Feature Subset Selection By Using Graph Based Clustering

K . Divya
M.Tech 2nd year
Dept. Of CSE
AITS
Tirupati, India

B . Ramana Reddy
Asst.Professor
Dept. Of CSE
AITS
Tirupati, India

***Abstract*—This paper describes selection of Feature Subset by using graph based clustering method. Feature selection is a process of identifying a subset of the most representative features means most useful features that features produces same result as that result produced by the entire set of original features. A feature selection algorithm can be evaluated in terms of Efficiency and Effectiveness. Efficiency is related to the time required to find a subset of features and Effectiveness is related to the quality of the subset of features. By considering all these criteria a fast clustering based feature subset selection algorithm is implemented. This FAST algorithm works in two steps, in first step all features are divided into clusters but the features in different clusters must be independent and in second step select the most representative features means strongly related to target classes from each cluster. In this algorithm for clustering we use graph based clustering method. To ensure the efficiency of FAST algorithm we use minimum spanning trees (MST) graph based clustering method.**

***Keywords—Feature* Selection, Distributed Clustering, filter method, graph based clustering, Time complexity.**

I. INTRODUCTION

Feature subset selection can be viewed as a preprocessing step. The main aim of Feature Selection is identifying subset of good features with respect to target class. Feature selection has been very effective in reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility. For machine learning applications we have many feature subset selection methods. They are: Embedded, Wrapper, Filter and Hybrid methods.

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 method. The wrapper model requires one predetermined learning algorithm in feature selection Based on the predictive accuracy of the learning algorithm, the wrapper method evaluate the ‘‘goodness” of the selected feature subset directly, which should intuitively yield better performance In spite of the good performance, the wrapper methods have limited applications due to the high computational complexity involved. Filter methods are independent of any learning algorithms; Filter methods typically make use of all the training data when selecting a subset of features. Their computational complexity is low, but the accuracy of the learning algorithms is not guaranteed. Hybrid method combines the advantages of both Filter and Wrapper methods, and avoids the high computational complexity of wrapper methods. 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.

Along with the filter feature selection methods we use the cluster analysis for selecting subset of the features. So it will be more effective than traditional feature selection algorithms.

In cluster analysis use graph-theoretic methods because 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. The best known graph-theoretic clustering algorithm is based on the construction of the minimal spanning Tree (MST) of the data, because they do not assume that the data points are grouped around centers or separated by a regular geometric curve.

By considering all these criteria, we propose a fast clustering based feature subset selection algorithm (FAST). This FAST algorithm works in two steps, in first step all features are divided into clusters by using graph theoretic clustering methods but the features in different clusters must be independent and in second step select the most representative features means strongly related to target classes from each cluster to form final subset of the features. The clustering-based strategy of FAST has a high probability of producing a subset of useful and independent features. The experimental results show that, compared with other five different types of feature subset selection algorithms, the proposed algorithm not only reduces the number of features, but also improves the performances of the four well known different types of classifiers. Good feature subset is one that contains features highly correlated with the target, yet uncorrelated with each other. For all the above mentioned terms this system is a fast filter method which can identify relevant features as well as redundancy among relevant features without pair wise correlation analysis, and iteratively picks features which maximize their mutual information with the class to predict, conditionally to the response of any feature already picked. Different from these algorithms, our proposed FAST algorithm employs clustering based method to choose features

II. EXISTING SYSTEM

In the past approach there are several algorithms which illustrate how to maintain the data into the database and how to retrieve it faster, but the problem here is no one cares about the database maintenance with ease manner and safe methodology. A Distortion algorithm, which creates an individual area for each and every word from the already selected transactional database, those are collectively called as dataset, which will be suitable for a set of particular words, but it will be problematic for the set of records. A Blocking algorithm make propagation to the above problem, and reduce the problems occurred in the existing distortion algorithm, but here also having the problem called data overflow, once the user get confused then they can never get the data back.

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 area 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.

A. DRAWBACKS OF EXISTING SYSTEM

- Lacks speed
- Security Issues
- Performance Related Issues
- 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.

So the focus of our new system is to enhance the throughput for any basis to eliminate the data security lacks there in and make a newer system prominent handler for handling data in an efficient manner.

III. PROPOSED SYSTEM

Feature selection is an effective way for identifying and removing as many irrelevant features and redundant features. Because (i) Irrelevant features do not provide the predictive accuracy (ii) Redundant features do not contribute to getting a better predictor for that they provide information which is already present in other feature For classification, Feature selection is used to find an “optimal” subset of relevant features such that the overall accuracy of classification is increased while the data size is reduced and the comprehensibility is improved. Feature selection methods consider two important issues: evaluation of a candidate feature subset and search through the feature space. So main aim of feature subset selection algorithm is removing of both irrelevant and redundant features. But some algorithms effectively eliminate the irrelevant features but fail to handle redundant features yet some of other algorithms can eliminate the irrelevant features while taking care of the redundant features. Our proposed algorithm falls into the second category.

Traditional feature subset selection algorithms focused on searching for relevant features. A well known algorithm that relies on relevance evaluation is In RELIEF, a subset of features in not directly selected, but rather each feature is given a relevance weighting indicating its level of relevance to the class label. It is important to note that this method is ineffective at removing redundant features as two predictive but highly correlated features are both likely to be given high Relevance weightings. Relief-F extends Relief, enabling this method to work with noisy and incomplete data sets and to deal with multi-class problems, but still cannot identify redundant features.

Learning algorithm speed and accuracy is affected by both irrelevant and redundant features so we should eliminate both irrelevant and redundant features.

Our Proposed FAST algorithm remove the both irrelevant features and redundant features and this algorithm uses minimum spanning tree based method to cluster features because MST does not assume that data points are grouped around centers or separated by a regular geometric curve and another important one is our proposed FAST does not limit to some specific types of data.

A. ADVANTAGES OF PROPOSED SYSTEM

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

B. FAST ALGORITHM

The Feature subset selection should be able to identify and remove as much of the irrelevant and redundant information as possible.”Good feature subsets contain features highly correlated with the class and uncorrelated with each other’s”.

Framework for FAST algorithm:

Framework contains two components they are: Irrelevant feature removal and Redundant feature elimination. The former obtains features relevant to the target concept by eliminating irrelevant ones, and the latter removes redundant features from relevant ones via choosing representatives from different feature clusters, and thus produces the final subset.

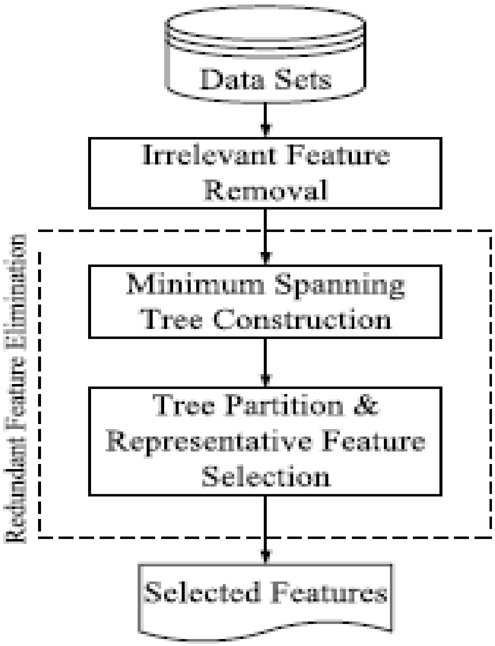
Data Sets
Irrelevant Feature Removal
Minimum Spanning Tree Construction
Tree Partition & Representative Feature Selection
Redundant Feature Elimination
Selected Features

Figure 1: Framework **of the Feature Selection Algorithm**

.

Dataset contain all features, FAST algorithm remove the irrelevant features, irrelevant features removal is a straightforward one the right relevance measure is defined or selected but redundant feature elimination is not a straightforward it is three step procedure: (i) the construction of the minimum spanning tree (MST) from a weighted complete graph; (ii) the partitioning of the MST into a forest with each tree representing a cluster(iii) the selection of representative features from the clusters.

SYSTEM ***ARCHITECTURE***

A. ***User Module***

In User module, It checks authentication of the Users and provides security to access the details which is presented in the system. Before accessing the data and searching the details user proves their authentication by providing account details.

B. Distributed ***Clustering***

In Distributional clustering, cluster the words into groups based 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 aggregated, and result in suboptimal word clusters so reducing high computational cost, For that proposed a new information-theoretic divisive algorithm for word clustering and applied this algorithm for text classification .This divisive algorithm cluster features using by using special metric of distance, and then makes use of the of the resulting cluster hierarchy to choose the most relevant features. But unfortunately, the cluster evaluation measured based on the distance does not identify a feature subset that improves the classifier original performance accuracy. Furthermore, even compared this one with other feature selection methods, the obtained performance accuracy is lower.

C. ***Subset Selection Algorithm***

Both the Irrelevant features and redundant features, affected the accuracy and speed of the learning machines. Thus, feature subset selection methods should be able to identify and remove as many irrelevant and redundant features as possible. Moreover, “good feature subsets always contain the features that highly correlated with the class, yet uncorrelated with each other. Keeping these in mind, we proposed a novel algorithm which can efficiently and effectively deal with both irrelevant and redundant features, and provide a good feature subset.

D. ***Feature*** *Selection* ***Algorithm***

Feature selection methods are subset of the more general field of feature extraction. Feature extraction method creates new features from functions of the entire set of original features, whereas feature selection method returns a subset of the features. Feature selection techniques used in many domains where there are many features and comparatively few samples (or data points).

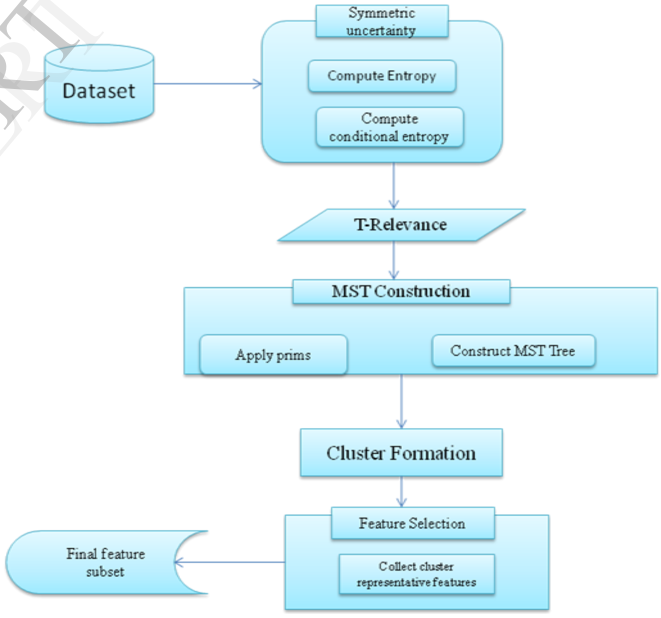

Figure 2: System **Architecture**

ALGORITHM ANALYSIS***:***

FAST algorithm logically works in three steps:

1. Remove Irrelevant features :Calculate T-relevance it means the relevance between the feature $A_i \epsilon A$ and the target concept C and denoted by $SU(A_i,C)$.If the T-relevance of the feature is less than the predetermined threshold value then remove that features and considered that feature is irrelevant feature.

2. Construct Minimum spanning tree by using Relevant features: Calculate F-correlation means the correlation between A_i and A_j features and denoted by SU (A_i, Aj)
3. Partitioning the MST and gathered Representative features.

Inputs: Entire Dataset: DS *(A1, A2... An, C)* - the given data set
θ - Predetermined threshold value i.e. the T-Relevance threshold.
Output: F*S* - selected feature subset.

```
// ***** Step1: Irrelevant Feature Removal ****
for i = 1 to n do
    T-Relevance = SU (Ai, C)
    if T-Relevance > θ then
        FS = FS ∪ { Ai};
//**** Step2: Minimum Spanning Tree Construction ****
G = NULL; //G is a completed graph
for each pair of features {A′i, A′k} ⊂ FS do
        F-Correlation = SU {A′i, A′k}
        Add A′i and/or A′k to G and use F-Correlation is the weight of the corresponding edge.
minSpanTree = Prim (G); //Using Prim Algorithm to generate the
Minimum spanning tree
//****Step3: Tree Partition and Representative Feature Selection****
Forest = minSpanTree
For each edge Eij ∈ Forest do
if SU {A′i, A′k}<( SU (Ai’, C) ∧ SU (Ak’, C)) then
Forest = Forest − Eij
FS = { }
For each tree Ti ∈ Forest do
 Select Ar (i.e. Ar is feature that having maximum T-Relevance)

FS = FS ∪{Ar}

return FS
```

CONCLUSION

In this paper, we have presented a Graph based clustering feature subset selection algorithm for high dimensional data. This algorithm involves three steps 1) Removing irrelevant features from entire Dataset, 2) constructing a minimum spanning tree from relevant features, and 3) partitioning the MST and selecting most representative features. In the proposed algorithm, a cluster consists of features and considers each cluster as a single feature, so dimensionality is drastically reduced. Generally, proposed algorithm produces the best proportion of selected features, the best runtime, and the best classification accuracy and the Win/Draw/Loss record.

The purpose of cluster analysis has been established to be more effective than feature selection algorithms. To evaluate the performance of the proposed algorithm and compare it with those of the five well-known feature selection algorithms FCBF, CFS, Consist, and FOCUS-SF on the publicly available image, microarray, and text data from the four different aspects of the proportion of selected features, runtime, classification accuracy of a given classifier, and the Win/Draw/Loss record. Generally, the proposed algorithm obtained the best proportion of selected features, the best runtime, and the best classification accuracy for Naive, and RIPPER, and the second best classification accuracy for IB1. The Win/Draw/Loss records confirmed the conclusions. We also found that FAST obtains the rank of 1 for microarray data, the rank of 2 for text data, and the rank of 3 for image data in terms of classification accuracy of the four different types of classifiers, and CFS is a good alternative. At the same time, FCBF is a good alternative for image and text data. Moreover, Consist, and FOCUS-SF are alternatives for text data.

For the future work, we plan to explore the entire Fast algorithm in hands with association rule implementation gives flexible results to users, like removing irrelevant features from the Original Subset, and constructing a minimum spanning tree from the relative subset whatever present in the data store. By partitioning the minimum spanning tree we can easily identify the text representation from the features. Association Rule Mining gives ultimate dataset with header representation as well as FAST algorithm with applied K-Means strategy provides efficient data management and faster performance. The revealing regulation set is significantly smaller than the association rule set, in particular when the minimum support is small. The proposed work has characterized the associations between the revealing regulation set and the non-redundant association rule set, and discovered that the enlightening regulation set is a subset of the non-redundant association rule set.

REFERENCES

[1]. Dash M. and Liu H., Feature Selection for Classification, Intelligent Data Analysis, 1(3), pp 131-156, 1997.

[2]. Molina L.C., Belanche L. and Nebot A., Feature selection algorithms: A survey and experimental evaluation, in Proc. IEEE Int. Conf. Data Mining, pp 306-313, 2002.

[3]. Das S., Filters, wrappers and a boosting-based hybrid for feature Selection, In Proceedings of the Eighteenth International Conference on Machine Learning, pp 74-81, 2001.

[4]. John G.H., Kohavi R. and Pfleger K., Irrelevant Features and the Subset Selection Problem, In the Proceedings of the Eleventh International Conference on Machine Learning, pp 121-129, 1994.

[5]. Baker L.D. and McCallum A.K., Distributional clustering of words for text classification, In Proceedings of the 21st Annual international ACM SIGIR Conference on Research and Development in information Retrieval, pp 96-103,1998.