MJRET
Open Access
Multidisciplinary Journal of Research in Engineering and Technology
Journal homepage: www.mjret.in
ISSN:2348 - 6953

IMPLEMENTATION OF CLUSTERING-BASED FEATURE SUBSET SELECTION ALGORITHM-FAST

Vinod Wagh[1],Pankaj Bemberkar[2],Sonam Nikade[3],Mahendra Naradhania[4]
Student, Department of Computer Engineering, University of Pune
Imperial College of Engineering and Research,
Pune, India.
[1]yashrajpatil77@gmail.com,[2]pankaj.bemberkar@gmail.com,
[3]nikade91@gmail.com,[4]naradhaniam@gmail.com

Abstract- *A FAST Algorithm produces the subset of most important features from the available set of features. Working of the FAST is done in two steps. In the primary step, the features are divided into clusters by using graphical method. In the secondary step, the most representative's features are selected from each cluster. The feature selection algorithm is implemented from both point of views, among that one is the efficiency which is nothing but the time required to find subsets of the features and another is effectiveness which is related to the quality of the subset of features. When we apply the FAST algorithm on microarray data or any text data then it will not only give required subsets of features but also improves the performance. Feature section means to identify a required and most useful data and that gives the only required features from the databases.*

Keywords- *Clustering, filter method, subset selection, graph-based clustering.*

1. INTRODUCTION

In the FAST Algorithm, we are choosing a subset of good feature with respect to the target classes then we are removing immaterial features from the hall entire set of original features. For the selection of a good subset of features with respect to the database, selecting the best features is an effective way for removing irrelevant data, reducing dimensionality and for better results. The extraction of hidden predictive information from large databases is a powerful new approach with great potential to help industries to focus on their databases. There are many feature selection methods have been implemented and applied on the various applications. They are Hybrid, Wrapper, Filter and Embedded methods. The other methods are experimentally expensive to implement than filter method. The filter method is usually a better one in the case of features is very large. Thus we are going to implement the

filter method in this paper. When any kind of business data was first stored on computers that improves a continuous data access that allows users to get the data in real time. In the case of filter feature selection methods, the clustering application has to demonstrate to be more reliable than conventional feature selection algorithms. For reducing the dimensionality of text data the distributed clustering of words can be used. Our proposed FAST algorithm uses minimum spanning tree based method to cluster features which can be derived from other clustering based algorithms.

The general graph-theoretic clustering is uncomplicated. We have to compute an adjacent graph of instances, then by deleting any edge in the graph that is shorter than its neighbors. The output can be in the form of a cluster. In this research paper, we concern graph theoretic clustering method on the features. Here assuming the minimum spanning tree (MST) based on clustering algorithm. By applying the MST method, Fast clustering based feature Selection algorithm (FAST) is proposed.

2. RELATED WORK

Selection of features subset is the method of capturing and removing a lot of unrelated or unnecessary features as probable. As per many features selection approaches, some of them can successfully removes irrelevant features but not able to handle redundant features [6], [8]. Our proposed FAST algorithm not only eliminate irrelevant data but also able to handle the redundant features. With the irrelevant features, redundant features also affect the accuracy and performance of learning algorithms. Some of the examples like CMIM [4] that taken into consider the redundant features. CMIM [4] takes the features which maximize the mutual data with the class to predict, to the response of any features that are already picked. Our proposed FAST algorithm applies clustering based method to choose features which varies from these algorithms.

Along with immaterial features, unwanted features can change the speed and accuracy of learning algorithms, and thus it could be extracted as well [5], [7], [10]. CFS [9] is also one example that capture into the consideration of unnecessary features. CFS [9] is achieved by assuming a good subset of features greatly correlated with the actual target, yet uncorrelated with each other.

Feature selection involves recognizing a subset of maximum of helpful features that produces attuned results as the unique set of features. The FAST algorithm can be implemented from mutually efficiency and effectiveness points of view. Distributed clustering has been applied onto cluster words into the distribution of class labels related with each word by Baker and MacCallum [4]. To select the features of spectral data the hierarchical clustering has been used. Van Dijk [3] proposed a hybrid filter subset selection algorithm for regression. In the context of word classification hierarchical clustering has been adopted.

Our proposed FAST algorithm is different than these hierarchical clustering based algorithms and it make use of minimum spanning tree based technique to cluster the features. In the meantime, it does not supposed to be data points are grouped near to the centers. The proposed FAST does not limit to some exact or specific types of data.

For the clustering of many numbers of features there are different ways such as embedded method which is usually specific to learning algorithms hence they are more efficient than

the other three categories .Some examples are: traditional machine learning algorithm such as decision tree or artificial neural networks. Second method is a wrapper method which use predetermined learning algorithm to identify the goodness of the selected feature subsets, accuracy of the learning algorithm is basically high but in this case the generation of selected feature are limited and computational complexity is large.

Third method is the filter method which is independent of the learning algorithm, and they have good generality. The computational complexity of the filter method is low, but accuracy of the learning algorithm is not guaranteed. Final method is the hybrid method which is the combination of filter and wrapper method, which is proposed to achieve best possible performance and to reduce search space of the learning algorithm, similarly time complexity of the filter method.

3. DISADVANTAGES OF EXISTING SYSTEM

1) The generality of the selected feature is limited and the computational complexity is large in wrapper method.
2) In filter the computational complexity is low, but the generation of the learning algorithm is not guaranteed.

4. FEATURE SUBSET SELECTION

The below figure shows the control flow of the FAST system. How the clustering is done and its formation and then way of feature subset selection is clear from the Fig.1

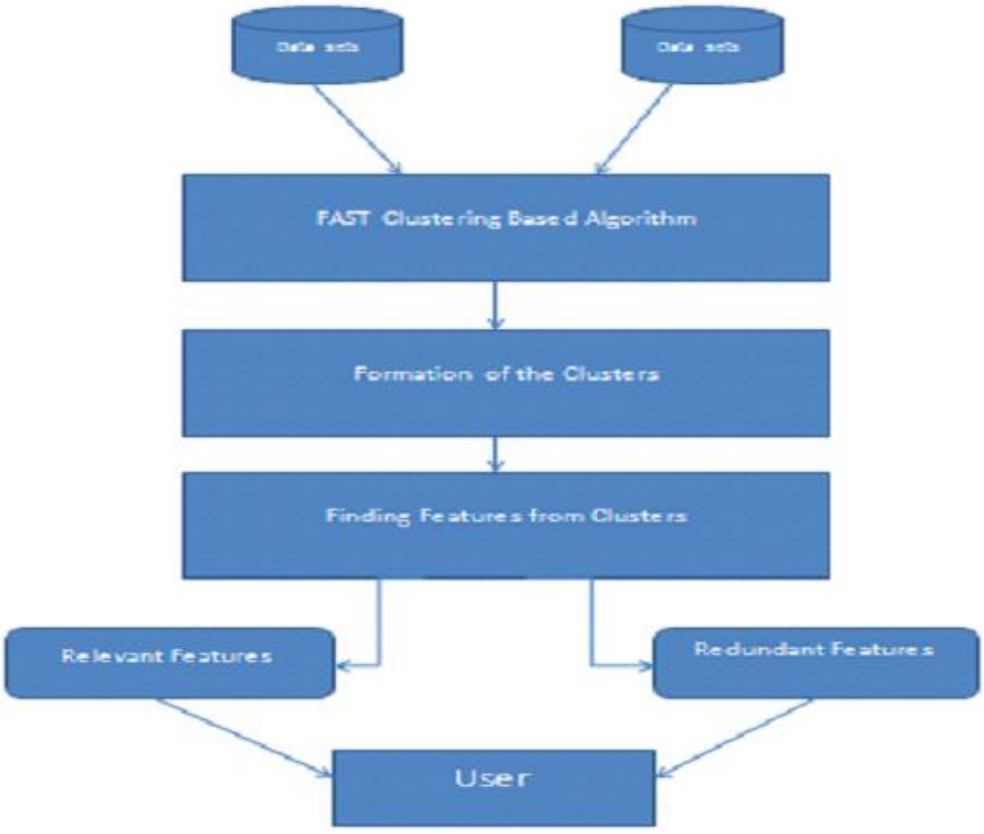

Fig.1 Control Flow of the FAST System

5. PROPOSED SYSTEM

We are going to implement the (FAST) Fast clustering based feature Selection algorithm to overcome the disadvantages of the existing system. Feature subset selection is the process of identifying and removing irrelevant and redundant features as many as possible. It is seen that many feature subset selection algorithms are fail to handle redundant features, and some other can eliminate the irrelevant features with the care of redundant features. FAST also comes under the same category.

Traditional feature subset selection research was focused on searching for relevant features, and that have some disadvantages, and that overcome by using FAST algorithm in our proposed system.

5.1 ADVANTAGES OF THE PROPOSED SYSTEM

1) Subsets of good feature contain features which are highly correlated with target class, and others are uncorrelated with each other.
2) The efficiency and effectiveness both are deal with irrelevant and redundant features, and obtain a good feature subset.

5.2 FRAMEWORK

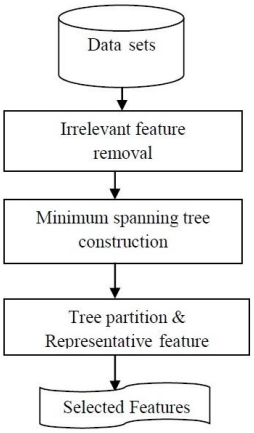

Fig.2 System Flow

5.3 DEFINITIONS AND MATHEMATICS MODULE

Symmetric Uncertainty is measured from mutual information by normalizing it to the entropies of the feature values or feature values and target class. According to the Zhao and Liu, the symmetric uncertainty as calculated of correlation between either two features or one feature and one target class.

The symmetric uncertainty is defined as follows

$$SU(X, Y) = 2*Gain(X/Y)/ (H(X) +H(Y)).$$

Where

H(X) is the entropy of a discrete random variable X.

Gain(X, Y) is the amount by which entropy of Y decrease.

$$\text{Gain}(X/Y) = H(X) - H(X/Y)$$

$$= H(Y) - H(Y/X)$$

Where H(X/Y) is the conditional entropy.

To calculate the Gain, we need to find the entropy and condition entropy.

$$H(X) = -\sum_{x \in X} p(x) log_2 p(x)$$

$$H(X|Y) = -\sum_{y \epsilon Y} p(y) \sum_{x \epsilon X} p(x|y)\, log_2 p(x|y)$$

p(x) is the probabilities density function for all values of X, p(x/y) is the condition probabilities density function.

a) **T-Relevance calculation:** The symmetric uncertainty between one feature and one target class is called as T-Relevance. The relevance between one feature Fi € F and one class C is referred to the T-Relevance of Fi and C.

T-Relevance = SU (Fi, C)

b) **F-Correlation calculation**: The symmetric uncertainty between any pair of features is called as F-Correlation. The Correlation between any pair of features Fi € F and Fj € F (i! =j) is referred to the F-Correlation of Fi and Fj.

F-Correlation = SU (Fi, Fj)

5.4 ALGORITHM AND ANALYSIS

The FAST algorithm logically divided into three steps

i) Removal immaterial features

ii) Construct a Minimum Spanning Tree (MST).

iii) Tree Partition and representative feature selection.

First Step: In this step, we have input data set D (F1, F2, F3…….Fn) and target class C.

We calculate the T-Relevance SU(Fi, C) value for every features Fi € F. If the T- Relevance are greater than predefined threshold Θ then we selected these features S = S U {Fi}.

Second Step: In the second step, first we determined the F-Correlation SU (Fi , Fj) value for each pair of features Fi and Fj then we seeing Fi and Fj as vertices and SU(Fi ,Fj) weight of edge between Fi and Fj. The weighted complete graph is constructed. The graph is undirected. The complete graph G has k vertices and

k(k-1)/2 edges. We construct a MST, which all vertices are connected but the sum of the weight of edge is the minimum, we are using Kruskal's algorithm.

Third Step: In the third step, we are removing the edge E= {(Fi, Fj)} whose weighted are smaller than both of the T-Relevance of Fi and the T-Relevance of Fj. After removing the result in two disconnected tree T1 and T2.

Suppose the MST shown in Fig.2 is calculated from complete graph G. first we check all six edges, and then who's weighted of the edge is smaller than T-Relevance of Fi and Fj. In Fig.2 the weight of edge SU (F0, F4) =0.3 is smaller than SU (F0, C) =0.5 and SU (F4, C) =0.7 then we are remove the edge SU (F0, F4). After remove the edge, the MST is divided into two clusters denoted as P (T1) and P (T2) is called as clustering.

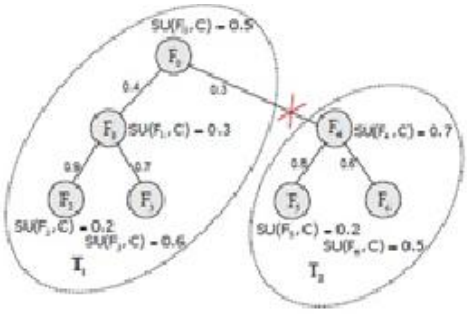

Fig.3 Example of clustering step

5.5 TIME COMPLEXITY

The amounts of time for working of algorithm are following

1) In the first step, the time complexity is O(n) because it is linear.
2) In the second step, the time complexity is O(k^2) because time complexity of Kruskal's algorithm is O(k^2).
3) In the third step, the time complexity is O(n) because it is also linear.
   Then the time complexity of the FAST Algorithm is O(m+k^2) when 1 < k < = m.

6. CONCLUSION

In this paper, we have compared the working of the proposed algorithm with some of the well-known feature selection algorithms like FCBF, CFS on the publically available microarray and text data from the four different aspects of the proportion of the selected features, runtime, and classification accuracy of a given classifier. We have presented a novel clustering based feature subset selection algorithm for large databases or high dimensional data.

Our proposed FAST algorithm involves 1) removing immaterial features, 2) constructing a MST from related ones, 3) selecting representative features. Each cluster is consisting of some class of features. Hence, each cluster is considered as a single feature and thus dimensionality is drastically reduced. For future work, we plan to explore different types of datasets like image for correlation measures and to study some feature space.

7. REFERENCES

[1] *H. Almuallim and T.G. Dietterich, are given "Algorithms for Identifying Relevant Features," In Proceedings of the 9th Canadian Conference of AI, pp 38-45,1992.*

[2] *H. Almuallim and T.G. Dietterich., are given "Learning Boolean Concepts in the Presence of Irrelevant Features and Artificial Intelligence and vol. 69 nos.1/2,pp 279-305, 1994.*

[3] *Baker L.D. and A.K.McCallum, are given "Distributional Clustering of Words for Text Classification", In Proceedings of the 21st Annual international ACM SIGIR Conference on Research and Development in information Retrieval and pp 96-103, 1998.*

[4] *D.A.Bell and H. Wang, are given "A Formalism for Relevance and Its Application in Feature Subset Selection", 41(2), pp 175-195, 2000.*

[5] *R. Butterworth, G. Piatetsky-Shapiro and D.A.Simovici, are given "On Feature Selection Through Clustering", In Proceedings of the Fifth IEEE international Conference on Data Mining, pp 581-584, 2005.*

[6] *C. Cardie has given "Used Decision Trees Improving Case-Based learning", In Proc. Of Tenth International Conference on the Machine Learning, pp 25-32,1993.*

[7] *P.Chanda, Y.Cho, A.Zhang and Ramanathan M., are given "Mining of Attribute Interactions Using Information Theoretic Metrics," In Proceedings of IEEE International Conference of Data Mining Workshops, pp 350-355, 2009.*

[8] *W. Cohen, has given "Fast Effective Rule Induction", In Procedure.12th international Conference on Machine Learning (ICML'95), pp 115-123, 1995.*

[9] *M.Dash and H.Lliu, are given "Feature Selection for Classification", Intelligent Data Analysis, 1(3), pp 131-156, 1997.*

[10] *M.Dash and H.Lliu and H.Motoda, are given " Consistency based feature Selection", Proceedings of the Fourth Pacific Asian Conference on Knowledge Discovery And Data Mining, pp 98-109, 2000.*

[11] *J.Demsar, has given "Statistical comparison of classifiers over multiple data sets", J.Mach, Learn. Res., 7 and pp 1-30, 2006.*

[12] *Fleuret F. has given"Fast binary feature selection with conditional mutual Information," In the Journal of the Machine Learning Research, in 2004.*