Implementation of Clustering-Based Feature

Kamepalli Sujatha

International Journal of Emerging Trends & Technology in Computer Science

International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 3, Issue 3, May – June 2014 ISSN 2278-6856

Implementation of Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data

Sujatha Kamepalli [1], Radha Mothukuri [2]

[1] Associate Professor,CSE Dept,
Malineni Lakshmaiah Engineering College,Singaraya Konda,Prakasam District, Andhra Pradesh

[2] Lecturer in CSE Dept
Bapatla Women's Engineering College, Bapatla, Guntur District, Andhra Pradesh,

Abstract: ***Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Feature selection is the process of selecting a subset of relevant features for use in model construction. The central assumption when using a feature selection technique is that the data contains many redundant or irrelevant features. Of the many feature subset selection algorithms, some can effectively eliminate irrelevant features but fail to handle redundant features yet some of others can eliminate the irrelevant while taking care of the redundant features. Our proposed FAST algorithm falls into the second group. Traditionally, feature subset selection research has focused on searching for relevant features. This paper provides the implementation of this algorithm on high dimensional data.***

Keywords: Data mining, Feature selection, FAST algorithm, relevant features, redundant features.

1. INTRODUCTION

1.1 Data mining:

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Data mining tasks are specified by its functionalities that tasks are classified into two forms:

1. Descriptive mining tasks: Portray the general properties of the data.
2. Predictive mining tasks: Perform the implication on the current data order to craft prediction.

Data mining Functionalities are:

- Characterization and Discrimination
- Mining Frequent Patterns
- Association and Correlations
- Classification and Prediction
- Cluster Analysis
- Outlier Analysis
- Evolution Analysis [1].

1.2 Feature Selection:

Feature selection is the process of selecting a subset of relevant features for use in model construction. The central assumption when using a feature selection technique is that the data contains many redundant or irrelevant features. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context. Feature selection techniques are a subset of the more general field of feature extraction. Feature extraction creates new features from functions of the original features, whereas feature selection returns a subset of the features. Feature selection techniques are often used in domains where there are many features and comparatively few samples (or data points) [2].

Research on feature selection has been done for last several decades and is still in focus. Reviews and books on feature selection can be found in [3, 4, and 5]. Recent papers such as [6, 7, 8, 9, 10, and 11] address some of the existing issues of feature selection.

Feature subset selection is an effectual way for dimensionality reduction, elimination of inappropriate data, rising learning accurateness, and recovering result unambiguousness. Numerous feature subset selection methods have been planned and considered for machine learning applications [12]. They can be separated into four major categories such as: the Wrapper, Embedded, and Filter and Hybrid methods. In particular, we accept the minimum spanning tree based clustering algorithms, for the reason that they do not imagine that data points are clustered around centers or separated by means of a normal geometric curve and have been extensively used in tradition [3].

2. RELATED WORK

Zheng Zhao and Huan Liu in "Searching for Interacting Features" propose to efficiently handle feature interaction to achieve efficient feature selection [13]. S.Swetha, A.Harpika in"A Novel Feature Subset Algorithm for High Dimensional Data", build up a novel algorithm that can capably and efficiently deal with both inappropriate and redundant characteristics, and get hold of a superior feature subset [14]. T.Jaga Priya Vathana,

C.Saravanabhavan, Dr.J.Vellingiri in"A Survey on Feature Selection Algorithm for High Dimensional Data Using Fuzzy Logic" proposed fuzzy logic has focused on minimized redundant data set and improves the feature subset accuracy [15]. Manoranjan Dash, Huan Liub in "Consistency-based search in feature selection", focuses on inconsistency measure according to which a feature subset is inconsistent if there exist at least two instances with same feature values but with different class labels. We compare inconsistency measure with other measures and study different search strategies such as exhaustive, complete, heuristic and random search that can be applied to this measure [16]. Mr. M. Senthil Kumar, Ms. V. Latha Jothi M.E in" A Fast Clustering Based Feature Subset Selection Using Affinity Propagation Algorithm"- Traditional approaches for clustering data are based on metric similarities, i.e., nonnegative, symmetric, and satisfying the triangle inequality measures using graph-based algorithm to replace this process a more recent approaches, like Affinity Propagation (AP) algorithm can be selected and also take input as general non metric similarities [1]. Priyanka M G in "Feature Subset Selection Algorithm over Multiple Dataset"- here a fast clustering based feature subset selection algorithm is used. The algorithm involves (i) removing irrelevant features, (ii) constructing clusters from the relevant features, and (iii) removing redundant features and selecting representative features. It is an effective way for reducing dimensionality. This FAST algorithm has advantages like efficiency and effectiveness. Efficiency concerns the time required to find a subset of features and effectiveness is related to the quality of the subset of features. It can be extended to use with multiple datasets [2]. Lei Yu, Huan Liu in" Efficient Feature Selection via Analysis of Relevance and Redundancy"- we show that feature relevance alone is insufficient for efficient feature selection of high-dimensional data. We define feature redundancy and propose to perform explicit redundancy analysis in feature selection. A new framework is introduced that decouples relevance analysis and redundancy analysis. We develop a correlation-based method for relevance and redundancy analysis, and conduct an empirical study of its efficiency and effectiveness comparing with representative methods [17]. Yanxia Zhang, Ali Luo, and Yongheng Zhao in" An automated classification algorithm for multi-wavelength data" we applied a kind of filter approach named Relief to select features from the multi-wavelength data. Then we put forward the naive Bayes classifier to classify the objects with the feature subsets and compare the results with and without feature selection, and those with and without adding weights to features. The result shows that the naive Bayes classifier based on Relief algorithms is robust and efficient to preselect AGN candidates [18]. N.Deepika, R.Saravana Kumar in" A Fast Clustering Based Flexible and Accurate Motif Detector Technique for High Dimensional Data", present an algorithm that uses FLAME as a building block and can mine combinations of simple approximate motifs under relaxed constraints. The approach we take in FLAME explores the space of all possible models. In order to carry out this exploration in an efficient way, we first construct two suffix trees: a suffix tree on the actual data set that contains counts in each node (called the data suffix tree), and a suffix tree on the set of all possible model strings (called the model suffix tree). To get effective and accurate motif detection [19].

3. EXISTING METHOD

The embedded methods incorporate feature selection as a part of the training process and are usually specific to given learning algorithms, and therefore may be more efficient than the other three categories. Traditional machine learning algorithms like decision trees or artificial neural networks are examples of embedded approaches. The wrapper methods use the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subsets, the accuracy of the learning algorithms is usually high. However, the generality of the selected features is limited and the computational complexity is large. The filter methods are independent of learning algorithms, with good generality. Their computational complexity is low, but the accuracy of the learning algorithms is not guaranteed. The hybrid methods are a combination of filter and wrapper methods by using a filter method to reduce search space that will be considered by the subsequent wrapper. They mainly focus on combining filter and wrapper methods to achieve the best possible performance with a particular learning algorithm with similar time complexity of the filter methods.

Disadvantages

1. The generality of the selected features is limited and the computational complexity is large.
2. Their computational complexity is low, but the accuracy of the learning algorithms is not guaranteed.

4. Proposed method

Feature subset selection can be viewed as the process of identifying and removing as many irrelevant and redundant features as possible. This is because irrelevant features do not contribute to the predictive accuracy and redundant features do not redound to getting a better predictor for that they provide mostly information which is already present in other feature(s). Of the many feature subset selection algorithms, some can effectively eliminate irrelevant features but fail to handle redundant features yet some of others can eliminate the irrelevant while taking care of the redundant features. Our proposed FAST algorithm falls into the second group. Traditionally, feature subset selection research has focused on searching for relevant features. A well-known example is Relief which weighs each feature according to its ability to discriminate instances under different targets

based on distance-based criteria function. However, Relief is ineffective at removing redundant features as two predictive but highly correlated features are likely both to be highly weighted. Relief-F extends Relief, enabling this method to work with noisy and incomplete data sets and to deal with multiclass problems, but still cannot identify redundant features.

Advantages:

1. Good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with each other.
2. The efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset.

4.1. User Module

In this module, Users are having authentication and security to access the detail which is presented in the ontology system. Before accessing or searching the details user should have the account in that otherwise they should register first.

4.2. Distributed Clustering

The Distributional clustering has been used to cluster words into groups based either on their participation in particular grammatical relations with other words by Pereira et al. or on the distribution of class labels associated with each word by Baker and McCallum . As distributional clustering of words are agglomerative in nature, and result in suboptimal word clusters and high computational cost, proposed a new information-theoretic divisive algorithm for word clustering and applied it to text classification. proposed to cluster features using a special metric of distance, and then makes use of the of the resulting cluster hierarchy to choose the most relevant attributes. Unfortunately, the cluster evaluation measure based on distance does not identify a feature subset that allows the classifiers to improve their original performance accuracy. Furthermore, even compared with other feature selection methods, the obtained accuracy is lower.

4.3. Subset Selection Algorithm

The Irrelevant features, along with redundant features, severely affect the accuracy of the learning machines. Thus, feature subset selection should be able to identify and remove as much of the irrelevant and redundant information as possible. Moreover, "good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other. Keeping these in mind, we develop a novel algorithm which can efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset.

4.4. Time Complexity

The major amount of work for this algorithm involves the computation of SU values for TR relevance and F-Correlation, which has linear complexity in terms of the number of instances in a given data set. The first part of the algorithm has a linear time complexity in terms of the number of features m. Assuming features are selected as relevant ones in the first part, when k ¼ only one feature is selected.

4.5. Flow chart:

The following Diagram shows the the flow chart for implementing the clustering based feature selection algorithm.

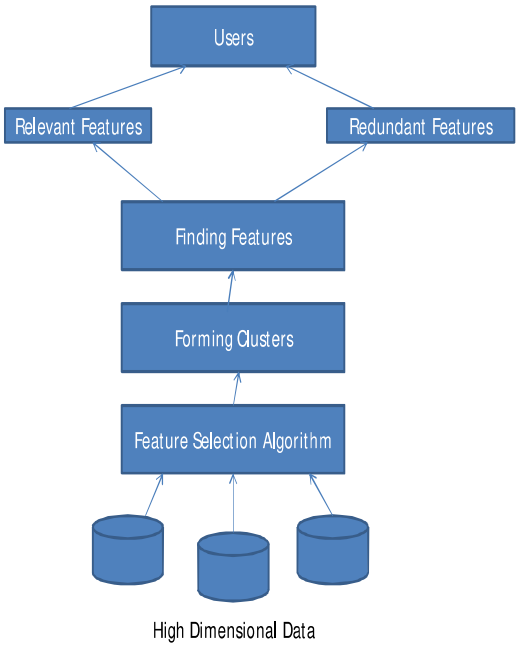
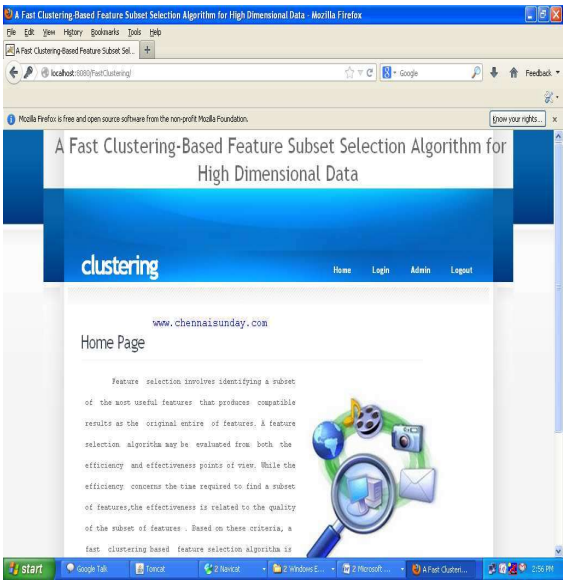
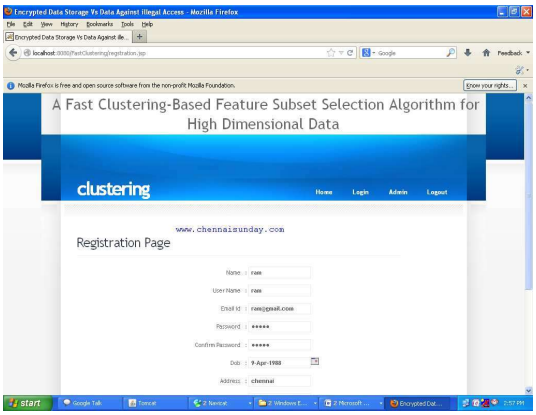
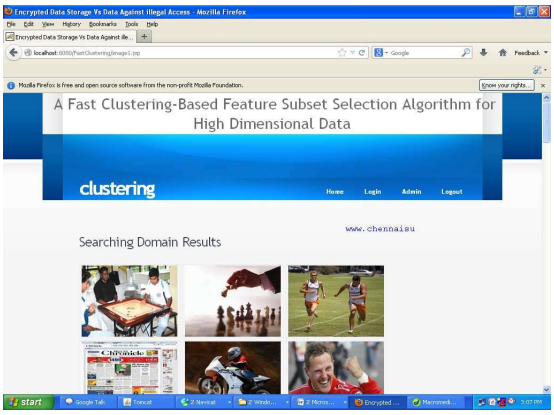


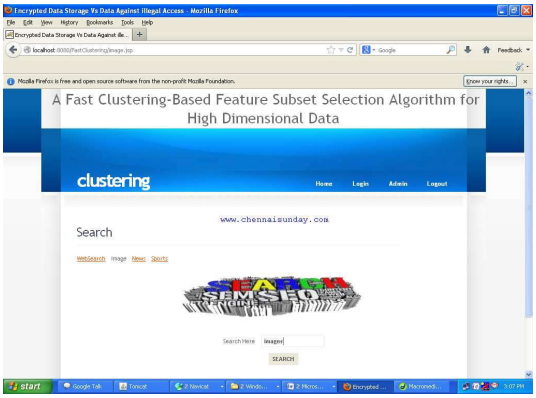
Figure 1 Flow Chart for Feature Selection

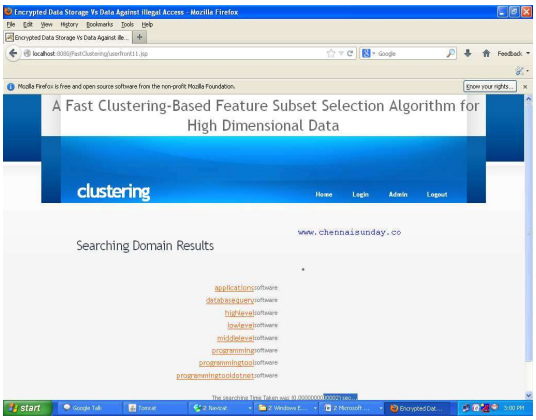
5. Implementation Results

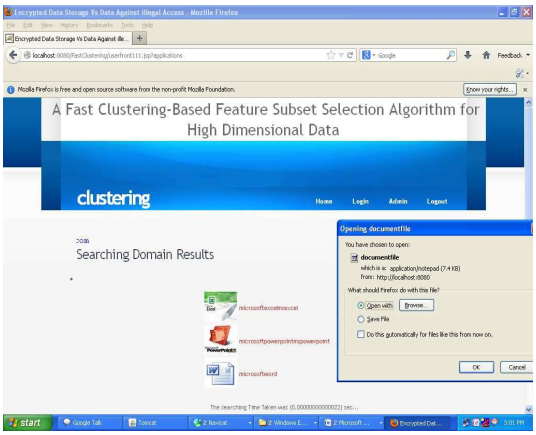












6. Conclusion

This paper explains about the data mining functionalities and also about the feature subset selection. In this we have explained different methods proposed for feature subset selection. The proposed method is used to extract the features based on clustering. This also provides the implementation details of the proposed algorithm. The implementation details include the modules User Module, Distributed Clustering, Subset Selection Algorithm.

References

[1] Mr. M. Senthil Kumar, Ms. V. Latha Jothi M.E,"A Fast Clustering Based Feature Subset Selection Using Affinity Propagation Algorithm", International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014 Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14).

[2] Priyanka M G "Feature Subset Selection Algorithm over Multiple Dataset" Proceedings of IRF International Conference, Goa, 16th March-2014, ISBN: 978-93-82702-65-8.

[3] M. Dash, H. Liu, Feature selection methods for classification, Intelligent Data Analysis: An Internat. J. 1 (3) (1997).

[4] H. Liu, H. Motoda (Eds.), Feature Extraction, Construction and Selection: A Data Mining Perspective, Kluwer Academic, Boston, MA, 1998.

[5] H. Liu, H. Motoda, Feature Selection for Knowledge Discovery and Data Mining, Kluwer Academic, Dordrecht, 1998.

[6] D.A. Bell, H. Wang, Amalism for relevance and its application in feature subset selection, Machine Learning 41 (2000) 175–195.

[7] S. Das, Filters, wrappers and a boosting-based hybrid for feature selection, in: Proceedings of the Eighteenth International Conference on Machine Learning (ICML), Williamstown, MA, 2001, pp. 74–81.

[8] M.A. Hall, Correlation-based feature selection for discrete and numeric class machine learning, in:

Proceedings of Seventeenth International Conference on Machine Learning (ICML), Stanford, CA, Morgan Kaufmann, San Mateo, CA, 2000, pp. 359–366.

[9] I. Inza, P. Larranaga, R. Etxeberria, B. Sierra, Feature subset selection by Bayesian network-based optimization, Artificial Intelligence 123 (2000) 157–184.

[10] R. Kohavi, G.H. John,Wrappers for feature subset selection, Artificial Intelligence 97 (1–2) (1997) 273–324.

[11] P. Soucy, G.W. Mineau, A simple feature selection method for text classification, in: Proceedings of IJCAI-01, Seattle, WA, 2001, pp. 897–903.

[12] Almuallim H. and Dietterich T.G., Learning boolean concepts in the presence of many irrelevant features, Artificial Intelligence, 69(1-2), pp 279-305,1994.

[13] Zheng Zhao and Huan Liu in "Searching for Interacting Features", ijcai07

[14] S.Swetha1, A.Harpika "A Novel Feature Subset Algorithm For High Dimensional Data" IJRRECS/October 2013/Volume-1/Issue-6/1295-1300 ISSN 2321-5461.

[15] T.Jaga Priya Vathana, C.Saravanabhavan , Dr.J.Vellingiri" A Survey On Feature Selection Algorithm For High Dimensional Data Using Fuzzy Logic" The International Journal Of Engineering And Science (IJES), Volume2, Issue 10, Pages 27- 38, 2013, ISSN (e): 2319 – 1813 ISSN (p): 2319 – 1805.

[16] Manoranjan Dash a,, Huan Liub "Consistency-based search in feature selection" Artificial Intelligence 151 (2003) 155–176.

[17] Lei Yu, Huan Liu," Efficient Feature Selection via Analysis of Relevance and Redundancy" Journal of Machine Learning Research 5 (2004) 1205–1224 Submitted 1/04; Revised 5/04; Published 10/04.

[18] Yanxia Zhang*a, Ali Luo*a, and Yongheng Zhao*a," An automated classification algorithm for multi-wavelength data" National Astronomical Observatories, Chinese Academy of Sciences, China.

[19] N.Deepika, R.Saravana Kumar," A Fast Clustering Based Flexible and Accurate Motif

Detector Technique for High Dimensional Data", International Journal of Innovations in Scientific and Engineering Research (IJISER) E-ISSN: 2347-971X P-ISSN: 2347-9728 Vol 1Issue 3 MAR 2014.

AUTHOR

K. Sujatha is pursuing her Ph.D. in Krishna University,Machilipatnam, A.P. She is interested doing research in data mining. Present she is carrying her research in Infrequent Pattern Mining. She has five international journal publications with high impact factors and indexing. She has two national journal publications. She attended for Two AICTE sponsored 2-week workshops. She is member in Indian Association of Engineers (IAE).She has a total of 10 years experience in teaching. She is working as an Associate Professor in CSE Department, Malineni Lakshmaiah Engineering College, Singaraya konda, Prakasam District. A.P.

M. Radha is Research Scholar in Data Mining. She is working as a Lecturer in CSE Department, Bapatla Women's Engineering College, Bapatla, Guntur District,Andhra Pradesh. She has a total of 12 Years Experience in Teaching. She has one International Journal Publication. She has participated in two National level conferences and attended many workshops. She attended for Two AICTE sponsored 2-week workshop on data mining. . She is member in Institution of Engineers (India) IEI.