J. Sagar Babu *et al*, International Journal of Computer Science and Mobile Applications,
Vol.2 Issue. 12, December- 2014, pg. 23-29 **ISSN: 2321-8363**

IMPLEMENTATION OF RELEVANT AND REDUNDANT FEATURES OF HIDDEN PREDICTIVE INFORMATION BASED ON DATA MINING

Mr.J.Sagar Babu[1]
Mr.R. Kiran Babu[2]
Associate Professor[1]
Assistant Professor[2]
Head, Department of Computer Science & Engineering[1]
Head, Department of Information Technology[2]
Princeton College of Engineering & Technology[12]
Ghatkesar[12]
Hyderabad.[12]

ABSTRACT— Clustering which tries to group a set of points into clusters such that points in the same cluster are more similar to each other than points in different clusters, under a particular similarity metric. In the generative clustering model, a parametric form of data generation is assumed, and the goal in the maximum likelihood formulation is to find the parameters that maximize the probability (likelihood) of generation of the data given the model. In the most general formulation, the number of clusters k is also considered to be an unknown parameter. Such a clustering formulation is called a "model selection" framework, since it has to choose the best value of k under which the clustering model fits the data. In clustering process, semi-supervised learning is a class of machine learning techniques that make use of both labeled and unlabeled data for training - typically a small amount of labeled data with a large amount of unlabeled data. Semi supervised learning falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data). Feature selection involves identifying a subset of the most useful features that produces compatible results as the original entire set of features. A feature selection algorithm may be evaluated from both the efficiency and effectiveness points of view. While the efficiency concerns the time required to find a subset of features, the effectiveness is related to the quality of the subset of features. 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 here we select more recent approaches, like Affinity Propagation (AP) algorithm can take as input also general.

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

I. INTRODUCTION

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. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive

information that experts may miss because it lies outside their expectations. Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

a) Massive data collection
b) Powerful multiprocessor computers
c) Data mining algorithms

The Scope of Data Mining

Data mining derives its name from the similarities between searching for valuable business information in a large database for example, finding linked products in gigabytes of store scanner data — and mining a mountain for a vein of valuablore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

A An Automated prediction of trends and behaviors:

Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data — quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.

B Automated discovery of previously unknown patterns:

Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

II. 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.

III. 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.

A 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.

B 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.

C Subset Selection Algorithm

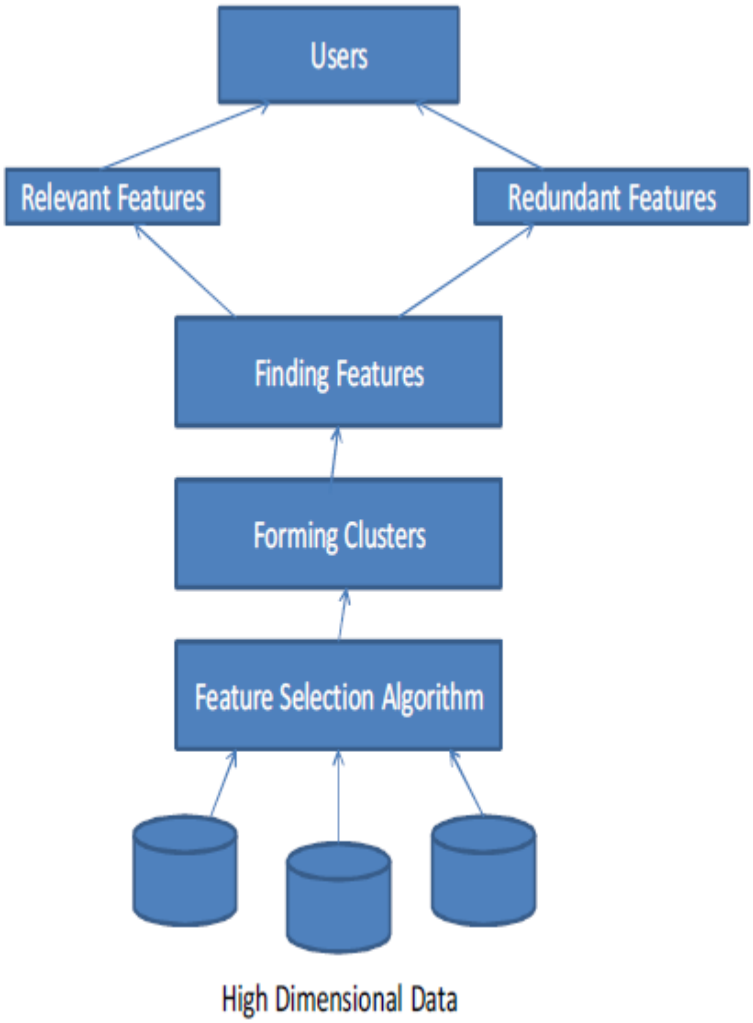
The Irrelevant features, along with redundant features, severely affect the accuracy of the learning machines. Thus, feature subset selection should be able to identify and remove as much of the irrelevant and redundant information as possible. Moreover, "good feature subsets contain features highly 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.

D TIME COMPLEXITY

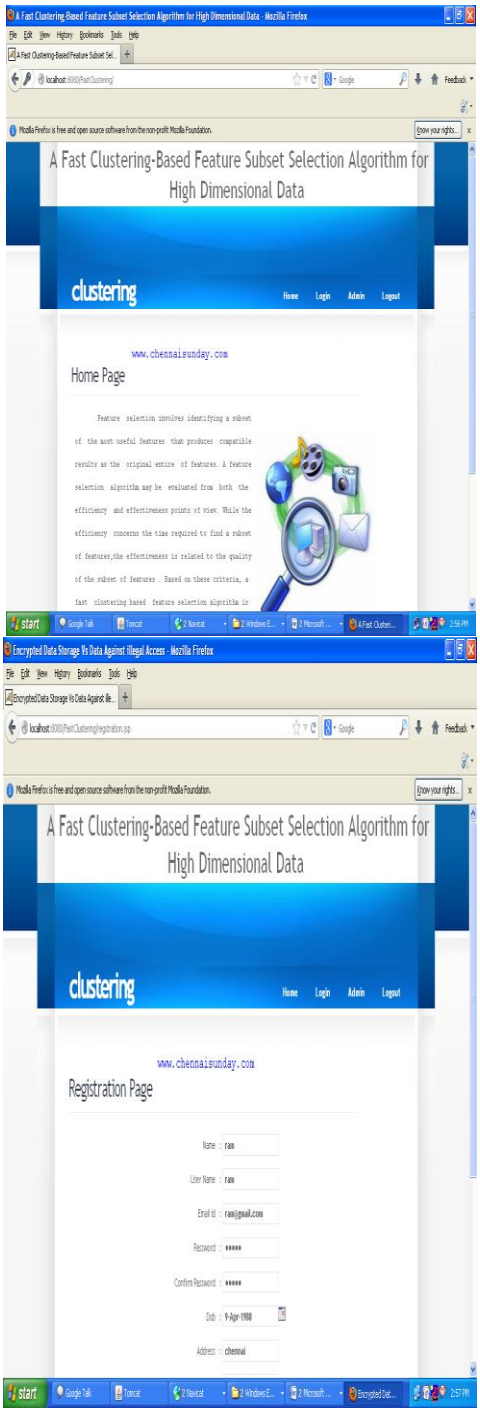
The major amount of work for this algorithm involves the computation of SU values for TR relevance and FCorrelation, which has linear complexity in terms of the number of instances in a given data set. The first part of the algorithm has a linear time complexity in terms of the number of features m.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

FLOW CHART:

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



IV. IMPLEMENTATION RESULTS



V. 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] Demsar J., Statistical comparison of classifiers over multiple data sets, J.Mach. Learn. Res., 7, pp 1-30, 2006.
[8] Fleuret F., Fast binary feature selection with conditional mutual Information,Journal of Machine Learning Research, 5, pp 1531-1555, 2004.
[9] Forman G., An extensive empirical study of feature selection metrics fortext classification, Journal of Machine Learning Research, 3, pp 1289-1305,2003.
[10] Garcia S and Herrera F., An extension on "Statistical Comparisons ofClassifiers over Multiple Data Sets" for all pairwise comparisons, J. Mach.Learn. Res., 9, pp 2677-2694, 2008.
[11] Guyon I. and Elisseeff A., An introduction to variable and feature selection,Journal of Machine Learning Research, 3, pp 1157-1182, 2003.
[12] Hall M.A., Correlation-Based Feature Selection for Discrete and NumericClass Machine Learning, In Proceedings of 17th International Conferenceon Machine Learning, pp 359-366, 200

Mr.J.Sagar Babu is Associate Professor & Head, Department of Computer science & engineering, Princeton College of Engineering & Technology, Ghatkesar, R.R-Dist, He has Working experience in teaching field since 2008.His qualification is B.Tech in Computer Science & Engineering from Dr.Samuel George Institute of Engineering & Technology, Markapur, Prakasam-Dist in 2005.M.Tech in Computer Science & Engineering from QIS College of Engineering & Technology, Ongole, Prakasam-Dist. Completed In 2008.

Mr.R.Kiran Babu is Asst. Professor & Head, Department of Information Technology, Princeton College of Engineering & Technology, Ghatkesar, R.R-Dist. He has Working experience in teaching field since 2012.His qualification is B.Tech in Information Technology from Prakasam Engineering College, Kandukur, Prakasam-Dist in 2007.M.Tech in Computer Science & Engineering from Princeton College of Engineering & Technology, Ghatkesar, R.R--Dist. Completed In 2012.

