ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 3, March 2014

Analysing the Efficiency of Data by Using Fast Clustering and Subset Selection Algorithm

R.Pavithra[1], J.Vinitha Grace[2], A.Arun Sethupathy Raja[3], V.Stalin[4], M.Ramakrishnan[5]

UG Scholar, Department of Computer science & Engineering, Sree Shakthi Engineering College, Coimbatore, Tamilnadu, India[1, 2, 3, 4]

Assistant Professor, Department of Computer science & Engineering, Sree Shakthi Engineering College, Coimbatore, Tamilnadu, India[5]

ABSTRACT: There are several algorithms applied to find the efficiency and effectiveness. Here we consider the efficiency as the time taken to retrieve the data's from the database and effectiveness is from the most datasets (or) subsets which are relevant to the users search. By using FAST algorithm we can retrieve the data's without the irrelevant features. Here the irrelevant features are carried out by means of various levels of the query input and the output the relevant information can be carried out in case of the subset selection and clustering methods. These can be formed in well equipped format and the time taken for retrieve the information will be short time and the Fast algorithm calculate the retrieval time of the data from the dataset. This algorithm formulates as per the data available in the dataset. In this paper, mainly focus about the micro array images which are not discussed in the previous work. By analyzing the efficiency of the proposed work and the existing work, the time taken to retrieve the data will be better in the proposed by removing all the irrelevant features which are gets analyzed.

KEYWORDS: Subset, Relevance, Fast Algorithm, Irrelevant features.

I. INTRODUCTION

The aim is to choose the relevant features with respect to user search; subset selection is an efficient way for removing irrelevant features (or) data. Here we use filter method as the feature subset selection method.

The subset selection is used in clustering of feature selection is the more efficient way. The known algorithms like FCBF, CFS are not efficient for high-dimensional data and feature selection.

The FAST algorithm also can be used with Minimum Spanning Tree. Using MST we can cluster the data's the important data's which are related data's which are related to the search is selected from each clustered datasets to from the selected datasets to form the selected datasets. The FAST algorithm has been tested with large datasets and high-dimensional data's. The FAST and CFS only reduces the time required to retrieve the datasets whereas FAST reduces the time required and also produces selected datasets.

II. LITERATURE SURVEY

A. An efficient approach to clustering in large multimedia databases with noise.

Several clustering algorithms can be applied to clustering in large multimedia databases. The effectiveness and efficiency of the existing algorithms, however is somewhat limited since clustering in multimedia databases requires clustering high-dimensional feature vectors and since multimedia databases often contains large amounts of noise. Using DENCLUE algorithm we can remove noise from the large databases. The main advantage includes, that the clustering can be done efficiently in high dimensional database.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 3, March 2014

B. Feature subset selection using the wrapper method: over fitting and dynamic search space topology

In the feature subset selection, a search for an optimal set of features is made using the induction algorithm as a black box. The best-first search is to find good feature subsets. The over fitting problems can be reduced by using the best first search. The relevant and optimal features can be easily selected in this method, and also the over fitting can be reduced.

C. A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining

Partitioning a large set of objects into homogeneous clusters is a fundamental operation in data mining. The k-means algorithm is best suited for implementing this operation because of its efficiency in clustering large data sets. Clustering methods partition a set of objects into clusters such that objects in the same cluster are more similar to each other than objects in different clusters. The most distinct characteristic of data mining is that it deals with very large data sets (gigabytes or even terabytes). This requires the algorithms used in data mining to be scalable. However, most algorithms currently used in data mining do not scale well when applied to very large data sets because they were initially developed for other applications than data mining which involve small data sets. We present a fast clustering algorithm used to cluster categorical data. The main advantages of this method is that, can easily categorize the objects and the dissimilar objects can be removed easily.

D. Fast and Effective Text Mining Using Linear-time Document Clustering

Clustering is a powerful technique for large-scale topic discovery from text. It involves two phases: first, feature extraction maps each document or record to a point in high-dimensional space, then clustering algorithms automatically group the points into a hierarchy of clusters. Document clustering helps tackle the information overload problem in several ways. One is exploration; the top level of a cluster hierarchy summarizes at a glance the contents of a document collection. Also the features are extracted efficiently.

E. Irrelevant Features and the Subset Selection Problem

We address the problem of ending a subset of features that allows a supervised induction algorithm to induce small high accuracy concepts. We examine notions of relevance and irrelevance_ and show that the definitions used in the machine learning literature do not adequately partition the features into useful categories of relevance. The features selected should depend not only on the features and the target concept_ but also on the induction algorithm. We describe a method for feature subset selection using cross validation that is applicable to any induction algorithm. In this the relevant features alone be extracted.

III. EXISTING SYSTEM

In the existing system they have used the CFS algorithm. It is available for small number of dataset and is very slow compared to FAST and subset selection algorithm. FCBF is very slow compared to FAST. In this major disadvantage follows that the Irrelevant features do not contribute to the predictive accuracy, not reduces data redundancy. Time consumption in data retrieval, CFS is not available on the 3 biggest datasets of the 35 datasets. It does not support the microarray data. In database, cannot able to download the existing only upload is possible.

IV. PROPOSED SYSTEM

The proposed system consists of various levels, in that the first is about the Fast algorithm which removes irrelevant features and redundant features. The feature subset selection algorithm helps in identifying the relevant datasets. The irrelevant features will not have any relation with the datasets. The redundant features can be separated from cluster and can be easily eliminated. The FAST has also a better runtime performance with high-dimensional data. The Subset selection algorithm (fig 1) invokes searching of the relevant datasets and cluster these

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 3, March 2014

datasets and hence by eliminating the relevant features. The qualities of datasets are also efficient which satisfies the users search (or) requirements. The Subsets are the sets containing another set that is important information which user needs during the data retrieval. Feature selection techniques provide three main benefits when constructing predictive models:

- improved model interpretability,
- shorter training times,
- Enhanced generalization by reducing over fitting.

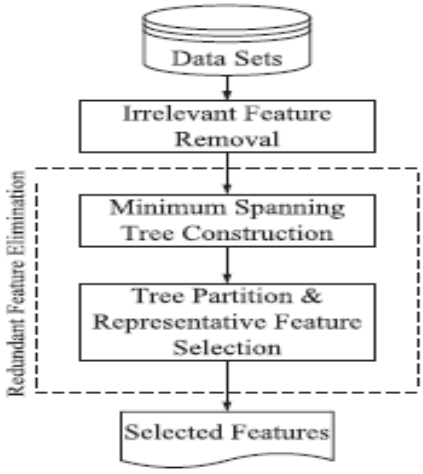

Fig 1: Subset selection Algorithm

Feature selection is also useful as part of the data analysis process, as it shows which features are important for prediction, and how these features are related.

Clustering is mainly used in grouping the datasets which are similar to the users search. The datasets which are irrelevant can be easily eliminated and redundant features inside the datasets are also removed. The clustering finally produces the selected datasets. The clustering uses MST for selecting the related datasets and finally the relevant datasets.

A **minimum spanning tree** (**MST**) or **minimum weight spanning tree** is then a spanning tree with weight less than or equal to the weight of every other spanning tree. More generally, any undirected graph (not necessarily connected) has a **minimum spanning forest**, which is a union of minimum spanning trees for its connected components.

It is the cluster analysis of data with anywhere from a few dozen to many thousands of dimensions. Such high-dimensional data spaces are often encountered in areas such as medicine, where DNA microarray technology can produce a large number of measurements at once, and the clustering of text documents, where, if a word-frequency vector is used, the number of dimensions equals the size of the dictionary.

V. EXPRIEMENTAL RESULT

In the user module, the 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. The Distributional clustering has been used to cluster words into groups based

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 3, March 2014

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. 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. The major amount of work for Algorithm 1 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.

VI. CONCLUSION

The overall function leads to the subset selection and FAST algorithm which involves, removing irrelevant features, constructing a minimum spanning tree from relative ones (clustering) and reducing data redundancy and also it reduces time consumption during data retrieval. It supports the microarray data in database; we can upload and download the data set from the database easily. Images can be downloaded from the database. Thus we have presented a FAST algorithm which involves removal of relevant features and selection of datasets along with the less time to retrieve the data from the databases. The identification of relevant data's is also very easy by using subset selection algorithm.

REFERENCES

1. Qinbao Song, Jingjie Ni and Guangtao Wang, "A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING VOL:25 NO:1 YEAR 2013.
2. H. Almuallim and T.G. Dietterich, "Algorithms for Identifying Relevant Features," Proc. Ninth Canadian Conf. Artificial Intelligence, pp. 38-45, 1992.
3. H. Almuallim and T.G. Dietterich, "Learning Boolean Concepts in the Presence of Many Irrelevant Features," Artificial Intelligence, vol. 69, nos. 1/2, pp. 279-305, 1994.
4. Arauzo-Azofra, J.M. Benitez, and J.L. Castro, "A Feature Set Measure Based on Relief," Proc. Fifth Int'l Conf. Recent Advances in Soft Computing, pp. 104-109, 2004.
5. L.D. Baker and A.K. McCallum, "Distributional Clustering of Words for Text Classification," Proc. 21st Ann. Int'l ACM SIGIR Conf. Research and Development in information Retrieval, pp. 96-103, 1998.
6. Christos Boutsidis, Michael W. Mahoney, Petros Drineas, "An Improved Approximation Algorithm for the Column Subset Selection Problem" S. Chen and J. Wigger," Fast Orthogonal Least Squares Algorithm for Efficient Subset Model Selection", IEEE TRANSACTIONS ON SIGNAL PROCESSING. VOL. 43. NO 1, JULY 1995 1713.
7. Charu C. Aggarwal Cecilia Procopiuc IBM T. J. Watson, "Fast Algorithms for Projected Clustering", Research Center Duke University Yorktown Heights, NY 10598 Durham, NC 27706.
8. Alexander Hinneburg, Daniel A.Keim," An efficient approach to clustering in Large Multimedia Databases with Noise" Institute of Computer Science, University of Halle, Germany.
9. Ron Kohavi and Dan Sommer_eld, "Feature Subset Selection Using the Wrapper Method: Over_tting and Dynamic Search Space Topology", Appears in the First International Conference on Knowledge Discovery and Data Mining (KDD-95)
10. Bjornar Larsen and Chinatsu Aone,"Fast and Effective Text Mining Using Linear-time Document Clustering" SRA International, Inc.
11. Zhexue Huang, "A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining",Cooperative Research Centre for Advanced Computational Systems.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 3, March 2014

12. Saowapak Sotthivirat and Jeffrey A. Fessler "Relaxed ordered-subset algorithm for penalizedlikelihood image restoration" Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan 48109 S. Sotthivirat and J. A. Fessler Vol. 20, No. 3/March 2003/J. Opt. Soc. Am. A 439.
13. Bin Zhang, Member, IEEE, and Sargur N. Srihari, Fellow, IEEE,"Fast k-Nearest Neighbor Classification Using Cluster-Based Trees", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 26, NO. 4, APRIL 2004.
14. Jihoon Yang and Vasant Honavar, "Feature Subset Selection Using A Genetic Algorithm Artificial Intelligence Research Group, Department of Computer Science, Iowa State University.

BIOGRAPHY

R.Pavithra is an Under Graduate Student in the Dept of Computer Science and Engineering, in Sree Sakthi Engineering College, Coimbatore, under Anna University. Area of Interest is Data Mining, Compiler Design, Database Management System and Web Scripting Languages.

J.Vinitha Grace is an Under Graduate Student in the Dept of Computer Science and Engineering, in Sree Sakthi Engineering College, Coimbatore, under Anna University. Area of Interest is Database Management System and Web Scripting Languages and Data Mining,

A.Arun Sethupathy Raja is an Under Graduate Student in the Dept of Computer Science and Engineering, in Sree Sakthi Engineering College of, Coimbatore, under Anna University. Area of Interest is Database Management System, Compiler Design and Data Mining.

V.Stalin is an Under Graduate Student in the Dept of Computer Science and Engineering, in Sree Sakthi Engineering College, Coimbatore, under Anna University. Area of Interest is Compiler Design and Data mining.

M.Ramakrishnan completed his PG in the Dept of Computer Science and Engineering, and currently working as the assistant professor in Sree Sakthi Engineering College, Coimbatore, under Anna University. He has guided many UG and PG student in his experience. His Area of Interest is Networking, Data Mining, Cloud Computing, Image Processing, Network Security, Compiler Design, Operating System, and Web Content Mining.