Diversifying Subset Feature With Ranking For High Dimensional Data

[1]Monalisa Lenka, [2]Priyanka Yadav, [3]Jyoti Kumari, [4]S.Venkata Lakshmi
[1,2]Dept. of Information Technology, GITAM University, AP, India
[3]Department of CS, GITAM University, AP, India
[4]Assistant Professor, Department of IT, GITAM University, AP, India

Abstract
Feature selection involves identifying a subset of the most representative features.Feature selection algorithm may be evaluated from both the efficiency and effectiveness points of view.Based on these criteria, a fast clustering-based feature selection algorithm (FAST) is proposed.The Feature Subset Selection generally works in two steps:Features are divided into clusters by using graph-theoretic clustering methods. The most representative feature that is strongly related to target class is selected.To ensure the efficiency of FAST, we adopt the efficient minimum-spanning tree (MST) clustering method.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
Feature Subset Selection, Fast Clustering-Based Feature Selection Algorithm,Minimum Spanning Tree, Cluster

I. Introduction

With the aim of choosing a subset of good features with respect to the target concepts, feature subset selection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility. Many feature subset selection methods have been proposed and can be divided into four broad categories: the Embedded, Wrapper, Filter, and Hybrid approaches.

The wrapper methods are computationally expensive and tend to over fit on small training sets. The filter methods, in addition to their generality, are usually a good choice when the number of features is very large. Thus, we will focus on the filter method in this paper. With respect to the filter feature selection methods, the application of cluster analysis has been demonstrated to be more effective than traditional feature selection algorithms.

In cluster analysis, graph-theoretic methods have been well studied and used in many applications. Their results have, sometimes, the best agreement with human performance. The general graph-theoretic clustering is simple: compute a neighborhood graph of instances, then delete any edge in the graph that is much longer/ shorter (according to some criterion) than its neighbors. The result is a forest and each tree in the forest represents a cluster. We apply graph-theoretic clustering methods to features. In particular, we adopt the minimum spanning tree (MST)-based clustering algorithms, because they do not assume that data points are grouped around centers or separated by a regular geometric curve and have been widely used in practice.

Based on the MST method, we propose a fast clustering based feature subset Selection algorithm (FAST). The FAST algorithm works in two steps. In the first step, features are divided into clusters by using graph-theoretic clustering methods. In the second step, the most representative feature that is strongly related to target classes is selected from each cluster to form the final subset of features. Features in different clusters are relatively independent; the clustering based strategy of FAST has a high probability of producing a subset of useful and independent features.

II. Overview

The rest of the paper is organized as follows: Section II contains literature review about related work. Section III contains description of the system design. Finally paper is concluded in the Section IV.

III. Literature Review

In [1] authors present an integrated approach to intelligent feature selection. They introduce a unifying platform which serves an intermediate step toward building an integrated system for intelligent feature selection and illustrate the idea through a preliminary system based on research. The unifying platform is one necessary step toward building an integrated system for intelligent feature selection. The ultimate goal for intelligent feature selection is to create an integrated system that will automatically recommend the most suitable algorithm to the user while hiding all technical details irrelevant to an application.

In [2] authors introduced a novel concept, predominant correlation, and propose a fast filter method which can identify relevant features as well as redundancy among relevant features without pair wise correlation analysis. The efficiency and effectiveness of our method is demonstrated through extensive comparisons with other methods using real-world data of high dimensionality.

In [3] authors proposed a new framework of efficient feature selection via relevance and redundancy analysis, and a correlation-based method which uses C correlation for relevance analysis and both C- and F-correlations for redundancy analysis. A new feature selection algorithm is implemented and evaluated through extensive experiments comparing with three representative feature selection algorithms. The feature selection results are further verified by two different learning algorithms.

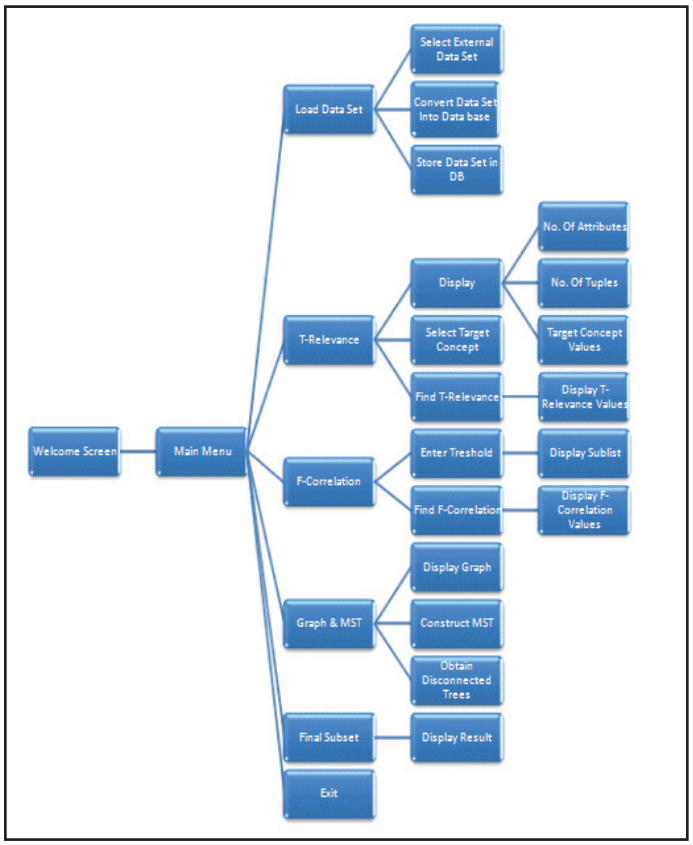
In [4] authors present a novel concept predominant correlation and propose a new algorithm that can effectively select good features based on correlation analysis with less than quadratic time complexity. A correlation based measure used in this approach. Two approaches classical linear correlation and Information theory are

used. The algorithm used is FCBF, Fast correlation based filter.
In [5] authors introduced the importance of removing redundant genes in sample classification and pointed out the necessity of studying feature redundancy. And proposed a redundancy based filter method with two desirable properties. It does not require the selection of any threshold in determining feature relevance or redundancy and it combines sequential forward selection with elimination, which substantially reduces the number of feature pairs to be evaluated in redundancy analysis.
In [6] authors generalized the ensemble approach for feature selection. So that it can be used in conjunction with many subset evaluation techniques, and search algorithms. A recently developed heuristic algorithm harmony search is employed to demonstrate the approaches. The key advantage of FSE is that the performance of the feature selection procedure is no longer depended upon one selected subset, making this technique potentially more flexible and robust in dealing with high dimensional and large datasets.
In [7] authors proposed a framework for feature selection which avoids implicitly handling feature redundancy and turns to efficient elimination of redundant features via explicitly handling feature redundancy. This framework composed of two steps: analysis of relevance determines the subset of relevant features by removing irrelevant ones, and analysis of redundancy determines and eliminates redundant features fromrelevant ones and thus produces the final subset. A novel clustering based feature subset selection algorithm for highdimensional data.
In [8] authors present an optimization tool for attribute selection. This paper formulates and validates a method for selecting optimal attribute subset based on correlation using Genetic algorithm, where genetic algorithm used as optimal search tool for selecting subset of attributes. Correlation between the attributes will decide the fitness of individual to take part in mating. Fitness function for GA is a simple function, which assigns a rank to individual attribute on the basis of correlation coefficients. Since strongly correlated attributes cannot be the part of DW together, only those attributes shall be fit to take part in the crossover operations that are having lower correlation coefficients.
In [9] authors identify the problems associated with clustering of gene expression data, using traditional clustering methods, mainly due to the high dimensionality of the data involved. For this reason, subspace clustering techniques can be used to uncover the complex relationships found in data since they evaluate features only on a subset of the data. Differentiating between the nearest and the farthest neighbors becomes extremely difficult in high dimensional data spaces. Hence a thoughtful choice of the proximity measure has to be made to ensure the effectiveness of a clustering technique.

III. Proposed System

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.The generality of the selected features is limited and the computational complexity is large.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.
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 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. 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.

A. System Flow Chart



B. High –Dimensional Data

The specificity of modern data mining is that huge amounts of data are considered. Compared to just a few years ago, we now use daily huge databases in medical research, imaging, financial analysis, and many other domains. Not only new fields are open to data analysis, but also it becomes easier, and cheaper, to collect large amounts of data. One of the problems related to this tremendous evolution is the fact that analyzing these data becomes more and more difficult, and requires new, more adapted techniques than those used in the past. A main concern in that direction is the dimensionality of data.

The Knowledge Discovery in Databases (KDD):

KDD process is commonly defined with the following stages:

1. Selection
2. Pre-processing
3. Transformation
4. Data mining
5. Evaluation

C. Data Pre-processing

Data pre-processing is an important step in the mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects.

Data gathering methods are often loosely controlled, resulting in out-of-range values (e.g., Income:-100), impossible data combinations, missing values, etc.

Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis. If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time. Data preprocessing includes cleaning, integration & transformation, Data Reduction, etc.

D. Data Reduction

Data reduction techniques can be applied to obtain a reduced representation of the. Data set that is much smaller in volume, yet closely maintains the integrity of the original

data. That is, mining on the reduced data set should be more efficient yet produce the same (or almost the same) analytical results.

Strategies for data reduction include the following:

1. Data cube aggregation
2. Feature subset selection
3. Dimensionality reduction
4. Numerosity reduction
5. Concept hierarchy generation

E. Feature Subset Selection

Feature subset selection is one of the data reduction technique. It involves identifying a subset of the most useful features that produces compatible results as the original entire set of features.

In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset 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 selection techniques are often used in domains where there are many features and comparatively few samples (or data points). The archetypal case is the use of feature selection in analyzing DNA microarrays, where there are many thousands of features, and a few tens to hundreds of samples.

With the aim of choosing a subset of good features with respect to the target concepts, feature subset selection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility.

Many feature subset selection methods have been proposed and studied for machine learning applications. They can be divided into four broad categories: the Embedded, Wrapper, Filter, and Hybrid approaches. 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 e x amp l e s of embedded approaches.

The wrapper methods use the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subset s, 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 t he 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 t he 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.

The wrapper methods are computationally expensive and tend to over fit on small training sets. The filter methods, in addition to their generality, are usually a good choice when the number of features is very large. 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. For n attributes, there are 2n possible subsets. An exhaustive search for the optimal subset of attributes can be prohibitively expensive, especially as n and the number of data classes increase. Therefore, heuristic methods that explore a reduced search space are commonly used for attribute subset selection. These methods are typically greedy in that, while searching through attribute space, they always make what looks to be the best choice at the time. Their strategy is to make a locally optimal choice in the hope that this will lead to a globally Optimal solution.

Such greedy methods are effective in practice and may come close to estimating an optimal solution. The "best" attributes are typically determined using tests of statistical significance, which assume that the attributes are independent of one another. Many other attribute evaluation measures can be used, such as the information gain measure used in building decision trees for classification. 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 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. However, along with irrelevant features, redundant features also affect the speed and accuracy of learning algorithms, and thus should be eliminated take into consideration the redundant features. Fast Correlation Based Filter Solution (FCBF) is a fast filter method which can identify relevant features as well as redundancy among relevant features without pair wise correlation analysis, iteratively picks features which maximize their mutual information with the class to predict, conditionally to the response of any feature already picked. . 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 .enabling this method to work with noisy and incomplete data sets and to deal with multiclass problems, but still cannot identify redundant features.

F. Problem Description

Feature subset selection can be viewed as the process of identifying and removing as many irrelevant and redundant features as possible. Along with irrelevant features, redundant features also affect the speed and accuracy of learning algorithms. 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.

Problems that are facing while using the existing system

1. Selected features contain redundant features.
2. No pair wise correlation
3. No Ranking concept is applied to Feature subset.
4. No Prediction Results are possible.
5. All require more CPU time and memory to identify the features that are required.

Fast Clustering based Feature Subset Selection Algorithm (FAST)

By implementing the FAST algorithm, the feature selection is based on Clustering-approach. In this, features are grouped into clusters by using graph-theoretic clustering methods and the most representative feature that is strongly related to target classes is selected from each cluster to form a subset of features. Feature Subset Selection Framework contains four phases:

1. Irrelevant Feature Elimination
2. Redundant Feature Removal
3. Rank For Feature Subset
4. Classification against Feature Subset

Algorithm Implementation

1. Removal of Irrelevant features
2. T-Relevance, F-Correlation calculation
3. MST construction
4. Relevant feature calculation

Advantages of Proposed System

1. Good feature subsets contain features highly correlated with 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.
3. Generally all the six algorithms achieve significant reduction of dimensionality by selecting only a small portion of the original features.
4. The null hypothesis of the Friedman test is that all the feature selection algorithms are equivalent in terms of runtime.

IV. 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 dataredundancy 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 datasetsalong 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.

V. Acknowledgement

We want to thanks our guide and our department of Information Technology, GITAM UNIVERSITY for their support and guidance.

References

[1] Huan Liu,"Toward Integrating Feature Selection Algorithms for Classification and Clustering", IEEE transactions on knowledge and data engineering, Vol. 17, No. 4, April 2005.

[2] Lei Yu, Huan Liu,"Efficiently Handling Feature Redundancy in High Dimensional Data", ACM, August 27, 2003.

[3] Lei Yu, Huan Liu,"Efficient Feature Selection via Analysis of Relevance and Redundancy", Journal of Machine Learning Research, 5, pp. 1205–1224, 2004.

[4] Lei Yu, Huan Liu,"Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution", Proceedings of the TwentiethInternational Conference on Machine Learning, 2003.

[5] Lei Yu, Huan Liu,"Redundancy Based Feature Selection for Microarray Data", ACM, August 2004.

[6] QiangShen, RenDiao, Pan Su,"Feature Selection Ensemble", Turing-100, Vol.10, pp. 289–306, 2012.

[7] Qinbao Song, Jingjie Ni, Guangtao Wang,"A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data", IEEEtransactions on knowledge and data engineering, Vol. 25, No. 1, 2013.

[8] Rajdev Tiwari, Manu Pratap Singh,"Correlation-based Attribute Selection using Genetic Algorithm", International Journal of ComputerApplications, Vol. 4, No. 8, August 2010.

[9] SajidNagi, Jugal K, Dhruba K. Bhattacharyya, Kalita, "A Preview on Subspace Clustering of High Dimensional Data", International journal ofcomputers & technology, Vol. 6, No. 3, May 2013.

[10] Butterworth R., Piatetsky-Shapiro G., Simovici D.A., On Feature Se-lection through Clustering", In Proceedings of the Fifth IEEE international Conference on Data Mining, pp. 581-584, 2005.

[11] Chikhi S., Benhammada S.,"ReliefMSS: A variation on a feature ranking Relief algorithm", Int. J. Bus. Intell. Data Min. 4(3/4), pp. 375-390, 2009.
[12] Cohen W.,"Fast Effective Rule Induction", In Proc. 12th international Conf. Machine Learning (ICML'95), pp. 115-123, 1995.

Monalisa Lenka received her B.Tech in Electronic And Telecommunication Engineering from Vignan Institute Of Technology And Management affiliated to Biju Patnaik University Of Technolgy(BPUT), Odisha, India,2012. She is currently pursuing her M.Tech at GITAM UNIVERSITY, Andhra Pradesh, India, in Information Technology. Her area of interest includes Implementation of LAR protocol using MANETS and Data mining- Clustering of Data.

Priyanka Yadav received her B.Tech from Avanthi Institute of Engineering And Technology affiliated to Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh, India, 2012. She is currently pursuing her M.Tech at GITAM UNIVERSITY, Andhra Pradesh, India, both in Information Technology. Her area of interest includes Implementation of LAR protocol using MANETS, Information security, Big Data and Data mining- Clustering of Data.

Jyoti Kumari received her B.Tech in Computer Science and Engineering from Greater Noida Institute Technology affiliated to Gautam Buddha Technological University, India, 2012. She is currently pursuing her M.Tech at Gitam UNIVERSITY, Andhra Pradesh, India, in Computer Science and Technology. Her area of interest includes Implementation of LAR protocol using MANETS, Information security and Data mining- Clustering of Data.

Mrs. S. Venkata Lakshmi M.Tech in Information Technology from Andhra University. Pursuing Ph.d from Andhra University, Asst.Prof in GITAM University. Over 6 years of teaching experience with GITAM University, 2 years of industry experience as a software engineer, handled courses for B.Tech, and M.Tech. Research areas include Data Mining and Databases.