Techniques, 3rd edn. Morgan Kaufmann Publishers, Burlington (2011)

An Evolutionary Method for Associative Local Distribution Rule Mining

Kaoru Shimada and Takashi Hanioka

Fukuoka Dental College, 2-15-1, Tamura, Sawara, Fukuoka, 814-0193, Japan
{shimada,haniokat}@college.fdcnet.ac.jp

Abstract. A method for rule mining for continuous value prediction has been proposed using a graph structure based evolutionary computation technique. The method extracts the rules named associative local distribution rule whose consequent part has a narrow distribution of continuous value. A set of associative local distribution rules is applied to the continuous value prediction. The experimental results showed that the method can bring us useful rules for the continuous value prediction. In addition, two cases of contrast rules are defined based on the associative local distribution rules. The performances of the contrast rule extraction were evaluated and the results showed that the proposed method has a potential to realize contrast analysis between two datasets.

Keywords: association rule, contrast rule, evolutionary computation, pattern recognition, regression.

1 Introduction

Association rule mining is the discovery of association relationships or correlations among a set of attributes in a database. Association rule in the form of '$X \rightarrow Y$ (if X then Y)' is interpreted as 'the set of attributes X are likely to satisfy the set of attributes Y'. The main advantage of the association rules is their interpretable form. When Y in the association rule is the class label, association rules are used to classification problems. Many techniques for associative classification techniques based on association rule mining have been proposed, which achieve quite effective performance [1,2,3]. Nowadays, interest for rule-based regression has increased, however, there is not much literature on the subject [4].

Recently, class association rule mining tools using a graph based evolutionary computation technique have been proposed and reported their applications [5,6]. The tools have been developed using a basic structure of Genetic Network Programming (GNP) and adopting a new evolutionary strategy to execute tasks accumulatively through generations. One of the advantages of GNP-based methods is flexible design of rule forms [7].

In this paper, we introduce an interesting rule named *associative local distribution rule*. In the form of associative local distribution rule, X is a combination of attributes and Y is the small localized distribution of continuous value. We

P. Perner (Ed.): ICDM 2013, LNAI 7987, pp. 239–253, 2013.

propose an extended GNP-based rule mining method for associative local distribution rules. The method discovers relations on the restricted local distribution and conditions given by combinations of attributes. An associative local distribution rule can work as a filter. Therefore, using a set of the rules, continuous value prediction can be executed. In this paper, a rule based prediction method for the continuous value is demonstrated. In addition, the rule mining method is extended to contrast rule mining to find interesting differences between two datasets [7,8].

Conventional Genetic Algorithm (GA) based methods extract a small number of rules optimizing a given fitness function [9]. On the other hand, in the proposed method, rules satisfying the given conditions are accumulated in a rule pool through GNP generations. GNP population evolves in order to store new interesting rules into the pool as many as possible, not to obtain the individual with highest fitness value. Therefore, the method is fundamentally different from the other evolutionary algorithms in its evolutionary way. The GNP-based method can quit the rule extraction anytime when enough number of rules, for example, for building a predictor are obtained.

This paper is organized as follows: In the next section, definitions on the associative local distribution rule and prediction methods using the rules are presented. In Section 3, an algorithm capable of finding local distribution rules is proposed. Experimental results for performance evaluations are presented in Section 4, and conclusions are given in Section 5.

2 Rules and Prediction Method

2.1 Associative Local Distribution Rule

Let A_k be an attribute in a database and its value be 1 or 0, and Y be the attribute with continuous values. Table 1 is an example of the database. In this paper, *associative local distribution rule* is defined as follows.

[Associative local distribution rule]

$$(A_j = 1) \wedge \cdots \wedge (A_k = 1) \rightarrow (m, s) \tag{1}$$

where, m and s is the mean value and standard deviation of Y of instances satisfying antecedent part of the rule, respectively. The rule is represented $X \rightarrow (m, s)$ in short, where, $X = (A_j = 1) \wedge \cdots \wedge (A_k = 1)$. X is represented briefly as $A_j \wedge \cdots \wedge A_k$.

If s is a small value satisfying given condition by users, the combination of attributes in the consequent part brings narrow distribution of the continuous value. For example, $(A_1 = 1) \wedge (A_2 = 1) \rightarrow (1995.73, 2.56)$ can be interpreted as "If A_1 and A_2 occur now, then the value of Y will be 1995.73 with standard deviation 2.56".

In this paper, frequency measurements of the associative local distribution rule r in the database is defined by

$$support(r) = \frac{N_X(r)}{N}, \tag{2}$$

Table 1. An example of database

ID	A_1	A_2	A_3	A_4	...	A_M	Y
1	1	1	0	1		1	1995
2	0	0	1	0		1	2010
3	0	1	0	0		0	1982
...	...	...	...	...	...	...	...
N	1	1	0	1		1	1996

where, N: total number of instances in the database,
$N_X(r)$: the number of instances satisfying the antecedent X of r,
$m(r)$: mean value of Y of the instances satisfying the antecedent of r,
$s(r)$: standard deviation of Y of the instances satisfying the antecedent of r.

Interestingness condition of rule r is defined using *support* and s as follows:

$$support(r) \geq sup_{min}, \tag{3}$$

$$s(r) \leq s_{max}, \tag{4}$$

where, sup_{min} and s_{max} are the threshold values of *support* and s given by users in advance, respectively.

2.2 Continuous Value Prediction

This subsection presents methods for building a predictor using a set of associative local distribution rules. Suppose that a test instance for prediction has the same data form as the training data. The followings are two methods for building the predictor using extracted rules.

[Input] A set of associative local distribution rules,
An instance to be predicted
[Output] $\hat{y}$: Predicted value for the instance

[First matched rule based method]
In this method, $\hat{y}$ is the mean value of the first matched rule in the predictor. If there is no rule that applies to the instance, no prediction of $\hat{y}$ is returned. Rules are sorted by following definition of precedence ($\succ$).
Given two rules, r_i and r_j, $r_i \succ r_j$ (also called r_i precedes r_j or r_i has a higher precedence than r_j) if
1) the $s(r_i)$ is smaller than $s(r_j)$, or
2) their s are the same, but the $support(r_i)$ is greater than $support(r_j)$, or
3) both the s and *support* of r_i and r_j are the same, but r_i is extracted earlier than r_j;

[Matched rule set based method]
1) Calculate the set R of suffixes of the rule whose antecedent part matches the new data.

2) When $|R| \neq 0$, compute $\hat{y}$ using all the rules in R as follows:

$$\hat{y} = \frac{\sum_{r \in R} m(r)}{|R|} \tag{5}$$

where, $m(r)$ is the mean value of the consequent part of r.
When $|R| = 0$, no prediction of $\hat{y}$ is returned.
3) In order to build a strict predictor, no prediction for $\hat{y}$ can be set in the case of $|R| \leq R_{min}$. R_{min} is the threshold number of rules given in advance. The method is an extension of the classification method described in [6].

2.3 Contrast Rules

In the case of data mining from the dense database, the contrasts between two data sets gathered by different conditions are interesting [7]. In this paper, the following combination of attributes (X) showing a contrast between two datasets are considered.

[Contrast rules]
Although $r_A : X \rightarrow (m_A, s_A)$ satisfies given importance conditions in Database A, however, the rule having the same antecedent part $r_B : X \rightarrow (m_B, s_B)$ does not satisfy the same conditions in Database B.

In this paper, following two cases of conditions are considered as definition of contrast rules.

[Case I: Difference between mean values is interesting]

$$|m_A - m_B| \geq d_{min}, \tag{6}$$

$$s_A \leq s_{max}, \quad s_B \leq s_{max} \tag{7}$$

where, d_{min} and s_{max} are the threshold values given by users in advance.

[Case II: Difference of distribution is interesting]

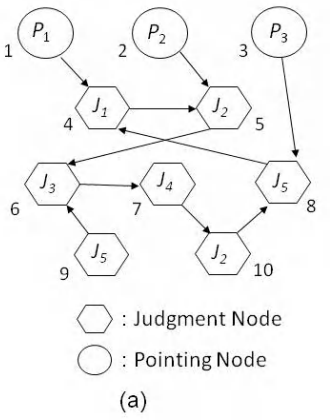
$$s_A \leq s_{max}, \quad s_B \geq s_{min} \tag{8}$$

where, s_{max} and s_{min} ($s_{min} > s_{max}$) are the threshold values given by users. In the both cases, the following conditions for *support* are also satisfied.

$$support(X \rightarrow (m_A, s_A)) \geq sup_{min}, \quad support(X \rightarrow (m_B, s_B)) \geq sup_{min} \tag{9}$$

3 Associative Local Distribution Rule Mining Method

In this section, methods for associative local distribution rule mining and contrast rule mining are proposed. In the proposed methods, the rules are extracted using evolutionary rule accumulation mechanism [6,7]. Candidate rules are symbolized by evolving network structure. Interpretation of rule representations and fitness function of GNP individual are designed based on the user's objects. Conditions of threshold values for interestingness are given by users in advance.



	NT_i	ID_i	C_i
1	P	1	4
2	P	2	5
3	P	3	8
4	J	1	5
5	J	2	6
6	J	3	7
7	J	4	10
8	J	5	4
9	J	5	6
10	J	2	8

(b)

Fig. 1. Basic structure of Genetic Network Programming individual for rule extraction

3.1 Candidate Rule Representation

Fig.1 shows a basic structure and genotype expression of GNP individual for rule mining. GNP is composed of two kinds of nodes: judgment node and pointing node (renamed processing node in [6,7]). Judgment nodes work as *if-then* type decision making functions and are the set of J_1, J_2, ..., J_p. Each judgment node has Yes-side connection and No-side connection. In Fig.1(a), connections for judgment nodes are simplified. On the other hand, pointing nodes are the set of P_1, P_2, ..., P_q. Each pointing node has an inherent numeric order and is connected to a judgment node. The practical roles of judgment nodes are predefined and stored in the library by supervisors. Once GNP is booted up, firstly the execution starts from P_1, secondly the next node to be executed is determined according to the connection from the current activated node. In Fig.1(b), NT_i describes the node type and ID_i is an identification number of judgment. C_i denote the node ID which are connected from node i.

Associative local distribution rules are represented as the connections of nodes in GNP individuals. Attributes and their values correspond to the functions of judgment nodes in GNP. Fig.2 shows a sample of the connection of nodes in GNP. '$A_1 = 1$', '$A_2 = 1$', '$A_3 = 1$' and '$A_4 = 1$' in Fig.2 denote the functions of judgment nodes. The connections of these nodes represent rules, for example, $(A_1 = 1) \wedge (A_2 = 1) \wedge (A_3 = 1) \wedge (A_4 = 1) \rightarrow (m_4, s_4)$. If this rule is interesting, then the rules symbolized by after changing the connections or functions of GNP nodes could be candidates of interesting ones. We can obtain these rule candidates effectively by GNP genetic operations, because mutation or crossover change the connections or contents of the nodes. In the next GNP generation after the operations, the candidates will be examined.

In the proposed method, combinations of judgment nodes work as filters and isolate the small distribution of the continuous values. In Fig.2, each judgment node separates the distribution of Y into two distributions. If the separated

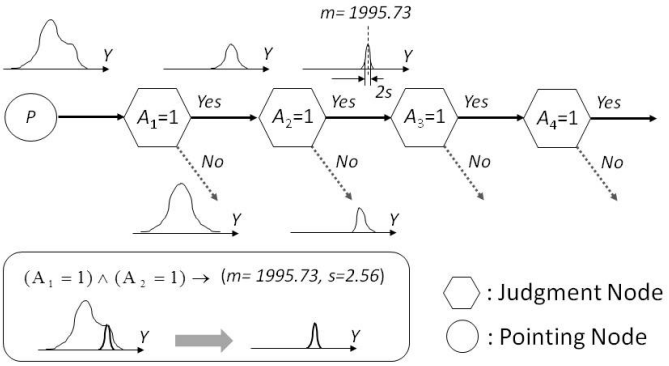


Fig. 2. Basic idea of associative local distribution rule extraction

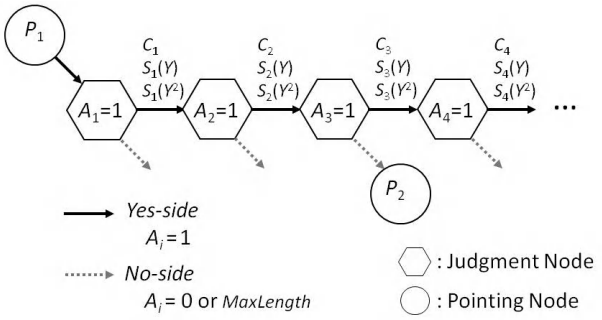


Fig. 3. An example of node connection in a GNP individual

distribution of Y satisfies the conditions of interesting local distribution rules, then the rule represented by the node connection is extracted.

The connections of nodes are represented as the antecedent part of associative local distribution rules. Fig.3 shows a sample of the connections of nodes in a GNP individual. P_1 is a pointing node and is a starting point of the antecedent part of rules. The connections of these nodes represent association rules, for example, $(A_1=1) \rightarrow (m_1, s_1)$, $(A_1=1) \wedge (A_2=1) \rightarrow (m_2, s_2)$, $(A_1=1) \wedge (A_2=1) \wedge (A_3=1) \rightarrow (m_3, s_3)$ and $(A_1=1) \wedge (A_2=1) \wedge (A_3=1) \wedge (A_4=1) \rightarrow (m_4, s_4)$.

Thus, node connections from the pointing node P_1 to each judgment node represent antecedent part of the rules. In this paper, we consider only the Yes-side connections of judgment nodes as rules. Yes-side of the judgment node is connected to another judgment node. Judgment nodes can be reused and shared with different rule representations because of GNP's features. No-side of the judgment node is connected to the next numbered pointing node. GNP individual generates many rule candidates using its graph structure. The kinds of the judgment node functions equal the number of attributes in the database.

Table 2. Measurements of associative local distribution rules

Associative local distribution rules	*support*	m	s^2
$A_1 \rightarrow (m_1, s_1)$	$\frac{C_1}{N}$	$\frac{S_1(Y)}{C_1}$	$\frac{S_1(Y^2)}{C_1} - m_1^2$
$A_1 \wedge A_2 \rightarrow (m_2, s_2)$	$\frac{C_2}{N}$	$\frac{S_2(Y)}{C_2}$	$\frac{S_2(Y^2)}{C_2} - m_2^2$
$A_1 \wedge A_2 \wedge A_3 \rightarrow (m_3, s_3)$	$\frac{C_3}{N}$	$\frac{S_3(Y)}{C_3}$	$\frac{S_3(Y^2)}{C_3} - m_3^2$
$A_1 \wedge A_2 \wedge A_3 \wedge A_4 \rightarrow (m_4, s_4)$	$\frac{C_4}{N}$	$\frac{S_4(Y)}{C_4}$	$\frac{S_4(Y^2)}{C_4} - m_4^2$

3.2 Calculation of Rule Measurements

GNP examines the attribute values of instances using judgment nodes. Judgment node determines the next node by a judgment result of Yes or No corresponding to Yes-side or No-side. For example, in Table 1, the instance $1 \in ID$ satisfies $A_1 = 1$, $A_2 = 1$ and $A_3 \neq 1$, therefore, the node transition from P_1 to P_2 occurs in Fig.3.

The total number of instances moving to Yes-side at each judgment node is calculated for every pointing node. The numbers are denoted as C_x in Fig.3 and these values are used as $N_X(r)$. In addition, when moving to Yes-side, the value of Y is examined in order to calculate m and s in the consequent of the rule. The sum of Y ($S_x(Y)$) and Y^2 ($S_x(Y^2)$) for the rules are stored and updated when the judgment result of instance moves to Yes-side.

Examinations of attribute values start from each pointing node. If the transition to Yes-side connection of judgment nodes continues and the number of the judgment nodes from the pointing node becomes a cutoff value (Maxlength: the maximum number of attributes in rules), then Yes-side connection is transferred to the next pointing node obligatorily.

When the examination of attribute values from the starting point P_s ends, then GNP examines the instance $2 \in ID$ from P_1 likewise. Thus, all the instances in the database are examined. In Fig.2, a sample distribution of Y of the instances moving to Yes-side at each judgment node is described.

When the examination of instances ends, then m and s of the candidate rules are calculated using the sum of Y and Y^2. The *support* is also calculated using C_x, that is, the total number of instances moving to Yes-side at each judgment node. Table 2 shows an example of the measurements of the rules obtained by the node connections in Fig.3.

3.3 Rule Extraction and Accumulation

In every GNP generation, the examinations are done from $1 \in ID$ and P_1 node. Examinations of attribute values start from each pointing node as described above. After all the instances in the database are examined, measurements of candidate rules of every pointing nodes are calculated and the interestingness of the rules are judged by given conditions like (3) and (4). The extracted interesting rules are stored in a rule pool all together through GNP generations.

When an associative local rule satisfying the conditions is extracted by GNP, the overlap of the attributes in the antecedent is checked and it is also checked whether the rule is new or not, i.e., whether it is in the pool or not.

3.4 Evolution of GNP

GNP individuals evolve in order to store new rules in the rule pool. Therefore, the fitness of GNP is set to express the potentiality for new rule extraction [6,7]. As the aim of the evolution of GNP is not to find an elite individual, the way of setting fitness functions is different from the case of conventional GA problems. In the experiments in the next section, the following fitness functions are used.

$$F_d = \sum_{r \in R_d} \{a \times support(r) + b/(s(r)+1) + n_X(r) + c_{new}(r)\} \tag{10}$$

The terms in (10) are as follows:
R_d: set of suffixes of extracted associative local distribution rules satisfying (3) and (4) in the GNP individual. $n_X(r)$: the number of attributes in the antecedent of rule r. a, b: constant. $c_{new}(r)$: additional constant defined by

$$c_{new}(r) = \begin{cases} c_{new} & \text{(rule } r \text{ is new)} \\ 0 & \text{(otherwise)} \end{cases} \tag{11}$$

Constants in (10) and (11) are set up empirically. $support(r)$, $s(r)$, $n_X(r)$ and $c_{new}(r)$ are concerned with the importance, complexity and novelty of rule r, respectively. $n_X(r)$ is also concerned with the evaluation of the good judgment node connections in the GNP individual.

All individuals in a population have the same number of nodes. The number of pointing nodes and judgment nodes in each GNP individual are determined by users. The connections of the nodes and the functions of the judgment nodes at an initial generation are determined randomly for each GNP individual. GNP individual needs not to include all the functions of judgment nodes and the number of each function is not fixed.

We use three kinds of genetic operators; crossover, mutation-1 (change the connection of nodes) and mutation-2 (change the function of judgment nodes) [6,7]. At each generation, individuals are replaced with new ones by a selection rule. The individuals are ranked by their fitnesses and upper 1/3 individuals are selected. After that, they are reproduced three times for the next generation, then three kinds of genetic operators are executed to them. These operators are executed for the gene of judgment nodes of GNP. All the connections of the pointing nodes are changed randomly in order to extract rules efficiently. The above number 1/3 is determined experimentally, which is not so sensitive to the results. Crossover operator affects two parent individuals. All the connections or contents of the uniformly selected corresponding nodes in two parents are swapped each other by crossover rate P_c. Mutation operator affects one individual. All the connections of each node are changed randomly by mutation

rate of P_m. $P_c = 1/5$, $P_{m1} = 1/6$ and $P_{m2} = 1/6$ is an effectual setting and was used in the experiments in Section 4. Information of the extracted rules like frequency of the appearances of attributes in the rules can be used for genetic operations to extract rules effectively [7,8].

3.5 Contrast Rule Mining

In Fig.3, C_x, $S_x(Y)$ and $S_x(Y^2)$ are the number of instances and sum of Y and Y^2 at each Yes-side of judgment node, respectively. When the database has a label for instance groups, we can calculate these values group by group at the same time. For example, when the database has two groups A and B, we can obtain values C_1^A, $S_1^A(Y)$, $S_1^A(Y^2)$, C_1^B, $S_1^B(Y)$ and $S_1^B(Y^2)$ at the same examination as in Table 2. Superscripts represent the two groups. Using these values, for example, measurements for contrast of $A_1 \rightarrow (m_1, s_1)$ between group groups A (r_a: $A_1 \rightarrow (m_{1A}, s_{1A})$) and B ($r_b$: $A_1 \rightarrow (m_{1B}, s_{1B})$) are obtained as follows:

$$support_A(A_1 \rightarrow (m_{1A}, s_{1A})) = \frac{C_1^A}{N^A}, \quad support_B(A_1 \rightarrow (m_{1B}, s_{1B})) = \frac{C_1^B}{N^B}$$

$$m_{1A} = \frac{S_1^A(Y)}{N^A}, \quad s_{1A}^2 = \frac{S_1^A(Y^2)}{C_1^A} - m_{1A}^2$$

$$m_{1B} = \frac{S_1^B(Y)}{N^B}, \quad s_{1B}^2 = \frac{S_1^B(Y^2)}{C_1^B} - m_{1B}^2$$

where, N^A and N^B are the number of instances of group A and B, respectively. When the candidate rule satisfy the conditions of contrast rules, the rule is accumulated in the contrast rule pool. The flow of contrast rule extraction is almost the same as associative local distribution rule mining.

GNP population evolves to discover contrast rules as many as possible. The fitnesses of GNP individual for contrast rule Case I and Case II described in **2.3** are defined as follows:

$$F_c^I = \sum_{r \in R_c} \{a \times (support_A(r_A) + support_B(r_B)) + b/(s_A(r_A) + 1)$$

$$+ b/(s_B(r_B) + 1) + n_X(r_A) + c_{new}(r_A)\} \tag{12}$$

where, R_c is the set of suffixes of extracted contrast rules satisfying (6), (7) and (9) in the GNP individual.

$$F_c^{II} = \sum_{r \in R_c} \{a \times support_A(r_A) + b/(s_A(r_A) + 1) + n_X(r_A) + c_{new}(r_A)\} \tag{13}$$

where, R_c is the set of suffixes of extracted contrast rules satisfying (8) and (9) in the GNP individual.

4 Experimental Results

4.1 Experimental Setting

The data set *YearPredictMSD* from UCI ML Repository [10,11] was used for the evaluation. The following is the abstract of the data set:
· Prediction of the release year of a song from audio features.
· 90 attributes ($T(j)$, ($j = 0, \ldots, 89$)) (12: timbre average, 78: timbre covariance)
· Target value (Y) is the year, ranging from 1922 to 2011.
· The instances are divided into 463715 instances for training and 51630 instances for testing.

Evaluations of the proposed methods were done on the following three points:

- Performance study of associative local distribution rule mining
- Evaluation of continuous value prediction using extracted rules
- Contrast rule mining between training data and testing data

The continuous attribute values ($T(j)$) are transformed to a set of attributes, whose value is 1 or 0. Discretization of the values is done as follows:
1) calculate the averaged value m_j and standard deviation s_j of $T(j)$ ($j = 0, \ldots, 89$), respectively.
2) discretize the value t_j of $T(j)$ using m_j and s_j and define attributes A_{3j}, A_{3j+1}, A_{3j+2} and their values as follows:

$$A_{3j} = \begin{cases} 1 & (t_j > m_j + s_j) \\ 0 & (\text{otherwise}) \end{cases}, A_{3j+1} = \begin{cases} 1 & (m_j - \frac{s_j}{2} < t_j < m_j + \frac{s_j}{2}) \\ 0 & (\text{otherwise}) \end{cases}$$

$$A_{3j+2} = \begin{cases} 1 & (t_j < m_j - s_j) \\ 0 & (\text{otherwise}) \end{cases}$$

The transformed *YearPredictMSD* dataset is complete and consists of 270 attributes and 1 continuous attribute for prediction. This transformation using above threshold has no scientific meaning, only intention was to make a dataset for the estimation use.

The GNP population size is 120. The number of pointing nodes and judgment nodes in each GNP individual are 10 and 100, respectively. The condition of termination is 100 generations. One trial of rule extraction by 100 generation is defined as 1 *round*. Extracted rules were stored in a pool for each *round*. After 500 (or 100 or 300) *rounds* trials, overlap of rules were checked and finally identified rules were obtained. $a = 20000$ and $b = 20$ in (10), $c_{new} = 30$ in (11), $a = 10000$ and $b = 10$ in (12) and (13) were used, respectively. These values were chosen by preliminary examination. The aim of the evolution is not to find an elite GNP individual. Therefore, settings of above constants for fitness function is not so strict for rule extractions. All algorithms were coded in C. Experiments were done on a 1.80GHz Intel(R) Core2 Duo CPU with 2GB RAM.

Table 3. Number of extracted associative local distribution rules. (90 Attributes, $sup_{min} = 0.0002$, $s_{max} = 4.0$)

Method	Identified rules (500 *rounds*)			Averaged value per *round*		
	Total	*support* ≥ 0.0003	$s \leq 3$	*Total*	*support* ≥ 0.0003	$s \leq 3$
Proposed Method	40828	16598	3170	4985.9	2726.4	585.3
Ring structure	39364	14798	2747	5000.0	2727.8	555.4
Random	19961	8700	829	156.7	90.2	4.3

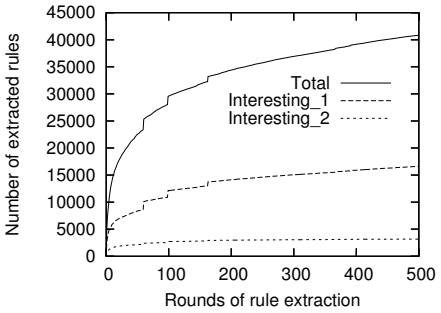


Fig. 4. Number of rounds for rule extraction vs. the number of extracted rules

The predictor was built up using the extracted rules. The predictor calculates $\hat{y}$ described in **2.2**. The prediction result J was defined as follows:

$$J = \begin{cases} success & (|y - \hat{y}| \leq d_{pr}) \\ failure & (\text{ otherwise }) \end{cases} \tag{14}$$

where, y is the real value. d_{pr} is the permissible range given in advance for the prediction. Therefore, d_{pr} is related to the accuracy of the prediction.

In this paper, accuracy (%) of the predictor is defined as

$$\frac{\text{Number of instances judged "}J = success\text{"}}{\text{Number of predicted instances}} * 100.$$

Cover rate (%) for the prediction is defined as

$$\frac{\text{Number of predicted instances}}{\text{Number of instances for the test}} * 100.$$

4.2 Performance Study of Rule Mining

First of all, completeness of associative local distribution rule extraction was estimated. As the GNP-based method cannot guarantee the extraction of all rules by the nature of evolutionary discovering policy, cover rate is an important factor of performance study. In addition, comparison of performance the proposed

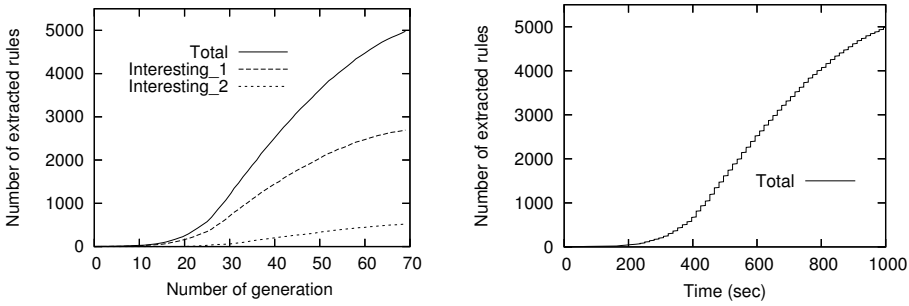


Fig. 5. An example of the number of extracted rules vs. GNP generation in one *round*.

method and the method without evolutionary mechanism by random generated GNP individuals. In this experiment, a part of the training dataset, that is, 90 transformed attributes $(A_1, \ldots, A_{90})$, attribute Y and 463715 instances was used. As the condition of rules, $sup_{min} = 0.0002$ in (3), $s_{max} = 4$ in (4) and $n_X(r) \leq 6$ were used. The condition $sup_{min} = 0.0002$ means that more than 93 instances in the training data should satisfy the rule. In each *round*, 5000 rules are the maximum number for rule accumulation. This means that the methods quit the rule extraction of one *round* when 5000 rules are extracted. Cover rate of rule extraction is estimated using the total number of identified extracted rules of independent 500 *rounds*.

Table 3 shows that the number of identified extracted associative local distribution rules of 500 *rounds* and the averaged value of extracted rules per *round*. *Ring structure* is the result using the same settings except the judgment node connection is restricted to make ring structure, that is, one ring form is composed using all the judgment nodes. *Random* is the result using the same settings except evolutionary mechanism of the proposed method. The connections and functions of judgment nodes were initialized every generation of GNP. Fig.4 shows that the number of identified accumulated rules versus the number of *rounds* for rule extraction. Interesting_1 and Interesting_2 in the Fig.4 are defined as the rules satisfying following additional conditions within the rule pool: $support \geq 0.0003$ and $s \leq 3$, respectively.

In the *Random* method, very small number of rules were extracted in each *round*. On the other hand, the proposed method extracted rules effectively by evolutionary mechanism. The ring structure method is behind the proposed method in total number of extracted rules. In the most of the *rounds*, 5000 rules were extracted around 70 generations in GNP evolution. It is found that almost the interesting rules can be extracted in a small number of combination of *rounds*. The results show that the rules with high measurement values tend to be extracted easily by the proposed method. Fig.5(a) shows a typical example of the rule extraction through GNP generation as one *round*. In the early generation, number of the extracted rule is very small. However, gradually the number of extracted rules increase by the evolutionary mechanism. Fig.5(b) shows the

Table 4. Prediction accuracy by First matched rule based method (270 Attributes)

s_{max} for used rules (#rule)	Cover rate (%)	Accuracy (%) Permissible ranges ($\leq d_{pr}$)			
		≤ 2.5	≤ 3.0	≤ 3.5	≤ 4.0
2.5 (1351)	2.37	60.07	70.29	77.33	80.69
3.0 (32253)	8.30	53.95	62.33	70.19	74.50
3.5 (188999)	21.68	47.79	56.03	63.61	69.69
4.0 (485159)	36.80	43.87	51.73	58.90	65.17

Table 5. Prediction accuracy by Matched rule set based method (270 Attributes)

(a) Number of Matched Rules ≥ 1 ($R_{min} = 0$)

s_{max} for used rules	Cover rate (%)	Accuracy (%) Permissible ranges ($\leq d_{pr}$)			
		≤ 2.5	≤ 3.0	≤ 3.5	≤ 4.0
2.5	2.37	59.82	69.15	77.58	80.85
3.0	8.30	52.93	61.56	70.05	75.39
3.5	21.68	46.62	55.36	62.93	70.03
4.0	36.80	42.44	50.29	57.65	64.45

(b) Number of Matched Rules ≥ 10 ($R_{min} = 9$)

s_{max} for used rules	Cover rate (%)	Accuracy (%) Permissible ranges ($\leq d_{pr}$)			
		≤ 2.5	≤ 3.0	≤ 3.5	≤ 4.0
2.5	0.65	66.17	73.95	82.34	84.43
3.0	3.88	57.36	66.83	75.71	80.20
3.5	9.13	51.45	60.19	68.04	75.33
4.0	22.06	45.63	54.00	61.59	68.54

run-time in the same experiment as Fig.5(a). Although the proposed method cannot extract all the rules meeting the given definition of importance, each trials extracts important rules from the large dataset. Run time of the GNP-based rule mining depends on the number of nodes in GNP individual and the number of extracted rules in the rule pool. Generally, the number of attributes in the data set is independent of calculation time of GNP-based method.

4.3 Evaluation of Continuous Value Prediction

Continuous value prediction using the extracted associative local distribution rules was evaluated. In this experiment, 270 transformed attributes, attribute Y and 463715 instances were used as training data. $sup_{min} = 0.0002$ in (3), $s_{max} = 4$ in (4) and $n_X(r) \leq 6$ were used. 51630 instances were used as test data for the evaluation. In each *round*, 5000 rules were the maximum number for rule accumulation. Evaluation was done using the total number of identified extracted rules of independent 300 *rounds*. Four predictors were built using the rules satisfying $2.5 \leq s_{max}$, $3.0 \leq s_{max}$, $3.5 \leq s_{max}$ and $4.0 \leq s_{max}$, respectively.

Table 6. Number of extracted contrast rules (Case I). ($sup_{min} = 0.0001$)

d_{min}	Identified rules (100 *rounds*)			Averaged value per *round*		
	$Total$	$m_A > m_B$	$(sup_A > 2 * sup_B)$ or $(2 * sup_A < sup_B)$	$Total$	$m_A > m_B$	$(sup_A > 2 * sup_B)$ or $(2 * sup_A < sup_B)$
0.5	66372	24727	1908	830.5	308.1	22.2
1.0	7309	2433	183	78.1	25.9	1.9
1.5	2120	697	38	21.8	7.1	0.4
2.0	843	326	17	8.6	3.3	0.2

Table 7. Number of extracted contrast rules (Case II). ($sup_{min} = 0.0001$)

	Identified rules (100 *rounds*)			Averaged value per *round*		
	$Total$	$\lvert m_A - m_B \rvert \leq 4$	$s_B \geq 12$	$Total$	$\lvert m_A - m_B \rvert \leq 4$	$s_B \geq 12$
$s_A \leq 4, \quad s_B \geq 4$	127997	118409	859	2134.8	1991.0	9.9
$s_A \leq 4, \quad s_B \geq 6$	13739	10619	550	172.3	133.5	5.9
$s_A \leq 4, \quad s_B \geq 8$	2357	1231	429	25.4	13.4	4.4

The total number of extracted associative local distribution rules was 485159 at 300 *rounds*. Table 4 shows that the prediction accuracy by First matched rule based method. It was found that the set of associative local distribution rules can predict continuous value. The number of rules satisfying strict condition like $2.5 \leq s_{max}$ was very small and cover rate for prediction was a low level. The trade-off for more cover rate is lower accuracy. Permissible range for prediction can be set as the same level as s_{max} in the rule extraction. Table 5 shows that the accuracy by Matched rule set based method. The condition $|R| \leq R_{min}$ improved the accuracy, however, the threshold R_{min} depends on problems.

4.4 Contrast Rule Mining between Two Labeled Data

In the evaluation of prediction, the cover rate and accuracy were not so high level. In order to analyze the difference between training data (Database A) and testing data (Database B), contrast rule extractions were done. In this experiment, more sensitive settings $sup_{min} = 0.0001$, $s_{max} = 4$ and $n_X(r) \leq 6$ were used. As the parameter for contrast rule Case I, $d_{min} = 0.5, 1.0, 1.5$ and 2.0 were used. As the parameter for Case II, $(s_{max}, s_{min}) = (4, 4), (4, 6)$ and $(4, 8)$ were used.

Table 6 shows that the number of contrast rules for Case I. $Total$ shows the number of identified rules at 100 *rounds*. "$m_A \geq m_B$" and "$sup_A > 2 * sup_B$ or $2 * sup_A < sup_B$" in Table 6 are defined as the rules satisfying additional conditions within the rule pool, respectively. Table 7 shows that the number of contrast rules for Case II. "$|m_A - m_B| \leq 4$" and "$s_B > 12$" in Table 7 are defined as the rules satisfying additional conditions within the rule pool, respectively. It was found that the many contrast rules exist between the training data and test data. This is one of the reason that using the strict condition $|R| \leq R_{min}$ is better in the prediction experiment.

5 Conclusions

A new rule mining method for continuous value prediction has been proposed using a graph structure based evolutionary method. The method extracts associative local distribution rules whose consequent parts have narrow distributions of continuous values. A set of associative local distribution rules is applied to the continuous value prediction. The performances of the rule extraction of the proposed method were evaluated using a large data set. The experimental results showed that the proposed method can bring us useful rules for the continuous value prediction. Users can define the conditions of rules for prediction flexibly. In addition, two cases of contrast rules are defined based on the associative local distribution rules. The performances of the contrast rule mining have been evaluated using a large number of instances. The results showed that the proposed method has a potential to realize contrast analysis for the datasets. We are studying applications of the proposed methods to the medical datasets.

Acknowledgment. This work was partly supported by JSPS KAKENHI Grant Number 24500191.

References

1. Liu, B., Hsu, W., Ma, Y.: Integrating Classification and Association Rule Mining. In: Proc. of the ACM Int'l Conf. on Knowledge Discovery and Data Mining, pp. 80–86 (1998)
2. Li, W., Han, J., Pei, J.: CMAR: Accurate and efficient classification based on multiple class-association rules. In: Proc. of the 2001 IEEE Int'l Conf. on Data Mining, pp. 369–376 (2001)
3. Li, J., Dong, G., Ramamohanarao, K., Wong, L.: DEEPS: A new instance-based lazy discovery and classification system. Machine Learning 54(2) (2004)
4. Janssen, F., Furnkranz, J.: Heuristic Rule-Based Regression via Dynamic Reduction to Classification. In: Proc. of International Joint Conferences on Artificial Intelligence 2011, pp. 1330–1335 (2011)
5. Mabu, S., Chen, C., Lu, N., Shimada, K., Hirasawa, K.: An Intrusion-Detection Model Based on Fuzzy Class-Association-Rule Mining Using Genetic Network Programming. IEEE Trans. on Syst., Man, and Cyber.-Part C 41, 130–139 (2011)
6. Shimada, K.: An Evolving Associative Classifier for Incomplete Database. In: Perner, P. (ed.) ICDM 2012. LNCS (LNAI), vol. 7377, pp. 136–150. Springer, Heidelberg (2012)
7. Shimada, K., Hirasawa, K.: Exceptional Association Rule Mining Using Genetic Network Programming. In: Proc. of the 4th International Conference on Data Mining, pp. 277–283 (2008)
8. Shimada, K., Hirasawa, K.: A Method of Association Rule Analysis for Incomplete Database Using Genetic Network Programming. In: Proc. of the ACM Genetic and Evolutionary Computation Conference 2010, pp. 1115–1122 (2010)
9. Ghosh, A., Jain, L.C. (eds.): Evolutionary Computing in Data Mining. STUDFUZZ, vol. 163. Springer, Heidelberg (2005)
10. Blake, C., Merz, C.: UCI repository of machine learning databases, `http://www.ics.uci.edu/~mlearn/MLRepository.html`
11. Bertin-Mahieux, T., Ellis, D.P.W., Whitman, B., Lamere, P.: The Million Song Dataset. In: Proc. of the 12th Int'l Society for Music Inf. Retrieval Conf. (2011)

Application of Data Mining Techniques on EMG Registers of Hemiplegic Patients

Ana Aguilera[1], Alberto Subero[1], and Ramón Mata-Toledo[2]

[1] Center of Analysis, Modelling and Treatment of Data, CAMYTD,
Facultad de Ciencias y Tecnología, Universidad de Carabobo, Valencia, Venezuela
{aaguilef,asubero}@uc.edu.ve
[2] Computer Science Department, James Madison University, Harrisonburg, Virginia, USA
matatora@jmu.edu

Abstract. Gait analysis provides a very large data volume coming from kinematic, kinetic, electromyographic (EMG) registers and physical examinations. The analysis and treatment of these data is difficult and time consuming. This work applies and explores exhaustively different analysis methods from data mining on these gait data. This study aims to provide a classification system based in gait patterns obtained from EMG records in children with spastic hemiplegia. The methods studied from data mining specifically for the classification task include SVM, neural networks, decision trees, regression logistic models and others. Different techniques of feature extraction and selection have been also employed and combined with classifications methods. The LMT algorithm provides the best result with 97% of instances classified correctly taking into account the indicators for 2 legs. A qualitative and quantitative validation were performed on the data.

Keywords: Gait Classifications, Data mining, EMG, Spastic Hemiplegia.

1 Introduction

Cerebral palsy is the result of a lesion or anomaly of the brain. Hemiplegia is the outcome of a lesion in or near the motor zone of the cerebral cortex. Spastic hemiplegia (SH) is a type of cerebral palsy characterized by muscular rigidity (paralysis) on one side of the body affecting the motor functions of the upper and lower limbs. SH produces neuron muscular communication failures that affect the efficient displacement of the person. In consequence, people affected by SH walk with a particular gait in which "the legs are held together and move in a stiff manner where the toes seem to drag and catch" [1]. This style of walking is due to a prolonged unilateral muscular contraction.

The loss of displacement efficiency produced by SH can be quantified by studying the energy consumption of the individual's gait seen through the electrical activity of the muscles. The Hospital Ortopédico Infantil (Orthopedic Children Hospital) in

P. Perner (Ed.): ICDM 2013, LNAI 7987, pp. 254–265, 2013.

Caracas (Venezuela), since 1977 has been conducting clinical analysis of gaits. This organization, within its Servicio de Diagnóstico Auxiliar (Diagnostic Auxiliary Services) and, at the only laboratory of its kind in the entire country, takes care of patients with walking difficulties as a result of SH, cerebral palsy, spina bifida, or some other related medical conditions. The lab uses a diagnostic tool that determines the physiological conditions of the patients who need rehabilitation to improve their walking. The main objective of gait analysis is to aid the clinicians in not only making the correct diagnosis but also choosing the most appropriate treatment strategy [2]. Gait analysis is now recognized as an useful clinical tool for the diagnosis and treatment of different types of cerebral palsy. However, in spite of its usefulness, gait analysis may be hampered by several factors such as large volume of data, the time and cost associated with the studies and their interpretations [3]. Gait analysis generates large volume of data because studying how the body moves while walking is a complex task which requires input from different nature and sources. Gait analysis can be considered a dynamical process that intends to discern a pattern from the continuous interaction of body and limbs while conducting a forward movement toward the center of gravity with minimal energy consumption. People affected by hemiplegia, while walking, have higher energy consumption than people with normal gait. The data used in this work was collected by clinicians of the Gait Laboratory of the Orthopedic Children Hospital between the years 1998 and 2004.

Several artificial intelligence and statistical approaches and techniques for the data evaluation of gait analysis have been used in recent years. A review of the former includes fuzzy systems, multivariate statistical techniques and fractal dynamics [4]. An additional analysis system, called adaptive neuro-fuzzy inference system, which combines neural network capabilities and fuzzy logic qualitative approaches was used to address the clinical problem of pes cavus and pes planus according to Xu [5]. Vector machines (SVM) have been used also to explore automated detection and classification of children with cerebral palsy using, as input features, two basic temporal-spatial gait parameters (stride length and cadence) as indicated by Kamruzzaman [6]. Statistical techniques such as the ANOVA test has been used in gait analysis and surface electromyograms of seven major lower limb muscles which were assessed for 5 years following multilevel surgery. A two-way ANOVA has considered the effect of the following factors: DIAGNOSIS (levels: diplegia and hemiplegia), TIME for repeated measurements (levels: examination before and after the surgery) and, the interaction between these factors for the norm-distance of the spatiotemporal, kinematic, kinetic and EMG parameters [7].

Works focusing on the discriminant feature extraction of gait parameters include principal component analysis (PCA) and hybrid feature reduction method based on the combination of feature ranking with PCA as described by Ming-Jing [8]. In this latter case, two gait analysis problems were considered using three feature reduction methods, namely, feature ranking based the value of signal to noise ratio (MSNR), PCA, and the proposed hybrid approach (MSNR & PCA). The first gait analysis problem was to differentiate the patients with neurodegenerative disease from the

controls based on the gait data collected by footswitches. The second gait analysis problem was to discriminate the patients with complex regional pain syndrome (CRPS) from controls based on the gait data collected by an accelerometer. An additional work used a kernel principal component analysis algorithm to extract gait features for initiating the training set of SVM via pre-processing [9]. A Bayesian classifier model was used to investigate the significance of two basic temporal-spatial gait parameters (stride length and cadence) using a normal and healthy group of 68 individuals versus a group of 88 individuals with spastic diplegia form of cerebral palsy [10]. Wolf [11] proposed the use of a methodological modular framework for automated assessment of gait patterns using different mathematical methods. Their application was based on the clinical problem of Botulinum Toxin A treatment of the spastic equinus foot. A set of 3670 parameters was ranked by relevance for classification of a group of 42 diplegic cerebral palsy patients.

Although there are plethora of useful and significant approaches for gait analysis, in this paper, the authors will concentrate in the use of several data mining techniques for the classification task. What makes data mining useful for gait analysis is its applicability for the extraction of hidden predictive information from large databases [3]. Gait analysis generates large volume of data because, as it was indicated before, gathers data of different natures and sources such as kinematic and kinetic through the readings of electromyographic (EMG) registers, heart's electrical activity through a EKGs, and corporal information through physical examinations and direct observations by clinicians. These tests allow the physicians to evaluate the function of the muscles and the position of the joints during gait while providing a more extensive and objective assessment of the neural deficit in patients who have spastic disorders [12].

Though different data mining techniques have been applied to extract information from each of these sources, in this paper, the authors, for simplicity and conciseness, have focused only on the study of data obtained through EMG registers.

2 Material and Methods

2.1 Subjects

The data corresponding to the records used in this work were collected by Gait Laboratory of Orthopedic Children´s Hospital [13] between 1998 and 2004. The data was acquired using a VICON © acquisition system (with twelve-bit transmission and a sampling frequency of 1.5 kHz. Signals were converted into an analog signals in the range of ± 10V). These data correspond to thirty SH patients with 99 EMG signals normalized and delimited to a gait cycle were considered in C3D format. These data contain registers with a diagnosis from medical staff of the laboratory according to the Gage's classification [12]. The distribution of cases is 51 for Group I, 16 for the Group II, 6 for the group III, 23 for the Group IV and 3 without classification.

2.2 Methods

Gage and his team have suggested that SH could be classified into at least four groups: group I, group II, group III and group IV taking into account the kinematic pattern of cerebral palsy patients [12], [14]. This classification for the four Groups of SH and the normal pattern was done in the sagittal plane kinematic patterns (Fig. 1). The kinematic classification goes to involve progressively the other joints such as the ankle (group I), knee (group II), hip (group III) and pelvis (group IV). The groups were characterized as follow: group I) drop foot, plantar flexion of the ankle in the swing phase, hip into increased flexion and lordosis; group II) group I characteristics and also plantar flexion of the ankle during the stance and swing phases, hyperextension of knee; group III) group II characteristics and also decreased knee movement; group IV) group III characteristics and also decreased hip movement and increased lordosis [12], [15].

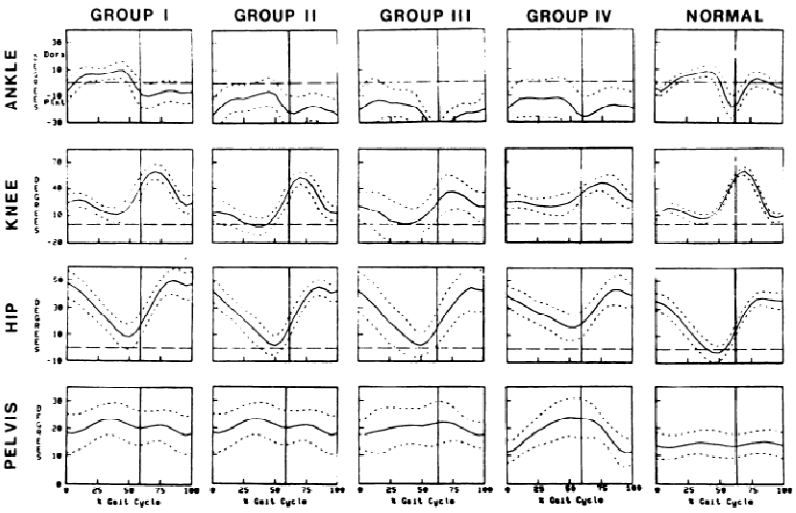


Fig. 1. Spastic Hemiplegia Clasiffication (Source : Winters, 1987)

Twenty-tree indicators were obtained from studying EMG records using Viloria's contributions [16] (Table 1). This study was conducted following the kinematcis classification from an electromyographical point of view, based on time and frequency domains. The twenty-tree indicators are : 2 indicators of instant energy, 3 obtained by fast Fourier transform (FFT), 16 from the computation of mean power frequency (MPF) and, others resulting from the energy spectrum of each component derived from the wavelet decomposition of the normalized EMG. The statistic behaviors for each indicator were determined computing the mean and standard deviation. The orders of magnitude from these indicators in time and frequency domain were obtained using computation of instant energy, significant frequencies and spectrum power [16]. The computed indicators for EMG including the two legs were considered. A total of 277 attributes have been selected and they correspond to 23 indicators x 6 muscle x 2 legs plus 1 for group type.

Table 1. Indicators to EMG signal (Source: Viloria, 2003)

INDICATOR	DESCRIPTION
BW	Bandwidth
MPF	Mean power frequency
ET	Spectrum power
MPFF	MPF from IDWT reconstruction (Inverse Discrete Wavelet Transform)
MPF1	MPF first component
MPF2	MPF second component
MPF3	MPF third component
MPF4	MPF fourth component
MPF5	MPF fifth component
MPF6	MPF sixth component
MPFLP	MPF lower pass component
PSD	Energy PSD
E1	Energy first component
E2	Energy second component
E3	Energy third component
E4	Energy fourth component
E5	Energy fifth component
E6	Energy sixth component
ELP	Energy lower pass component
ANL	Energy nonlinear
ANE	Accumulated energy nonlinear
EF	Spectrum power from IDWT reconstruction

Table 2. Different classifiers used to EMG signal

Classifier	Number of Correctly Classified Instances	Number of Incorrectly Classified Instances	% Correctly Classified
LMT	69	2	97,18%
FT	68	3	95,77%
RF	64	7	90,14%
NBTree	63	8	88,73%
LADTree	57	14	80,28%
BFTree	56	15	78,87%
SC	56	15	78,87%
J48	55	16	77,46%
REPTree	52	19	73,24%
J48graf	50	21	70,42%
RT	48	23	67,61%
DS	46	25	64,79%

Different data mining algorithms were used for classification purposes. The following classifiers were used for testing purposes: a best-first decision tree classifier (BFTree), DecisionStump, functional trees (FT), J48, a grafted (pruned or unpruned) C4.5 decision tree (J48graft), a multi-class alternating decision tree (LADTree), Logistic Model Tree (LMT), A Naïve Bayes/Decision tree (NBTree), RandomForest, RandomTree, Fast decision tree learner (REPTree) and SimpleCart. Table 2 shows the results of classification for 71 instances studied. Percentages represent the correctly classified the instances for each algorithm.

LMT = logistic model tree (LMT) [17] is an algorithm for supervised learning tasks that is combined with linear logistic regression and tree induction.
FT = functional trees is a tree learner of Jaoa Gama [18] that incorporates oblique splits and functions at the leaves.
RF = Random forest (or random forests) are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [19]. The algorithm for inducing a random forest was developed by Leo Breiman. It is able to classify big quantities of data with great accuracy.
NBTree = A Naïve Bayes/Decision tree. It is a decision tree with naive Bayes classifiers at the leaves.
LADTree = a multi-class alternating decision tree using the LogitBoost strategy [20].
BFTree = a best-first decision tree classifier [21]. This algorithm uses binary split for both nominal and numeric attributes. For missing values, the method of fractional'instances is used.
SC = SimpleCart is a decision tree learner that implements minimal cost-complexity pruning [22].
J48 = J48 algorithm is the Weka implementation of the C4.5 top-down decision tree learner proposed by Quinlan [23]. The algorithm uses the greedy technique and is a variant of ID3, which determines at each step the most predictive attribute, and splits a node based on this attribute. Each node represents a decision point over the value of some attribute. J48 attempts to account for noise and missing data. It also deals with numeric attributes by determining where thresholds for decision splits should be placed. The main parameters that can be set for this algorithm are the confidence threshold, the minimum number of instances per leaf and the number of folds for reduced error pruning.
REPTree = Fast decision tree learner. Builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning (with backfitting). Only sorts values for numeric attributes once. Missing values are dealt with by splitting the corresponding instances into pieces (i.e. as in C4.5).
J48graf = a grafted (pruned or unpruned) C4.5 decision tree [24].
RT = RandomTree builds a tree that considers K randomly chosen attributes at each node. Performs no pruning.
DS = DecisionStump [25] is a single-level decision tree with a categorical or numeric class label.

3 Results

The LMT algorithm [17] provides the best result with 97% of instances classified correctly taking into account the indicators for 2 legs. This algorithm was able to successfully classify a total of 96 instances with only three failures representing 96.96% of success. Of the 96 instances classified, 51 were classified as Group 1, 16 as Group 2, 6 as Group 3 and 23 instances as Group 4. It was also found that the 71 trial cases for the training model are well adjusted to the classification obtained from

kinematics. The feature selection for EMG has not improved the classification results. The LMT corresponds to a tree with only one leave as follows:

Number of Leaves : 1 Size of the Tree : 1

LM_1:

Group 1 : 2.9 + [MPFLP_RA_DE] * 0.12 + [MPF3_RA_DE] * 0.03 + [EF_RA_DE] * 0.17 + [E5_VL_DE] * -56.14 + [E4_VL_DE] * -12.82 + [E4_MH_DE] * 1.02 + [MPFF_S_DE] * 0 + [EF_S_DE] * 0.07 + [EPCF_RA_IZ] * -0.79 +[PSD_VL_IZ] * -0.17 + [E4_VL_IZ] * -2.82 + [EF_VL_IZ] * -0.35 + [ET_TA_IZ] * -0.08 + [AE_TA_IZ] * -0.11 + [ANE_TA_IZ] * -0.06 + [E3_G_IZ] * 0.04 + [MPF6_S_IZ] * -0.08 + [E4_S_IZ] * -0.3

Group 2 : 12.28 + [MPF3_RA_DE] * -0.04 + [BW_MH_DE] * -0.08 + [EF_MH_DE] * -0.07 + [ET_TA_DE] * -0.29 + [E3_RA_IZ] * 2.69 + [EF_RA_IZ] * 0.75 + [MPF_VL_IZ] * -0.01 + [E2_VL_IZ] * 2.07 + [EF_VL_IZ] * 0.79 + [BW_MH_IZ] * -0.01 + [ELP_MH_IZ] * 399.16 + [MPF_TA_IZ] * -0.02 + [E1_G_IZ] * -0.14 + [MPFF_S_IZ] * -0.01 + [E4_S_IZ] * 0.57

Group 3: -22.15 + [MPFF_RA_DE] * 0.01 + [BW_MH_DE] * -0.19 + [MPFF_S_DE] * 0.01 + [BW_RA_IZ] * 0.04 + [BW_TA_IZ] * -0.08 + [MPF_G_IZ] * 0.05 + [MPFF_G_IZ] * 0.03

Group 4: -18.35 + [MPFLP_RA_DE] * -0.78 + [MPF2_RA_DE] * 0.03 + [PSD_VL_DE] * 0.39 + [E5_VL_DE] * 153.29 + [E4_VL_DE] * 5.52 + [E3_VL_DE] * 3.07 + [E5_TA_DE] * 39.4 + [ET_TA_DE] * 0.4 + [MPFLP_VL_IZ] * 0.14 +[BW_TA_IZ] * 0.02 + [E6_G_IZ] * 1.32 + [MPFLP_S_IZ] * -0.29 + [MPF4_S_IZ] * 0.07

In this work the authors used two types of validation for evaluating the effectiveness of the models: a) A qualitative validation and verification directly from the experts and b) other quantitative validation using measures of validity and reliability described before.

3.1 Qualitative Validation

A group of patients with pathology of SH treated in hospital were selected randomly. Two-dimensional videos in two planes (sagittal and frontal) of each one of these patients and their respective EMGs were presented to specialist physicians. Afterwards, the results issued by each automatic classifier were compared with those diagnoses generated by physicians. A total of 3 physicians were consulted, 5 patients were selected, each one with 3 or 4 Trials. Table 3 shows the trials of patients which were selected with the results issued by classifier and their physicians' diagnosis.

Table 3. Comparison between automatic classification and expert classification

Trials	Medical diagnosis	Aprox. EMG (%)				
		Group of SH (%)				Classif
		I	II	III	IV	
50404, 50405, 50406, 50408	Group 4	17,5	0,03	0	100	Group 4
62008, 62010, 62015	Group 4	3,92	0	2,46	54,43	Group 4
53407, 53408	Group 3	95,67	2,5	100	53,51	Group 3
51403, 51407, 51408	Group 1	99.89	0.01	0	0,09	Group 1
38605	Group 2	0,03	99,96	0	96,81	Group 2

The column for the automatic classifiers has been divided into two sub-columns. One of them shows the percentages obtained by each type of SH, and the other subcolumn shows the final classification of each classifier by patient. These results have been also compared with the results of kinetics and kinematics classifiers. The results obtained with EMG records have a success of 100%.

3.2 Quantitative Validation

The validity and accuracy of a model is determined, first by the degree in which the model classifies what it has to classify and, second by the absence of systematic errors [26]. Reliability is the degree of stability achieved when the test is repeated under similar conditions. The measures commonly used to measure the reliability are: the global concordance index, the Kappa concordance index, coefficient of variation and interclass correlation coefficient. The validity of the model is also calculated using sensitivity and specificity measures. These measures involve combinations of variables such as true positive, true negatives, false positives and false negatives. Also, the positive verisimilitude rate (LR+) and negative verisimilitude rate (LR-) are other measures used for determining validity of the model. ROC curves (Receiver Operating Characteristic) are a very useful visual tools when comparing two models of classification. They show the comparison between the true positive rate and false positive rate for a model. Perfect accuracy will have an area of 1.0.

Table 4 shows the numerical results of the LMT algorithm. 71 instances of the total (99) were used for training of model. This model classified 69 instances correctly. Test mode used was the Cross-Validation with 10 Folds. From the confusion matrix it is possible to observe that only for group 1 and 4, 1 instance was classified incorrectly in each group (Table 5). The table 6 presents the measures of sensitivity (S), specificity (E), positive verisimilitude (LR+), negative verisimilitude (LR-) and ROC curves.

Table 4. LMT algorithm results

Total Number of Instances 71	EMG	Percentage
Correctly Classified Instances	69	97.1831 %
Incorrectly Classified Instances	2	2.8169 %
Kappa statistic	0.9569	
Mean absolute error	0.0459	
Root mean squared error	0.1468	
Relative absolute error	13.8763 %	
Root relative squared error	36.1886 %	

Table 5. Confusion Matrix

```
a  b  c  d   <-- classified as
33  1  0  0 |  a = Group 1
 0 16  0  0 |  b = Group 2
 0  0  3  0 |  c = Group 3
 1  0  0 17 |  d = Group 4
```

Table 6. Quantitative measures for LMT

CLASS		Group 1	Group 2	Group 3	Group 4
Classifier LMT-277	S	0.971	1	1	0.944
	E	0.971	0.941	1	1
	LR+	0.971	1	1	0.944
	LR-	0.027	0.018	0	0
	ROC	0.982	0.988	1	0.996

4 Application to Aid Decision

The LMT was implemented and integrated in a computer stand- alone application made in java The purpose of this application is to aid physicians in making their diagnosis. The application used the patterns learned from data analysis studies to approximate a classification of patients into the four group of SH. Figure 1 shows the basic data of patient and his/her video and at the bottom the classification using the patterns. At the left, the application shows classifications based on kinetic and kinematics signals and at the right, classifications based on EMG signals. Figure 3 presents a view of EMG signal for the twelve muscle studied.

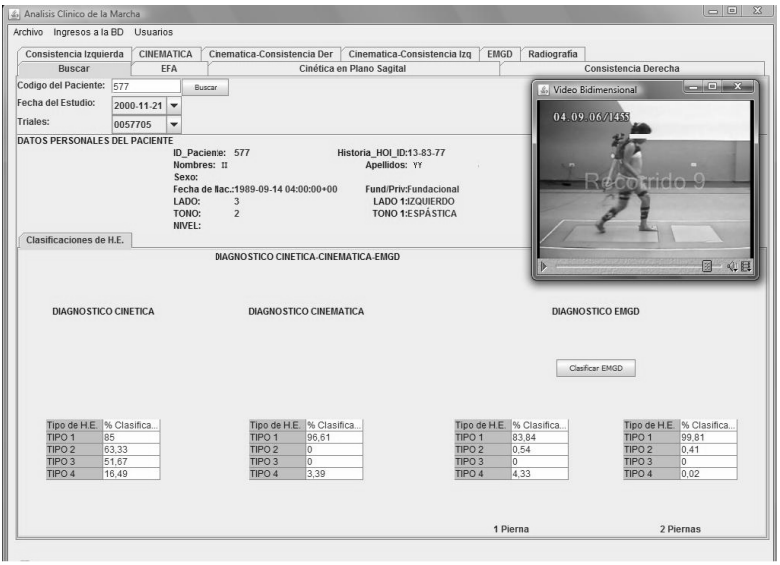


Fig. 2. Application integrating the LMT model for EMG signals

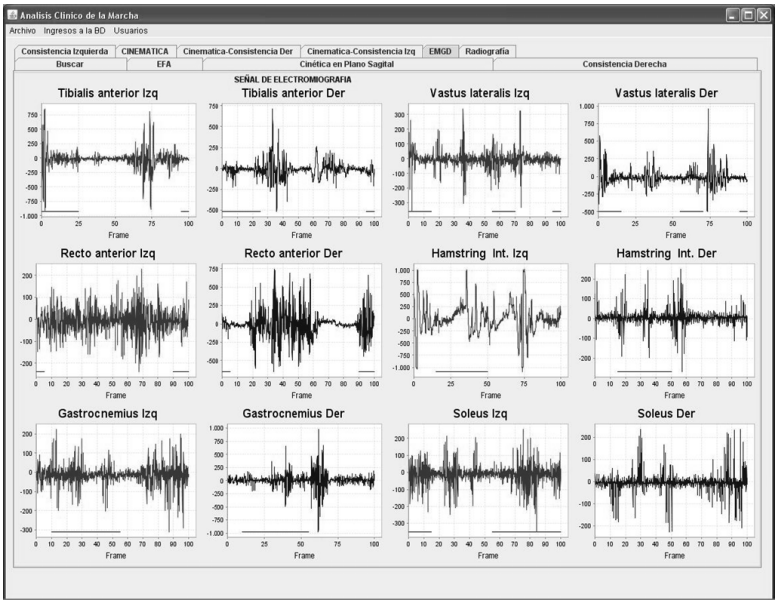


Fig. 3. View of EMG signals from the application

5 Discussion

Gait analysis is a useful clinical tool for the diagnosis and treatment of different types of cerebral palsy. However, this analysis generates large volume of data from

different natures and sources such as kinematic , kinetic and electromyographic (EMG) registers, heart's electrical activity through a EKGs, and corporal information through physical examinations and direct observations by clinicians. This data is difficult to treat directly or manually to obtain useful patterns of data behavior to aid physicians in their habitual task. For this reason, the authors have explored a variety of data mining techniques, particularly, the classification task, to extract useful knowledge about spastic hemiplegia diagnosis. In this work, the authors have focused on EMG records analysis with Twenty-tree indicators as defined by Viloria's contribution. Two legs were considered for the study. It is assumed that the unaffected side limb (known as contralateral limb) compensates gait deviations due to the abnormal pattern of the ipsilateral limb. But when considering human gait the behavior of both limbs is highly correlated so analysis of the contralateral side should prove useful [27].

The LMT algorithm provided the best result with 97% of instances classified correctly taking into account the indicators for the 2 legs. This algorithm was able to successfully classify a total of 96 instances with only three failures representing 96.96% of success. These results were validated from a qualitative and quantitative point of view; the obtained results are very promising and worth pursuing in a future study. The LMT algorithm was also included in a user-end application with the results of kinetics and kinematics analysis, similar to the present work.

In the future, the authors will consider other pathologies which will be analyzed also using other data mining techniques with a particular interest in the clustering analysis with unsupervised methods [28].

Acknowledgments. The authors appreciate the financial support of Fulbright program at James Madison University, USA and the CDCH of Carabobo University, Venezuela.

References

1. Farlex: The Free Dictionary by Farlex, http://medical-dictionary.thefreedictionary.com/gait
2. Loose, T., Mikut, R., Malberg, H., Simon, J., Schablowski, M., Rupp, R., Döderlein, L.: A Computer Based Method to Assess Gait Data. In: 2nd European Medical and Biological Engineering Conference, EMBEC (2002)
3. Simon, S.: Quantification of human motion: gait analysis—benefits and limitations to its application to clinical problems. J. of Biomechanics 37(12), 1869–1880 (2004)
4. Chau, T.: A review of analytical techniques for gait data. Part 1: fuzzy, statistical and fractal methods. Gait & Posture 13(1), 49–66 (2001)
5. Xu, S., Zhou, X., Sun, Y.: A novel gait analysis system based on adaptive neuro-fuzzy inference system. J. Expert Systems with Applications 37(2), 1265–1269 (2010)
6. Kamruzzaman, J., Begg, R.K.: Support Vector Machines and Other Pattern Recognition Approaches to the Diagnosis of Cerebral Palsy Gait. IEEE Transactions on Biomedical Engineering 53(12), 2479–2490 (2006)
7. Patikas, D., Wolf, S., Schuster, W., Armbrust, P., Dreher, T., Döderlein, L.: Electromyographic patterns in children with cerebral palsy: Do they change after surgery? Gait & Posture 26, 362–371 (2007)

8. Yang, M.-J., Zheng, H.-R., Wang, H.-Y., Mcclean, S., Harris, N.: Combining feature ranking with PCA: An application to gait analysis. In: International Conference on Machine Learning and Cybernetics (ICMLC), pp. 494–499 (2010)
9. Jianning, W.: Kernel-Based Feature Extraction for Automated Gait Classification Using Kinetics Data. In: Fourth International Conference on Natural Computation, ICNC 2008, vol. 4, pp. 162–166 (2008)
10. Zhang, B., Zhang, Y., Begg, R.: Gait classification in children with cerebral palsy by Bayesian approach. J. Pattern Recognition 42(4), 581–586 (2009)
11. Wolf, S., Loose, T., Schablowski, M., Döderlein, L., Rupp, R., Jürgen, H., Bretthauer, G., Mikut, R.: Automated feature assessment in instrumented gait analysis. Gait & Posture 23(3), 331–338 (2006)
12. Winters, T.F., Gage, J.R., Hicks, R.: Gait patterns in spastic hemiplegia in children and young adults. J. Bone Joint Surg (Am) 69, 437–441 (1987)
13. Fundación Hospital Ortopédico Infantil. Caracas Venezuela (2006), http://www.ortopedicoinfantil.org
14. Robb, J.: Gage, J.R. (ed.): The Treatment of Gait Problems in Cerebral Palsy. Clinics in Developmental Medicine No. 164–165. Mac Keith Press, London (2004); Gait & Posture 24(1), 130–130
15. Perry, J.: Distal Rectus Femoris Transfer. Developmental Medicine and Child Neurology 29, 153–158 (1987)
16. Viloria, N.: Electromyographic Evaluation of kinematics classification in Spastic Hemiplegic Patients with pathological gait. Msc Thesis in Biomedical Engineering, Simón Bolívar University, Caracas, Venezuela (2003)
17. Landwehr, N., Hall, M., Frank, E.: Logistic Model Trees (2003), http://www.cs.waikato.ac.nz/~ml/publications/2003/landwehr-etal.pdf
18. Gama, J.: Functional Trees. Machine Learning 55(3) (2004)
19. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001), http://www.cs.colorado.edu/~grudic/teaching/CSCI5622_2004/RandomForests_ML_Journal.pdf
20. Holmes, G., Pfahringer, B., Kirkby, R., Frank, E., Hall, M.: Multiclass alternating decision trees. In: Elomaa, T., Mannila, H., Toivonen, H. (eds.) ECML 2002. LNCS (LNAI), vol. 2430, pp. 161–172. Springer, Heidelberg (2002)
21. Shi, H.: Best-first decision tree learning. Hamilton, NZ (2007)
22. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Wadsworth International Group, Belmont (1984)
23. Quinlan, R.J.: C4.5: Programs for machine learning. Morgan Kaufmann (1993)
24. Webb, G.: Decision Tree Grafting From the All-Tests-But-One Partition. In: Sixteenth Int. Joint Conf. on Artificial Intelligence, pp. 702–707. Morgan Kaufmann, San Francisco (1999)
25. Oliver, J.J., Hand, D.: Averaging Over Decision Stumps. In: Bergadano, F., De Raedt, L. (eds.) ECML 1994. LNCS, vol. 784, pp. 231–241. Springer, Heidelberg (1994)
26. Bermejo, B.: Epidemiología Clínica aplicada a la toma de decisiones en medicina (2001), http://www.cfnavarra.es/salud/docencia.investigacion/textos/Monograf_1/Epidemiologia_clinica.pdf
27. Bravo, R.J., De Castro, O.C., Salazar, A.J.: Spastic hemiplegia gait characterization using support vector machines: Contralateral lower limb. Rev. Fac. Ing. UCV 21(2), 111–119 (2006)
28. Xu, G., Zhang, Y., Begg, R.: Mining Gait Pattern for Clinical Locomotion Diagnosis Based on Clustering Techniques. In: Li, X., Zaïane, O.R., Li, Z.-h. (eds.) ADMA 2006. LNCS (LNAI), vol. 4093, pp. 296–307. Springer, Heidelberg (2006)

Configurations and Couplings: An Exploratory Study

Warwick Graco and Hari Koesmarno

Corporate Analytics, Australian Taxation Office
Canberra Australia
warwick.graco@ato.gov.au

Abstract. This is an exploratory study to see if configurations that were coupled to an output variable could be found in data. The focus in this study was on the modal configurations, which are profiles of best fit for clusters, and their average cluster scores for an output variable. A multistage procedure explained in the paper below was applied to a crime dataset to identify the modal configurations for a sample of cities and towns of the USA and their links to the incidence of violent crime. Three coupled configurations were found including one that was indicative of an African American Configuration having the highest rate of violent crime followed by one indicative of a High Divorce Configuration and one indicative of an Economic Hardship Configuration. The results indicated that using this multistage procedure is feasible for finding modal configurations and their couplings in data. The advantages of this approach are discussed and future directions with the research are outlined.

1 Introduction

Different patterns can be found in data. Affinities [1] are one example. These are associations, sequences, episodes and other arrangements in data. Images, motifs, mosaics, drawings and similar are also examples of patterns [2]. Another example is score configurations. These are sets of scores such as the sets of physical and psychological attribute scores of athletes. They can be represented as rows in a table or alternatively as profiles in terms of graphs [2]. In this paper the terms 'configurations' and 'profiles' are used interchangeably.

A coupling is a pattern that is linked to an output measure such as a performance score, classification or rating. An example is an athlete's physical and psychological profile and the time taken to run 1500 meters.

There are different ways of showing the relationships between configurations and output measures. This is illustrated with relationships between height and weight of people and their performance in various events. One way of doing this is to calculate the correlations between these variables. Another is to partition a population of people into body types or morphologies of tall and heavy build, tall and medium build and tall and light build and similar and to link these morphologies to performance in various athletic events such as high jumping, sprinting and long distance running. For example, athletes who are small, wiry and have short legs tend to do well in long

P. Perner (Ed.): ICDM 2013, LNAI 7987, pp. 266–279, 2013.

distance events while those who are tall, heavy and have long arms tend do well in throwing events [4] & [5]. Here different body morphologies are coupled to different athletic performances. That is, different body morphologies predispose athletes to perform well in certain athletic events.

The couplings between configurations and output measures are not likely to show perfect relationships. That is, people with identical profiles are not likely to have the same output score. There will be differences between profile scores and output measures but the variations are expected to be small if there are casual links between the input scores in the profiles and the corresponding output measures.

In this paper we report an exploratory study to see if this multistage procedure identifies modal configurations and their couplings in data and what changes and refinements that should be made to improve its effectiveness in the future. This is another way of saying this is a proof-of-concept study to show that this approach works.

First, we describe the background and the aim of our research study in Section 2. T The dataset employed for our study is presented in Section 3. In Section 4 we explain the methods used to identify configurations and couplings, while the results obtained and their implications are described in Section 5. In Section 6 we discuss our results and draw conclusions. Finally some future directions with the research are outlined in Section 7.

2 Background and Aim

The couplings of score configurations to an output measure are not a recent development. Meehl [6] provided an example of two binary variables that were uncorrelated with an output measure but a high score on one and a low score on the other predicted the output perfectly.

Configural Frequency Analysis [7] is a method used to identify configurations and their couplings. The aims of this method are firstly, to find the most frequent patterns of binary scores coupled with a binary output in data and secondly, to determine if the frequent patterns are statistically significant in that they occur with a frequency greater than chance. Various statistical tests such as the Chi Square test, the binomial test or the hyper geometric test of Lehmacher can be applied to see which configurations are statistically significant.

An example is the different combinations of symptoms that are associated with flu. Some combinations are more frequent and significant than others. Those that occur with the highest frequencies are expected to indicate the symptoms of this disease.

Configural Frequency Analysis was originally developed to measure the couplings involving binary variables such as having or not having the various symptoms that are judged to be indicative of flu. It has subsequently been applied to configurations involving continuous variables [7].

Another approach to identifying configurations and their couplings is clustering where a dataset is clustered so that the configurations with similar score patterns are placed in the same cluster. Their couplings with output measures can also be included

in the clustering [3]. The profile of best fit can be found for each cluster. This is the modal profile of scores linked to the modal output score.

For example, the psychological profiles of policemen and women and ratings of their effectiveness as detectives can be clustered and the modal profile of each cluster can be identified. The modal profiles of interest will be those that have high modal effectiveness ratings as these indicate the attributes of good detectives.

More than one modal profile maybe linked to an output measure. The authors introduce the term 'tassel'. This can be used to indicate where two or more modal profiles are linked to the same output. They are called this because the different modal patterns are 'knotted' to the same output measure. Each configuration is a string of input variable scores. The advantage of tassels is they reveal where more than one pattern has a relationship with the same output such as say the different combinations of economic and social conditions that are coupled with economic growth. This occurrence can have multiple combinations of causes.

Clustering algorithms are normally applied to all variables (i.e. columns) in a dataset. Not all variables assist with finding clusters. Irrelevant variables mask the clusters by hiding them in noisy data. Feature selection methods can be used to identify variables that discriminate clusters. By doing this they help to remove irrelevant and redundant variables from the analysis.

These distinguishing variables enable the identification of subspaces in the data. A subspace is defined by the distinguishing variables (i.e. the columns of the dataset) and the cases (i.e. rows that have similar profiles) for these variables.

There are various ways [8] to identify subspaces in a dataset. One strategy is to first distinguish the variables which describe the same class in data. For example, apples, oranges and pears are classes of fruit. They each have certain distinguishing characteristics. These characteristics are found in the columns of datasets. Techniques which can discriminate the defining variables of classes are called 'taxometric' or 'taxonic methods'. Meehl and associates [9] have developed a number of techniques to do this task. The term 'taxonic' is used hereafter in this paper.

Once the distinguishing variables of a class are identified, the rows defined by the class variables can be clustered. For example, apples can be placed in different buckets (i.e. clusters) based on attributes such as size, color, texture and taste. Furthermore, apples in the different buckets can be coupled to the average price they would fetch in the marketplace.

This is the approach pursued in the research reported in this paper. A multistage procedure is employed. In the first step a taxonic method is applied to a dataset to identify the classes and their defining variables. An output variable is also included in the taxonic analysis to see which class (or classes) includes this variable. In the second step, principal component analysis [10] is applied to the classes that have many defining variables to identify which ones load on the same component. This includes the output variable. In the third step the resulting components that include the coupled variable are clustered to see which clusters have modal profiles with high scores on the output variable.

The aim is to identify configurations of modal variable scores in the crime dataset that are coupled to high levels of per capita crime variable. It is expected that these modal configurations will help reveal the defining attributes of US cities and towns that have high levels of violent crime.

3 The Data

A community and crime dataset was used in this study. It was chosen from the Machine Learning Repository at the University of California Irvine (UCI) [11]. It was selected because it had 128 attributes and 1994 instances to do the analysis reported here. The dataset includes socio-economic data from the 1990 US Census [12], law enforcement data from the 1990 US Law Enforcement Management and Administrative Statistics (LEMAS) Survey [13] and crime data from the 1995 FBI Uniform Crime Report [14].

The instances were selected cities and towns in the USA. The input variables includes community ones such as the percent of the population considered urban and the median family income and law-enforcement ones such as per capita number of police officers, and percent of officers assigned to drug units. Details of the input variables and their basic statistics can be at the UCI website [11]. There are too many to list in this paper. These variables were selected to see how well they predict the level of violent crime in each city or town. This was the per capita violent crimes variable.

Violent crimes include murder, rape, robbery and assault. There were discrepancies counting of rapes in some states and in turn resulted in incorrect values for per capita violent crime variable. These cities are not included in the dataset. Many of these omitted communities were from the Middle Western USA.

All numeric data was normalized into the decimal range 0.00-1.00 using an unsupervised, equal-interval binning method. Attributes retain their distribution and skew (hence for example the population attribute has a mean value of 0.06 because most communities are small).

The normalization preserves rough ratios of values within an attribute. All values more than three standard deviations above the mean are normalized to 1.00 and all values more than three standard deviations below the mean are normalized to 0.00. The normalization does not preserve relationships between values between attributes. For example, it would not be meaningful to compare the value for per capita income for Caucasians (white Per Cap) with the value for per capita income for African Americans (black Per Cap) for a community.

A limitation of the LEMAS survey [13] was that it was restricted to police departments with at least 100 officers. It also included a random sample of smaller departments. Communities not found in both census and crime datasets were also omitted thus many are missing from this dataset.

Some of the variables had missing values and these were deleted from the study. This reduced the number of input variables to from 128 to 99.

4 Methodology

There were a number of steps carried out in the analysis of the crime dataset. They included the sequence of identifying:

Classes -> Components -> Clusters -> Configurations -> Couplings.

Four classes were identified in the data and the 1st of these consisted of 74 defining variables including the output variable of per capita violent crime. The decision was taken to identify the principal components of these defining variables. One of the principal components consisted of variables that included a high loading on the output variable. This was clustered to identify the modal profiles of these clusters which had high scores on the output variable. The results obtained from using these methods are shown in the Results Section 5 below. The specific methods used in this study are explained next.

4.1 Taxonic Analysis

Previously it was indicated that Meehl and associates devised methods [9] to identify class structure in data. Another algorithm written in Base SAS was employed in this study. It was used because it discovers multiple classes in data whereas the methods developed by Meehl and associates [9] were devised to identify single classes.

The taxonic method employed is called 'outlier table analysis. How it works is explained in [15]. Basically it is analogous to identifying the defining signature of each case in a dataset and sorting the cases so that those with very similar signatures are placed in the same class. A signature is a profile of scores that all cases in the class share. In practice not all cases in a class have identical signatures and instead will have small variations such as one case may have ten of the 12 defining variables while another has ten but they are slightly different variables. Outlier table analysis accounts for these differences.

How this is done in practice is as follows. The first step is to convert the scores for all variables in the dataset to percentile ranks so that they all have the same range. The second step is that a high and a low percentile rank cutoff score are specified for the dataset such as the 70th percentile rank for high scores and the 30th percentile rank for low scores. This is done because the discriminatory variables of a class tend to have either high or low percentile rank scores. The third step is that outlier table analysis does frequency counts using each variable in the dataset as an anchor. That is, for each anchor variable score for each case if it is above the 70th percentile rank than add '1' to a frequency count for both the anchor variable and the variable that has a percentile rank score above the cutoff. Table 1 below shows frequency counts where the anchor and other variables had scores above the cutoff.

Table 1. High-High Outlier Table Frequency Count

Variables	*1*	*2*	*3*	*4*	*5*
1	100	95	85	99	70
2	95	100	90	86	79
3	85	90	100	91	87
4	99	86	91	100	93
5	70	79	87	93	100

The anchor variables are in the diagonal of the table from top left to bottom right (i.e. those with a frequency count of 100). If anchor variable 1 is taken as an example, there were 95 instances where variable 2 was above the high cutoff when the anchor variable 1 was above the same cutoff.

The fourth step is to repeat the process with anchor variable scores above the high cutoff and the other variable scores below the low cutoff to derive a High-Low table. This shows variables that have low percentile rank scores when the anchor variable has a high percentile rank score. The fifth step is to subtract the High-Low scores from their corresponding High-High scores in the respective tables to derive a difference table of scores. If variables are discriminatory they will have either very high frequency counts in one table and very low in the other or vice versa. If they are non-discriminatory in that they are not characteristic of a class they will tend to have similar frequency counts in both tables. Therefore the difference scores will be low.

The sixth step is to produce a correlation matrix using the columns of scores in the difference matrix. This is done using a nonparametric correlation coefficient such as Spearman Rho [16]. A nonparametric coefficient is employed because the scores in the difference matrix are not normally distributed. The last step is to do a principal component analysis of the correlation matrix. How this method works is described in the next section. Principal component analysis [10] (hereafter referred to as component analysis) is conducted to identify columns of the correlation matrix that are redundant in that they show the same pattern of correlations as other columns. That is, they have the same defining variables of a class. The rotated component solution shows the non-redundant variables that define each class. SAS JMP Principal Component procedure [17] was applied to identify the uncorrelated patterns in the difference matrix.

It is highly advisable to repeat this taxonic analysis using different cutoffs such as the 60th and 40th, 75th and 25th, 80th and 20th and 90th and 10^{th} percentile ranks to check for consistency of results. The cutoffs selected depend upon the number of cases in the dataset. If it is very large, say in the tens of thousands of cases, one can be

liberal with the cutoffs such as using 80th and 20th percentile ranks and above. If the number of cases is low, say as small as 300, it is wise to use conservative cutoffs of 51st and the 49th percentile ranks to test for taxonicity of variables.

4.2 Components Analysis

It pays to do principal component analysis again where a class has a large number of defining variables. This will reveal the class defining variables in each component for a class. The principal component analysis will also show which components have high loadings on the violent crime output variable. SAS JMP Principal Component procedure [17] was used for this purpose.

4.3 Cluster Method

Components which have a high loadings on the output crime variable can be clustered to see which clusters have modal profiles linked to a high incidence of crime. A number of cluster methods were experimented with and the decision was taken to use a grid approach. This is because it gave clusters that were similar in size. The approach used is a modification of self-organizing maps (SOM) [18]. The grid approach used resembles k-means clustering more that the neural network learning used by SOM. The goal is to form clusters on a cluster grid, such that points in clusters that are near each other in the grid are also near each other in multivariate space.

The SAS JMP Cluster Package [17], where the SOM option was selected, was used to do the grid clustering. Ten cluster solutions were generated using the component defining variables. There was no need to standardize scales are they were already scaled between 0.00 and 1.00. The Johnson Transform [19] option was also employed to balance highly skewed variables, or to bring outliers closer to the center of the rest of the values. A number of the variables used in the analysis had skewed distributions.

5 Results

Only selected results are shown in this paper because of the large number of variables employed in this analysis. There were four classes found in the crime dataset with the first accounting for approximately 65 percent, the second 28 percent, the third 2.65 percent and the fourth 2.35 percent of the variance. The 1st and 3rd classes had loadings of 0.58 and 0.55 for the crime per capita output variable. The other classes had very low-to-zero loadings on this variable. There were 74 defining variables for the 1st class including the crime output variable and five for the 3rd class including the crime output variable. Because the 3rd class is so small with few defining variables, attention is focused on the 1st class in the remainder of this paper.

Component analysis was performed on the 74 variables of the 1st class and results for the component with the highest loading with the crime output variable are shown in Table 2. The other components had loadings that were from low-to-zero in value on this variable.

The results in Table 2 reveal a component consisting of cities and towns containing low income African Americans that have high rates of violent crime. There is high percentage of broken families with public assistance income and their living in squalid conditions with no telephone. Many are unemployed. This could be called an 'African American Crime Pattern or Configuration'.

The variables in Table 2 were used in the cluster analysis and the resulting cluster means are shown in Table 4. These represent the modal profiles for the clusters.

Clusters 1 to 3 in Table 4; which have cluster sizes of 136, 116 and 183 cities and towns; have the highest mean scores for the crime output variable with values of 0.68, 0.42 and 0.51 respectively. They are above the population mean score of 0.24 for this variable shown in Table 3.

Table 2. Defining Variables for Principal Component 2 for the 1st Class

loading	***Attributes***
0.88	Percentage of population that is African American
-0.76	Percentage of population that is White American
-0.51	Percentage of households with investment / rent income
0.53	Percentage of households with public assistance income
0.48	Percentage of people under the poverty level
0.44	Percentage of people 16 and over, in the labor force, and
0.56	Percentage of males who are divorced
0.60	Percentage of females who are divorced
0.60	Percentage of population who are divorced
-0.78	Percentage of families (with kids) that are headed by two parents
-0.80	Percentage of kids in family housing with two parents
-0.74	Percent of kids 4 and under in two parent households
-0.77	Percent of kids age 12-17 in two parent households
0.83	Percentage of kids born to never married
0.53	Percentage of vacant housing that is boarded up
0.51	Percentage of occupied housing units without phone
Outcome Variable Loading	
0.70	Total number of violent crimes per 100K population

The cluster means for Clusters 1 to 3 that indicate patterns in the clusters are shaded in Table 4. What is noticeable with the cluster 1 is that the cities and towns display the African American Crime Pattern noted above in Table 2. That is, the cities and towns in this cluster contain low income African Americans with high unemployment, high levels of poverty, high public assistance and many broken families.

The other two clusters have the mean profiles of cities and towns having lower percentages of African Americans. The mean profile for Cluster 2 has cities and towns with households with high unemployment, high percentage of public assistance, lower investment and rental income and citizens under the poverty level. This can be labeled an 'Economic Hardship Pattern or Configuration'.

Cluster 3 has cities and towns with a higher incidence of divorce but lower levels of economic hardship (e.g. less poverty, less unemployment and less requiring public assistance) compared to the patterns evident in Clusters 1 and 2. This is labeled a 'High Divorce Pattern or Configuration'.

Together these three clusters reveal three different coupled configurations with Cluster 1 the African American Configuration having the highest rate of violent crime per 100,000 population of an average of 0.68, Cluster 3 the High Divorce Configuration having the second highest rate of violent crime of an average of 0.51 and Cluster 2 the Economic Hardship Configuration having the third highest rate of an average of 0.42. The results do not reveal the reasons for these patterns.

6 Discussion

The aim of this study was not to develop specific insights and understandings into the drivers of violent crime in cities and towns in the USA. It was instead to see if the multistage procedure described in this paper for identifying coupled configurations in data is a feasible one. The results obtained show that it works.

The results also indicate that the multistage procedure has four advantages. The first is that it used what amounts to two-mode clustering to identify subspaces in data. The subspaces consist of the defining attributes of a component and cases which have similar profiles. Each subspace represents a pattern of behavior.

The second advantage is that this procedure showed the gains that can be made by including the output variable in the taxonic analysis. It resulted in the identification of two classes that contained this variable. One class had many variables and the other few. Attention was focused on the larger class and it was found to have one component that had a high loading on the violent-crime output variable. It was shown to have cities and towns that have an African Crime Pattern.

When the defining variables for this component were clustered it showed that there were three clusters of cities and towns that had sizeable scores on the output variable. One matched the African Crime Pattern identified by component analysis while the other two revealed a configuration of cities and towns having a High Divorce Pattern and those having an Economic Hardship Pattern. The advantage of the clustering was that it revealed two other patterns that were not shown by the results of the component analysis.

From this perspective, it is suggested that if suitable data was collected on individuals and gangs who commit violent crimes, rather than cities and towns where crime occurs, it is likely that other modal configurations of violent criminals would be identified.

A few possible ones include those who are young, rich, spoilt and bored and who see violent crime as an escape. Other young ethnic groups besides traditional white Caucasian ones and African American ones might also be identified as being involved in violent crimes. Each of these will have a modal configuration scores.

The concept of a tassel was introduced earlier in the paper. If a rating rather than a score was used to indicate high crime rates it would allow the different modal configurations to be knotted to this rating. This would show the different typologies of criminals associated with violent crime.

The third advantage is that taxonic analysis, when combined with component analysis for classes with large number of variables, is an effective way of discovering class structure in data.

What was learned from various taxonic analyses [eg 15] conducted by the authors of this paper is the need to have a broad and representative number of variables to discriminate classes. If important variables are left out of the data they can make a difference between finding a class or not finding one. It pays to be over inclusive with what the person doing taxonic analyses judges to be relevant variables when doing this type of analysis.

The fourth advantage is that including the output variables in the taxonic analysis and the component analysis is a way of short circuiting the process of identifying which input variables are linked to an output variable. It reduces the search space that has to be explored to find coupled configurations.

In terms of steps that can be taken to improve the multistage research results reported here, one is that independent components [20] rather than principal components analysis should have been employed in this study. This is because a number of variables had skewed distributions. Independent components analysis is better suited to finding components in these variables.

The study can also be repeated by including the variables with missing values. The missing values can be imputed using different techniques [21].

In terms of future directions, it is intended recode the methods used to find classes, components and clusters into a single R package. This will provide integrated environment for doing the analysis reported in this paper. It will also allow those doing this analysis to use other R techniques besides those employed in this study to find class structure and other constructs covered in this research.

The outlier table analysis has the limitation that it is currently restricted to using continuous variables. It is intended to extend it to include categorical variables in the identification of classes and clusters. An example of a categorical variable is gender and whether some classes and clusters have a strong male or female presence.

The research reported in this paper used one clustering algorithm. There are moves to use ensemble approaches to clustering where the final clusters are based on the results of multiple clustering algorithms [22]. This can be researched further to see if it is better suited to recovering clusters in data.

Table 3. Population Statistics for Variables in Principal Component 2 for the 1st Class

Population Statistics	**Variables**	% population African American	% population White American	% households with investment / rent income	%with public assistance income	% under poverty level	% Over 16 & Unemployed	% male divorced	% female divorced	% divorced	% family two parents	% Kids with two parents	% Teen With two parents	% Kids born to never married parents	% Vacant housing Boarded up	% occupied home with no phone	**Violent crimes per population**
Mean		0.18	0.75	0.50	0.32	0.30	0.36	0.46	0.49	0.49	0.61	0.66	0.58	0.24	0.20	0.26	0.24
Stand Dev		0.25	0.24	0.18	0.22	0.23	0.2	0.18	0.18	0.18	0.2	0.22	0.19	0.23	0.22	0.24	0.23
Mdn		0.06	0.85	0.48	0.26	0.25	0.32	0.47	0.5	0.5	0.63	0.7	0.61	0.17	0.13	0.185	0.15
Mode		0.01	0.98	0.41	0.1	0.08	0.24	0.56	0.54	0.57	0.7	0.91	0.6	0.09	0.00	0.01	0.03

Table 4. Cluster Means for Principal Component 2 for the 1st Class

	Variables	% population African American	% population White American	% households with investment / rent income	%with public assistance income	% under poverty level	% Over 16 & Unemployed	% male divorced	% female divorced	% divorced	% family two parents	% Kids with two parents	% Teen With two parents	% Kids born to never married par-	% Vacant housing Boarded up	% occupied home with no phone	**Violent crimes per population**
Cluster	**Size**																
1	1	0.86	0.22	0.26	0.6	0.6	0.65	0.6	0.6	0.7	0.1	0.1	0.1	0.8	0.5	0.60	0.68
2	1	0.24	0.48	0.26	0.7	0.6	0.72	0.4	0.5	0.5	0.4	0.4	0.4	0.4	0.3	0.62	0.42
3	1	0.43	0.56	0.36	0.4	0.4	0.46	0.6	0.6	0.6	0.4	0.4	0.3	0.4	0.3	0.47	0.51
4	1	0.32	0.52	0.40	0.3	0.3	0.40	0.4	0.5	0.5	0.5	0.5	0.5	0.3	0.2	0.26	0.34
5	1	0.10	0.82	0.48	0.3	0.4	0.41	0.3	0.4	0.4	0.6	0.6	0.5	0.2	0.1	0.28	0.16
6	2	0.09	0.85	0.41	0.4	0.4	0.48	0.5	0.6	0.6	0.5	0.5	0.5	0.2	0.2	0.44	0.20
7	1	0.90	0.82	0.50	0.2	0.2	0.29	0.6	0.6	0.6	0.5	0.5	0.5	0.2	0.1	0.22	0.24
8	2	0.07	0.87	0.55	0.1	0.1	0.24	0.4	0.5	0.5	0.6	0.7	0.6	o.1	0.1	0.13	0.14
9	3	0.05	0.90	0.62	0.1	0.1	0.24	0.3	0.3	0.3	0.7	0.8	0.7	0.9	0.1	0.07	0.08
10	2	0.03	0.92	0.78	0.0	0.0	0.17	0.1	0.2	0.2	0.8	0.9	0.8	0.5	0.0	0.02	0.05

Lastly, the approach used in this paper will also be tested on larger and far more diverse industrial strength datasets to see how it performs with finding classes, components, clusters, configurations and couplings. This will provide further evidence of the effectiveness of this approach.

7 Conclusion

A multistage procedure that consisted of the sequence of a taxonic method, followed by principal component analysis and then cluster analysis, which was a modification of a SOM, was applied to a crime dataset to see if configurations that were coupled to incidence of violent crime in the cities and towns of the USA could be discerned. Three coupled configurations were found including one that was indicative of an African American Configuration having the highest rate of violent crime followed by one indicative of a High Divorce Configuration and one indicative of an Economic Hardship Configuration. The results indicated that using this sequence is a feasible approach to finding configurations and their couplings in data.

Coupled configurations are an alternative to correlations for showing the relationships between input and output variables. Meehl [6] showed that a coupled configuration was found in data where there were zero correlations between variables. This indicates that it is advisable that researchers check their data for coupled configurations if the correlations between variables are weak. Another advantage of coupled configurations is that they can reveal the different modal patterns that are associated with an output such as the different combinations of conditions that contribute to economic output, managerial performance and sporting prowess to name a few examples. This assists to understand the various causes of issues thus providing further insights that can result in better decisions by decision makers. For these reasons coupled configurations are considered to be of value in finding in data.

References

1. Han, J., Kamber, M.: Data Mining: Concepts and Techniques, Section 6.3. Morgan Kaufmann Publishers, San Francisco (2006)
2. Perner, P. (ed.): Data Mining on Multimedia Data. LNCS (LNAI), vol. 2558, pp. 1–11. Springer, Heidelberg (2002)
3. Jaenichen, S., Perner, P.: Conceptual Clustering and Case Generalization of two-dimensional Forms. Computational Intelligence 22(3/4), 178–193 (2006)
4. Quigley, B.: Assessing the Athlete – Potential and Progress in Sports Coaching, pp. 78–79. Australian Government Printing Service, Canberra (1976)
5. Hemery, D.: The Pursuit of Sporting Excellence, pp. 28–30. Willow Books, London (1986)
6. Meehl, P.E.: Configural Scoring. Journal of Consulting Psychology 14, 165–171 (1950)
7. von Eye, A.: Introduction to Configural Frequency Analysis: The search for types and antitypes in cross-classifications. Cambridge University Press, Cambridge (1990)

8. Kriegel, H.-P., Kroger, P., Zimek, A.: Clustering High-Dimensional Data: A Survey on Subspace Clustering, Pattern-Based Clustering, and Correlation Clustering. ACM Transactions on Knowledge Discovery from Data (March 2009)
9. Waller, N.G., Meehl, P.E.: Multivariate Taxometric Procedures: Distinguishing Types from Continua. Safe, Thousand Oaks (1998)
10. Jolliffe, I.T.: Principal Component Analysis, 2nd edn. Springer, Berlin (2002)
11. See crime site, http://archive.ics.uci.edu/ml
12. U. S. Department of Commerce, Bureau of the Census, Census of Population and Housing. Summary Tape File 1a & 3a (Computer Files), United States (1990)
13. U.S. Department of Justice, Bureau of Justice Statistics, Law Enforcement Management And Administrative Statistics (Computer File) U.S. Department Of Commerce, Bureau Of The Census Producer and Inter-university Consortium for Political and Social Research, Washington, DC, Ann Arbor, Michigan (1992)
14. U.S. Department of Justice, Federal Bureau of Investigation, Crime in the United States (Computer File) (1995)
15. Koesmarno, H.K., Graco, W.J., He, H., Cooksey, R.W.: MAMBAC versus Outlier tables for Identifying Classes in Data. In: Proceedings of the Third International Conference on the Practical Applications of Knowledge Discovery and Data Mining, pp. 107–125. The Practical Application Company, London (1999)
16. Kendall, M.G.: Rank Correlation Methods, 4th edn. Griffin, London (1970), http://en.wikipedia.org/wiki/Spearman_Rho
17. Sall, J., Lehman, A., Stephens, M., Creighton, L.: JMP Start Statistics: A Guide to Statistics and Data Analysis using JMP, 5th edn. SAS Institute, Cary (2012), http://www.jmp.com
18. Kohonen, T.: Self Organizing Maps, 3rd edn. Springer, Berlin (2001)
19. Chou, Y., Polansky, A.M., Mason, R.L.: Transforming Nonnormal Data to Normality in Statistical Process Control. Journal of Quality Technology 30, 133–141 (1998)
20. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. Wiley, New York (2001)
21. Enders, C.K.: Applied Missing Data Analysis, 1st edn. Guildford Press, New York (2010), http://en.wikipedia.org/wiki/Missing_data
22. Wang, C., She, Z., Cao, L.: Coupled Clustering Ensemble: Incorporating Couplings Relationships Both between Base Clusterings and Objects. In: Paper to be presented to the 29th IEEE International Conference on Data Engineering (2013)

Author Index