Chapter 1
Data Mining Techniques in Clustering, Association and Classification

Dawn E. Holmes[1], Jeffrey Tweedale[2], and Lakhmi C. Jain[3]

[1] Department of Statistics and Applied Probability
University of California Santa Barbara
Santa Barbara
CA 93106-3110
USA
[2] School of Electrical and Information Engineering
University of South Australia
Adelaide
Mawson Lakes Campus
South Australia SA 5095
Australia
[3] School of Electrical and Information Engineering
University of South Australia
Adelaide
Mawson Lakes Campus
South Australia SA 5095
Australia

1 Introduction

The term *Data Mining* grew from the relentless growth of techniques used to interrogation masses of data. As a myriad of databases emanated from disparate industries, management insisted their information officers develop methodology to exploit the knowledge held in their repositories. The process of extracting this knowledge evolved as an interdisciplinary field of computer science within academia. This included study into statistics, database management and Artificial Intelligence (AI). Science and technology provide the stimulus for an extremely rapid transformation from data acquisition to enterprise knowledge management systems.

1.1 Data

Data is the representation of anything that can be meaningfully quantized or represented in digital form, as a number, symbol or even text. We process data into information by initially combining a collection of artefacts that are input into a system which is generally stored, filtered and/or classified prior to being translated into a useful form for dissemination [1]. The processes used to achieve this task have evolved over many years and has been applied to many situations using a magnitude of techniques. Accounting and pay role applications take center place in the evolution of information processing.

D.E. Holmes, L.C. Jain (Eds.): Data Mining: Found. & Intell. Paradigms, ISRL 23, pp. 1–6.
springerlink.com

Data mining, expert system and knowledge-based system quickly followed. Today we live in an information age where we collect data faster than it can be processed. This book examines many recent advances in digital information processing with paradigms for acquisition, retrieval, aggregation, search, estimation and presentation.

Our ability to acquire data electronically has grown exponentially since the introduction of mainframe computers. We have also improved the methodology used to extract information from data in almost every aspect of life. Our biggest challenge is in identifying targeted information and transforming that into useful knowledge within the growing collection of noise collected in repositories all over the world.

1.2 Knowledge

Information, knowledge and wisdom are labels commonly applied to the way humans aggregate practical experience into an organised collection of facts. Knowledge is considered a collection of facts, truths, or principles resulting from a study or investigation. The concept of knowledge is a collection of facts, principles, and related concepts. Knowledge representation is the key to any communication language and a fundamental issue in AI. The way knowledge is represented and expressed has to be meaningful so that the communicating entities can grasp the concept of the knowledge transmitted among them. This requires a good technique to represent knowledge. In computers symbols (numbers and characters) are used to store and manipulate the knowledge. There are different approaches for storing the knowledge because there are different kinds of knowledge such as facts, rules, relationships, and so on. Some popular approaches for storing knowledge in computers include procedural, relational, and hierarchical representations. Other forms of knowledge representation used include *Predicate Logic, Frames, Semantic Nets, If-Then rules and Knowledge Inter-change Format*. The type of knowledge representation to be used depends on the AI application and the domain that Intelligent Agents (IAs) are required to function [2]. Knowledge should be separated from the procedural algorithms in order to simplify knowledge modification and processing. For an IA to be capable of solving problems at different levels of abstraction, knowledge should be presented in the form of frames or semantic nets that can show the *is-a* relationship of objects and concepts. If an IA is required to find the solution from the existing data, Predicate logic using IF-THEN rules, Bayesian or any number of techniques can be used to cluster information [3].

1.3 Clustering

In data mining a cluster is the resulting collection of similar or same items from a volume of acquired facts. Each cluster has distinct characteristics, although each has a similarity, its size is measured from the centre with a distance or separation from the next [4]. Non-hierarchical clusters are generally partitioned by class or clumping methods. Hierarchical clusters produce sets of nested groups that need to be progressively isolated as individual subsets. The methodology used are described as: partitioning, hierarchical agglomeration, Single Link (SLINK), Complete Link (CLINK), group average and text based document methods. Other techniques include [5]:

- A Comparison of Techniques,
- Artificial Neural Networks for Clustering, and
- Clustering Large Data Sets, and
- Evolutionary Approaches for Clustering, and
- Fuzzy Clustering, and
- Hierarchical Clustering Algorithms, and
- Incorporating Domain Constraints in Clustering, and
- Mixture-Resolving and Mode-Seeking Algorithms, and
- Nearest Neighbour Clustering, and
- Partitional Algorithms, and
- Representation of Clusters, and
- Search-Based Approaches.

Where clustering can typically be applied in Image Segmentation, Object/Character Recognition, Information Retrieval and Data Mining.

1.4 Association

Data is merely a collection of facts. To make sense of that collection, a series of rules can be created to sort, select and match a pattern of behavior or association based on specified dependancies or relationships. For instance a collection of sales transaction within a department store can hold a significant volume of information. If a cosmetics manager desired to improve sales, knowledge about existing turnover provides an excellent base-line (this is a form of market analysis). Similarly, using the same data set, the logistics manager could determine inventory levels (this concept is currently associated with trend analysis and prediction). Association rules allow the user to reveal sequences, links and unique manifestations that emerge over time [6]. Typically cross-tabulation can be used where items, words or conjunctions are employed to analyse simple collections that are easily classified, such as age, cost or gender.

1.5 Classification

Data bases provide an arbitrary collection of facts. In order to make sense of the random nature of such collections, any number of methods can be used to map the data into usable or quantifiable categories based on a series of attributes. These subsets improve efficiency by reducing the noise and volume of data during subsequent processing. The goal is to predict the target class for each case. An example would be to measure the risk management of an activity, as either low, high or some category in between. Prior to classification, the target categories must be defined before the process is run [7]. A number of AI techniques are used to classify data. Some include decision-trees, rule-based, Bayesian, rough sets, dependency networks, Support Vector Machines (SVM), Neural Networkss (NNs), genetic algorithms and fuzzy logic.

2 Data Mining

There are many commercial data mining methods, algorithms and applications, with several that have had major impact. Examples include: *SAS*[1], *SPSS*[2] and *Statistica*[3]. Other examples are listed in sections 2.1 and 2.2. Any number can be found on-line, and many are free. Examples include: *Environment for DeveLoping KDD-Applications Supported by Index-Structures (ELKI)*[4], *General Architecture for Text Engineering (GATE)*[5] and *Waikato Environment for Knowledge Analysis (Weka)*[6].

2.1 Methods and Algorithms

- Association rule learning,
- Cluster analysis, and
- Constructive induction, and
- Data analysis, and
- Decision trees, and
- Factor analysis, and
- Knowledge discovery, and
- Neural nets, and
- Predictive analytics, and
- Reactive business intelligence, and
- Regression, and
- Statistical data analysis, and
- Text mining.

2.2 Applications

- Customer analytics,
- Data Mining in Agriculture, and
- Data mining in Meteorology, and
- Law-enforcement, and
- National Security Agency, and
- Quantitative structure-activity relationship, and
- Surveillance.

[1] See `http://www.sas.com/`
[2] See `http://www.spss.com/`
[3] See `http://www.statsoft.com/`
[4] See `http://www.dbs.ifi.lmu.de/research/KDD/ELKI` from Ludwig Maximillian University.
[5] See `gate.ac.uk` from the University of Sheffield.
[6] See `http://www.cs.waikato.ac.nz/~ml/weka/` from the University of Waikato.

3 Chapters Included in the Book

This book includes twelve chapters. Each chapter is described briefly below. Chapter 1 provides an introduction to data mining and presents a brief abstract of each chapter included in the book. Chapter 2 is on clustering analysis in large graphs with rich attributes. The authors state that a key challenge for addressing the problem of clustering large graphs with rich attributes is to achieve a good balance between structural and attribute similarities. Chapter 3 is on temporal data mining. A temporal association mining problem, based on similarity constraint, is presented. Chapter 4 is on Bayesian networks with imprecise probabilities. The authors report extensive experimentation on public benchmark data sets in real-world applications to show that on the instances indeterminately classified by a creedal network, the accuracy of its Bayesian counterpart drops.

Chapter 5 is on hierarchical clustering for finding symmetries and other patterns in massive, high dimensional datasets. The authors have illustrated the powerfulness of hierarchical clustering in case studies in chemistry and finance. Chapter 6 is on randomized algorithm of finding the true number of clusters based on Chebychev polynomial approximation. A number of examples are used to validate the proposed algorithm. Chapter 7 is on Bregman bubble clustering. The authors present a broad framework for finding k dense clusters while ignoring rest of the data. The results are validated on various datasets to demonstrate the relevance and effectiveness of the technique.

Chapter 8 is on *DepMiner*. It is a method for implementing a model for the evaluation of item-sets, and in general for the evaluation of the dependencies between the values assumed by a set of variables on a domain of finite values. Chapter 9 is on the integration of dataset scans in processing sets of frequent item-set queries. Chapter 10 is on text clustering with named entities. It is demonstrated that a weighted combination of named entities and keywords are significant to clustering quality. The authors present implementation of the scheme and demonstrate the text clustering with named entities in a semantic search engine.

Chapter 11 is on learning from imbalanced data. Using experimentations, the authors made some recommendations related to the data evaluation methods. Finally Chapter 12 is on regional association rule mining and scoping from spatial data. The authors have investigated the duality between regional association rules and regions where the associations are valid. The design and implementation of a reward-based region discovery framework and its evaluation are presented.

4 Conclusion

This chapter presents a collection of selected contribution of leading subject matter experts in the field of data mining. This book is intended for students, professionals and academics from all disciplines to enable them the opportunity to engage in the state of art developments in:

- Clustering Analysis in Large Graphs with Rich Attributes;
- Temporal Data Mining: Similarity-Profiled Association Pattern;
- Bayesian Networks with Imprecise Probabilities: Theory and Application to Classification;
- Hierarchical Clustering for Finding Symmetries and Other Patterns in Massive, High Dimensional Datasets;
- Randomized Algorithm of Finding the True Number of Clusters Based on Chebychev Polynomial Approximation;
- Bregman Bubble Clustering: A Robust Framework for Mining Dense Clusters;
- DepMiner: A method and a system for the extraction of significant dependencies;
- Integration of Dataset Scans in Processing Sets of Frequent Itemset Queries;
- Text Clustering with Named Entities: A Model, Experimentation and Realization;
- Regional Association Rule Mining and Scoping from Spatial Data; and
- Learning from Imbalanced Data: Evaluation Matters.

Readers are invited to contact individual authors to engage with further discussion or dialog on each topic.

References

1. Moxon, B.: Defining data mining, the hows and whys of data mining, and how it differs from other analytical techniques. DBMS 9(9), S11–S14 (1996)
2. Bigus, J.P., Bigus, J.: Constructing Intelligent Agents Using Java. Professional Developer's Guide Series. John Wiley & Sons, Inc., New York (2001)
3. Tweedale, J., Jain, L.C.: Advances in information processing paradigms. In: Watanabe, T. (ed.) Innovations in Intelligent Machines, pp. 1–19. Springer, Heidelberg (2011)
4. Bouguettaya, A.: On-line clustering. IEEE Trans. on Knowl. and Data Eng. 8, 333–339 (1996)
5. Jain, A., Murty, M., Flynn, P.: Data clustering: A review. ACM Computing Surveys 3(3), 264–323 (1999)
6. Hill, T., Lewicki, P.: Statistics: Methods and Applications, StatSoft, Tulsa, OK (2007)
7. Classification, clustering, and data mining applications. In: Banks, D., House, L., McMorris, F., Arabie, P., Gaul, W. (eds.) International Federation of Classification Societies (IFCS), Illinois Institute of Technology, Chicago, p. 658. Springer, New York (2004)