IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 3, May 2014.
www.ijiset.com
ISSN 2348 – 7968

Negation of Immaterial and Duplicate Features in High Dimensional Data Using Clustering

Ranjani A[1], Jayashree G[2]
[1] **Computer Science and Engineering,Saveetha Engineering College,,Chennai, TamilNadu, India**
[2] **Computer Science and Engineering,Saveetha Engineering College,Chennai, TamilNadu, India**

Abstract
Feature Selection means finding a subset that is very useful to produce compatible results as the original set of features. A feature Selection algorithm can be concerned from the efficiency and effectiveness of features obtained were the efficiency is calculated from the time required to find the subset and effectiveness is concerned with quality of the subset of features. Based on this criteria, a new clustering based feature selection algorithm is proposed. The algorithm works by two steps. Features are divided into clusters by using clustering methods in the first step, the representative features that is more related to the target classes is selected from each cluster in the second step. This forms a small subset of features that is more related to the target class by eliminating the irrelevant and redundant features. Features formed from different cluster are relatively independent. In order to ensure the efficiency of our new clustering technique minimum Spanning tree is adopted. Compared with other feature selection algorithms with respect to four different types of classifiers before and after feature selection our new technique provides improved performance and relatively small subset that is most related to the target class.

Index Terms –Feature Selection , Clustering, Target classes.

I. Introduction

The aim of dimensionality reduction is choosing the subset of good features that is most related to the target class. The former feature Selection algorithms are proposed and studied for machine learning. They can be categorized as the Embedded, Filtering, Wrapper, and Hybrid approaches.

Examples of Embedded approaches are decision trees or artificial neural networks. This method involves feature selection as a training process and it is specific to the learning algorithm. The wrapper method is used to determine the goodness of the selected subset of features[3]. The accuracy of the algorithm is usually high, the generality of the selected features is limited and the computational complexity is large. The filter methods have good generality and are not dependent on the learning algorithm, the computational complexity is low, but the accuracy is not guaranteed. The hybrid methods works with the combination of filter and wrapper methods. The wrapper methods are more expensive. The filter methods are more advantageous than other methods because the generality of the selected features are good and the filter methods are a good choice when the number of features is very large.

With respect to the filter feature selection methods, the combination of cluster analysis has been experimented and found to be more effective than the traditional feature selection method. The cluster formation is simple: Compute a graph with number of edges , then delete any edge in the graph that is more longer/shorter than its neighbors. The minimum spanning tree is adopted based on the clustering algorithm. A new clustering Technique is proposed, it works by two steps, Features are divided into clusters by using clustering methods in the first step, the representative features that is more related to the target classes is selected from each cluster in the second step. This forms a small subset of features that is more related to the target class by eliminating the irrelevant and redundant features.

2. RELATED WORK

Feature subset selection is viewed as the process of finding and eliminating irrelevant and redundant features as much as possible.

This is because (i) Features that are irrelevant to the target class do not contribute to the predictive accuracy [7] (ii) redundant features or duplicate features do not provide a better predictor since they provide the information which is already present in other features. Many feature selection algorithm focuses on eliminating irrelevant features but fail to handle the redundant features[6]. The proposed clustering technique efficiently handles both the irrelevant and redundant features.

Relief works are based on the distance criteria[8]. it weighs each feature according to the ability of discriminating instances under different concepts. Relief is not recommended because it is not effective in handling redundant features. Relief is extended as Relief-F which works with noisy and incomplete data sets, but still it is not efficient in indentifying the redundant features. The time and quality of features is affected with the presence of irrelevant and redundant features. Hierarchical clustering is employed in word selection in the task of text classification. Distributional clustering works by the participation of words in grammatical relations. However they are agglomerative and ends up with high computational cost.

3 FEATURE SUBSET SELECTION

3.1 Proposed System

The learning machines are affected by the presence of irrelevant and redundant features from the accuracy points of view. Therefore our algorithm should be efficient enough to accurately identify and eliminate the irrelevant and redundant features with time and quality of the feature concerned.

We develop a novel framework which is composed of two components. The first component called irrelevant feature removal obtains the features which are relevant to the target class thereby eliminating irrelevant features. The second component called the redundancy elimination will eliminate all the redundant features. This framework of two connected

ISSN 2348 – 7968

components produces reduced number of final subset. The working of the first component is easier if the correct measure is selected to identify the relevancy. The time taken to function the second component is bit complicated.

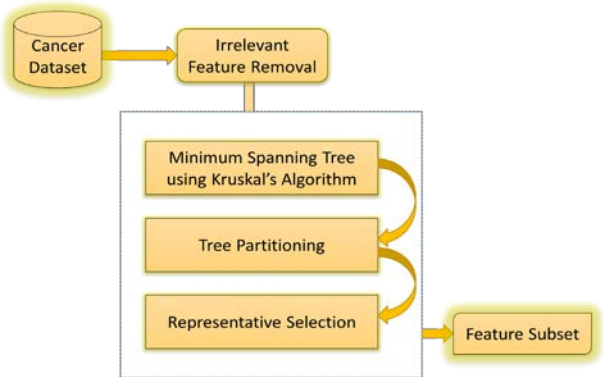

Fig 1 Proposed Data Flow Diagram

In order to precisely explain the functionality of the second component, it is carried out in three steps. First step involves the development of the minimum spanning tree, second step involves the partitioning of the constructed minimum spanning tree into group of cluster. Thirdly , the representative feature is selected from each cluster. The traditional definitions of irrelevant and redundant feature removal is explained below.

Definition 1(Relevant feature): A Feature F_i said to be relevant feature if it contributes more to the target class. Otherwise the feature F_i is irrelevant to the target class. It should be eliminated.

Definition 2(Redundant feature): Features F_i and F_j are said to be redundant if the values of the redundant features are completely correlated with each other.

Thus Feature Relevance can be defined by feature-target class correlation and Feature Redundancy can be defined by the correlation between features.

Definition 3(Symmetric uncertainty): It is defined as correlation between two features or a feature and target class.

Symmetric uncertainty can be defined with entropy conditional entropy and the information gain.

Definition 4(T-Relevance): The correlation between the feature F_i that belongs to the set S and the target class C is called as T-Relevance.

Definition 5(F-correlation): The dependence between any pair of features F_i and F_j is called F-correlation.

3.2 Algorithm

The proposed Algorithm basically includes three steps (i) removing the irrelevant features by calculating the T-relevance (ii) Developing the minimum spanning tree (iii)partitioning the minimum spanning tree and selecting representative features.

3.3 Analysis

For a dataset D with m features and the last column representing the class C. we compute the T-relevance i.e the symmetric uncertainty between the feature and the target class as a first step. Features whose values are greater than a threshold values are selected. The threshold value is taken as a correlation measure to eliminate the irrelevant feature. The features whose values are lesser than the threshold value are not selected for further process.

In the second step the F-correlation, which provides the correlation between any pair of features is calculated. Then the features are viewed as vertices and the Symmetric uncertainty between the features are viewed as the weight of the edge between a pair of feature. This will create a graph G that is undirected. The graph reflects the correlation between all the target relevant features. Graph G has k vertices and k(k-1)/2 edges. The graph will be heavily denser and the edges will have different edges. Having the graph G the minimum spanning tree is constructed which connects all the vertices such that the sum of the weights of the edges is minimum using the known algorithms. In the third step, the edges whose weights are smaller than both of the T-relevance is removed from the graph. The Proposed system flow diagram depicted below represents the steps of t-relevance calculation through entropy, condition entropy and information gain. Then the F-correlation is calculation through feature relation between the relevant feature. The minimum Spanning tree is constructed to eliminate the edges with more weights in case of high dimensional data.

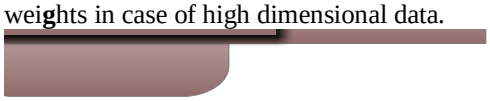

Fig 2 Proposed Architecture

ISSN 2348 – 7968

4 EXPERIMENTAL PROCEDURE

To evaluate the performance of our proposed new clustering algorithm and compare with other feature selection algorithms in a very simple way ,the experimental study is setup as follows.

(1) The proposed new algorithm is compared with different types of feature selection algorithm. Some of them evaluate the features individually. Some of them are based on subset evaluation. Such a subset consists of features highly correlated with target concept. The new clustering technique for feature selection heuristically set a threshold to be the t-relevance value to eliminate irrelevant features and produces a feature subset.

(2) Different types of classification algorithms are used to classify data sets before and after the feature selection. Naive Bayes algorithm utilizes the independence between the features and even then provides excellent results. The instance based learning algorithm is a closest nearest adjacent algorithm, it classifies features by taking the class of the closest associated vectors in the training set. It is the simplest among the algorithms used.

The new clustering technique is evaluated against the following metrics (i) the probability of the selected features (ii) the time required to obtain the feature subset (iii) the accuracy of classified feature subset (iv) the number of dataset for the proposed clustering technique achieves better ,equal and worst performance than other five feature selection algorithms

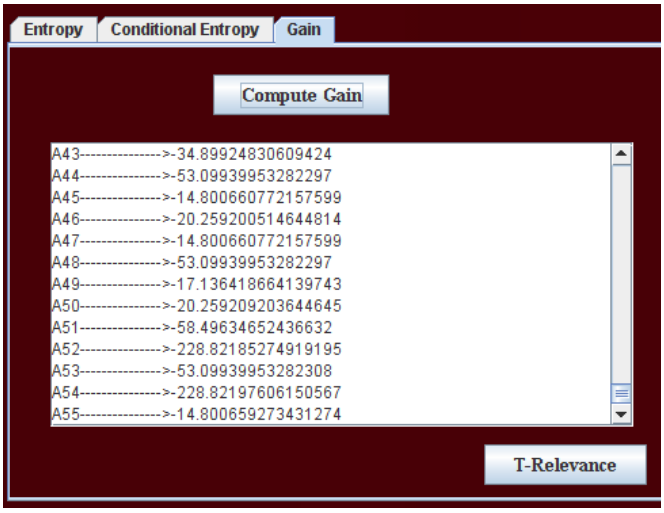

Fig 3 Result

5 PERFORMANCE EVALUATION

In order to use the data in a best way and obtain good results traverse evaluation strategy is used. That is, for multiple data set, different feature subset selection algorithm and each classification algorithm, the traverse evaluation strategy is iterated 5 times with each time the order of instances of the dataset being randomized. This is due to many of the feature selection algorithms exhibiting ordering effects. The ordering effect may sometime drastically improve or decrease the performance. When the input is randomized the ordering effect may be reduced.

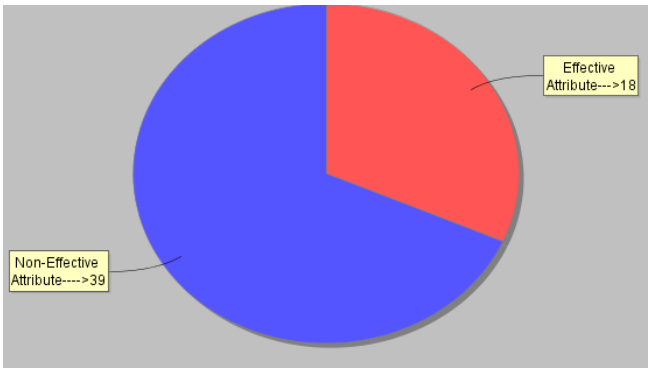

Fig 4 Evaluation Result

In the experiment ,for each feature selection algorithm, the feature subset and the corresponding runtime with each dataset is obtained. For each classification algorithm, a different classification accuracy is obtained.

6 RESULTS AND ANALYSIS

(1) Generally the algorithms for feature selection obtains drastic reduction in dimensionality by selecting only a small portion of the original features. The new clustering technique achieves the best proportion of selected features.

(2) The proportion of selected features for image data of each algorithm has improved compared with the corresponding proportion of selected features on the given dataset.

(3) For microarray data, the selected feature subset has been improved by most of the feature subset selection algorithm. The proposed new clustering technique employs still better methods to obtain good results.

(4) The proportion of selected features for text data of the new clustering technique achieves better results.

ISSN 2348 – 7968

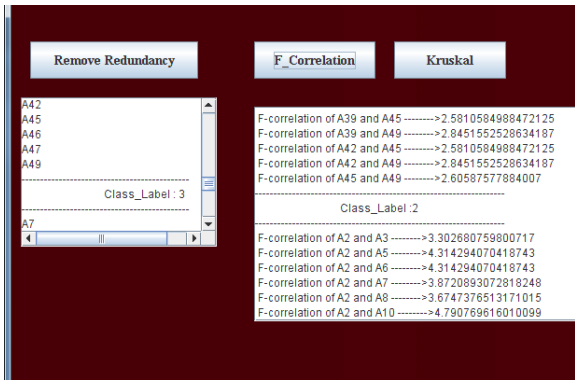

Fig 5. F-correlation Calculation

It is clear that the new clustering technique provides much better results for text, image and microarray data. when compared with other techniques of feature selection ,the new technique obtains a rank of 1. For all types of data the new technique ranks first and it is undisputed first choice. Microarray data is handled efficient because of the characteristics of the dataset itself and the property of the proposed algorithm.

The nature of the microarray data is that it is of small sample size which cannot fit in the training data. In the presence more number of features , researchers found that it is common that not all the features are informative since they may contain irrelevant and redundant features with respect to target concept. Therefore for successful sample classification only a small subset of discriminative features are selected.

6.1 Sensitivity Analysis

The proposed clustering technique also has a parameter called threshold like other feature selection algorithms which is the feature relevance measure. The classification accuracy changes more with different threshold value.

In order to find which parameter value could result in best classification accuracy for a specific classification problem with a given classifier the traverse evaluation strategy is used to demonstrate the accuracy changes with value of the parameter.

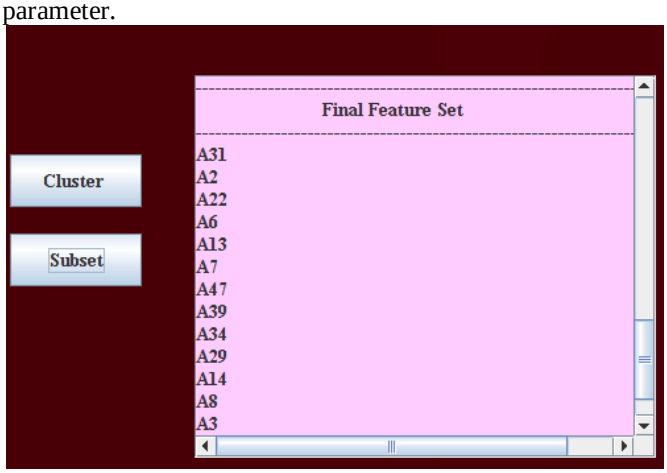

Fig 6 Reduced Subset

7 CONCLUSION

In this paper a new clustering based feature subset selection algorithm for multidimensional data is presented. The algorithm involves (i) removing features that are not relevant to the target class (ii) constructing a minimum spanning tree from the relevant features , and (iii) partitioning the minimum spanning tree and selecting the representative features. Since the representative features only cannot provide the exact prediction of the class label, some of the features which provide more probability with the class label which forms a subset are also chosen.

It is found that the newly proposed clustering technique performs well when compares with other feature selection algorithms, it is not specific to any particular type of data and obtains 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 different types of classifiers. For the future work, different types of correlation measures and some properties of the feature space can be studied.

REFERENCES

[1] Almuallim H and Dietterich T.G., Algorithms for identifying relevant Features, In Proceedings of the 9th Canadian conference on AI, pp 38-45,1992

[2] Almuallim H and Dietterich T.G., Learning Boolean concepts in the presence of many irrelevant features, Artificial Intelligence,69(1-2),pp 279-305,1994.

[3] Dash M. and Liu H., Feature selection for classification, Intelligent data Analysis, 1(3),pp 131-156,1997.

[4] Fleuret F., Fast binary feature selection with conditional mutual information ,journal of Machine Learning Research ,5,pp 1531-1555,2004.

[5] Hall M.A.,Correlation-Based Feature subset selection for Machine Learning, ph.D. dissertation Waikato ,New Zealand : univ.Waikato,1999.

[6] Hall M.A.,Correlation-Based Feature selection for Discrete and Numeric class Machine Learning, In Proceedings of 17th International Conference on Machine Learning ,pp 359-366, 2000.