ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 3, pp: (366-372), Month: July - September 2015, Available at: www.researchpublish.com

Implementation of Clustering Based Feature Subset Selection Algorithm for High Dimensional Data

[1]S Natarajan, [2]Parimala Anand, [3]D S Shanmukh, [4]Mohammed Saneen, [5]Darshan W M

[1,2,3,4,5] Department of ISE PESIT Bangalore

Abstract: **Feature Selection is an essential step in successful data mining applications, which can effectively reduce data dimensionality by removing the irrelevant and redundant features. Feature Selection is often an essential data prior to applying a learning algorithm. Machine learning algorithms are known to degrade in performance when faced with many features that are not necessary for predicting the desired output. The removal of irrelevant and redundant information often improves the performance of the machine learning algorithms. Feature selection techniques aim at reducing the number of unnecessary features in classification rules. The proposed Feature Subset Selection using clustering for high dimensional data works in two steps. First step, features are divided into clusters by using partitioning clustering methods. Second step, the representative feature that is strongly related to target class is selected from each cluster to form a subset of features. These features are then used for training using a machine learning algorithm and the results are compared with the original set of features. This is followed by comparison of the efficiency and accuracy of other feature selection algorithms using various combinations of machine learning algorithms.**

Keywords: **Data mining, Feature Selection, Relevant features, redundant features, FAST Algorithm.**

1. INTRODUCTION

Feature selection has been an active research area in pattern recognition, statistics, and data mining communities. The main idea of feature selection is to choose a subset of input variables by eliminating features with little or no predictive information. Feature selection can significantly improve the comprehensibility of the resulting classifier models and often build a model that generalizes better to unseen points. Further, it is often the case that finding the correct subset of predictive features is an important problem in its own right.

Data mining is one of the most prominent fields in this era. The amount of data generated in real time and the analysis involving that data to extract the meaningful information from this is what makes the field most challenging. Feature selection involves identifying a subset of the most useful features that produces compatible results as the original entire set of features. 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. Based on these criteria, the proposed system is experimentally evaluated.

Feature selection is applied to reduce the number of features in many applications where data has hundreds or thousands of features. Existing feature selection methods mainly perform explicit redundancy analysis in feature selection. For real-world concept learning problems, feature selection is important to speed up learning and to improve concept quality.

The high dimensionality of data poses challenges to learning tasks due to the curse of dimensionality. In the presence of many irrelevant features, learning models tend to over fit and become less comprehensible. Feature selection is one executive means to identify relevant features for dimensionality reduction.

The proposed 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 a subset of features. Features in different clusters are relatively independent; the clustering-

based strategy has a high probability of producing a subset of useful and independent features. The efficiency and effectiveness of the proposed algorithm are evaluated through an empirical study.

2. RELATED WORK

Feature subset selection can be viewed as the process of identifying and removing as many irrelevant and redundant features as possible. This is because: (i) irrelevant features do not contribute to the predictive accuracy [33], and (ii) redundant features do not redound to getting a better predictor for that they provide mostly information which is already present in other feature(s).

Traditionally, feature subset selection research has focused on searching for relevant features. A well-known example is Relief [34], which weighs each feature according to its ability to discriminate instances under different targets based on distance-based criteria function, but still cannot identify redundant features. However, along with irrelevant features, redundant features also affect the speed and accuracy of learning algorithms, and thus should be eliminated as well. Lei Yu and Huan Liu [2] showed that feature relevance is alone insufficient for effective feature selection.

CFS [3] is achieved by the hypothesis that a good feature subset is one that contains features highly correlated with the target, yet uncorrelated with each other.

Recently, hierarchical clustering has been adopted in word selection in the context of text. Distributional clustering has been used to cluster words into groups based either on their participation in particular grammatical relations with other or on the distribution of class labels associated with each word.

Quite different from these hierarchical clustering based algorithms, the proposed FAST algorithm uses minimum spanning tree based method to cluster features. Meanwhile, it does not assume that data points are grouped around centers or separated by a regular geometric curve. Moreover, our proposed FAST does not limit to some specific types of data.

3. EXISTING SYSTEM

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 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. 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 overfit on small training sets. The filter focus on finding relevant features. Feature relevance alone is insufficient for efficient feature selection of high-dimensional data. We define feature redundancy and propose to methods, in addition to their generality, are usually a good choice when the number of features is very large. This paper focuses on the filter method for feature selection.

4. PROPOSED METHOD

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 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. 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

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 3, pp: (366-372), Month: July - September 2015, Available at: www.researchpublish.com

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.

Advantages:

1. Good feature subsets contain features highly correlated with (predictive of) 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.

4.1 Flow Chart:

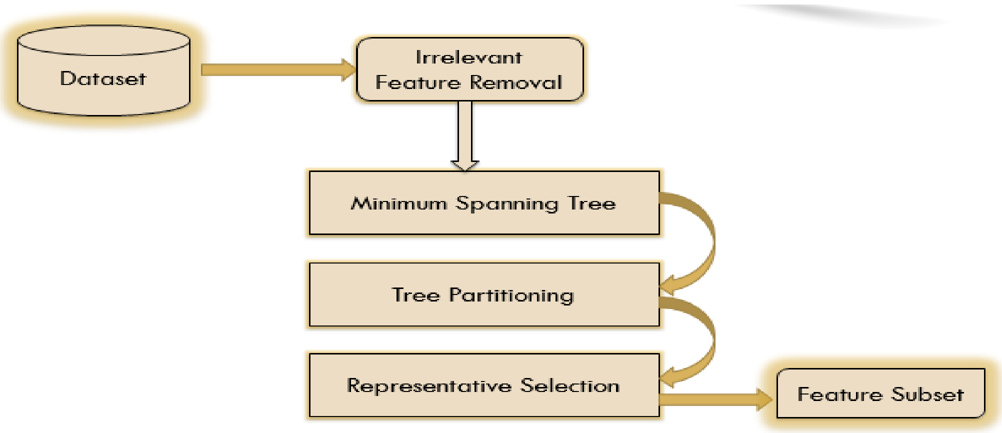


Fig. 1 Framework of the proposed feature subset selection algorithm

4.2 Algorithm:

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.

```
Algorithm 1: FAST
inputs : D( 1, 2, ..., , ) - the given data set - the T-
         Relevance threshold.
output: S - selected feature subset .
//==== Part 1 : Irrelevant Feature Removal ====
for i = 1 to m do
    T-Relevance = SU (  ,  )
    if T-Relevance >  then
        S = S ∪ {  };
//==== Part 2 : Minimum Spanning Tree Construction ====
G = NULL; //G is a complete graph
for each pair of features {  ,  } ⊂ S do
    F-Correlation = SU (  ,  )
    /    &    F-Correlation    &    &
    &    ;
minSpanTree = Prim (G); //Using Prim Algorithm to generate the
    minimum spanning tree
//==== Part 3 : Tree Partition and Representative Feature Selection ====
Forest = minSpanTree
for each edge ∈ Forest do
    if SU(  ,  ) < SU(  ,  ) ∧ SU(  ,  ) < SU(  ,  ) then
        Forest = Forest −
S =
for each tree ∈ Forest do
    = argmax  ∈  SU(  ,  )
    S = S ∪ {  };
return S
```

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 3, pp: (366-372), Month: July - September 2015, Available at: www.researchpublish.com

4.3 Implementation:

Implementation is the stage of the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective.

The implementation stage involves careful planning, investigation of the existing system and it's constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods.

4.4 Time complexity:

Time complexity analysis: The major amount of work for Algorithm 1 involves the computation of values for *T-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 O(m) in terms of the number of features. Assuming k (1<=k<=m) features are selected as relevant ones in the first part, when k= 1, only one feature is selected. Thus, there is no need to continue the rest parts of the algorithm, and the complexity is O(m). When 1<k<=m , the second part of the algorithm firstly constructs a complete graph from relevant features and the complexity is $O(k^2)$, and then generates a MST from the graph using Kruskal algorithm whose time complexity is $O(k^2)$. The third part partitions the MST and chooses the representative features with the complexity of O(k). Thus when 1<k<=m, the complexity of the algorithm is $O(m+k^2)$. This means when $k<=m^{1/2}$, FAST has linear complexity O(m).

5. IMPLEMENTATION RESULTS

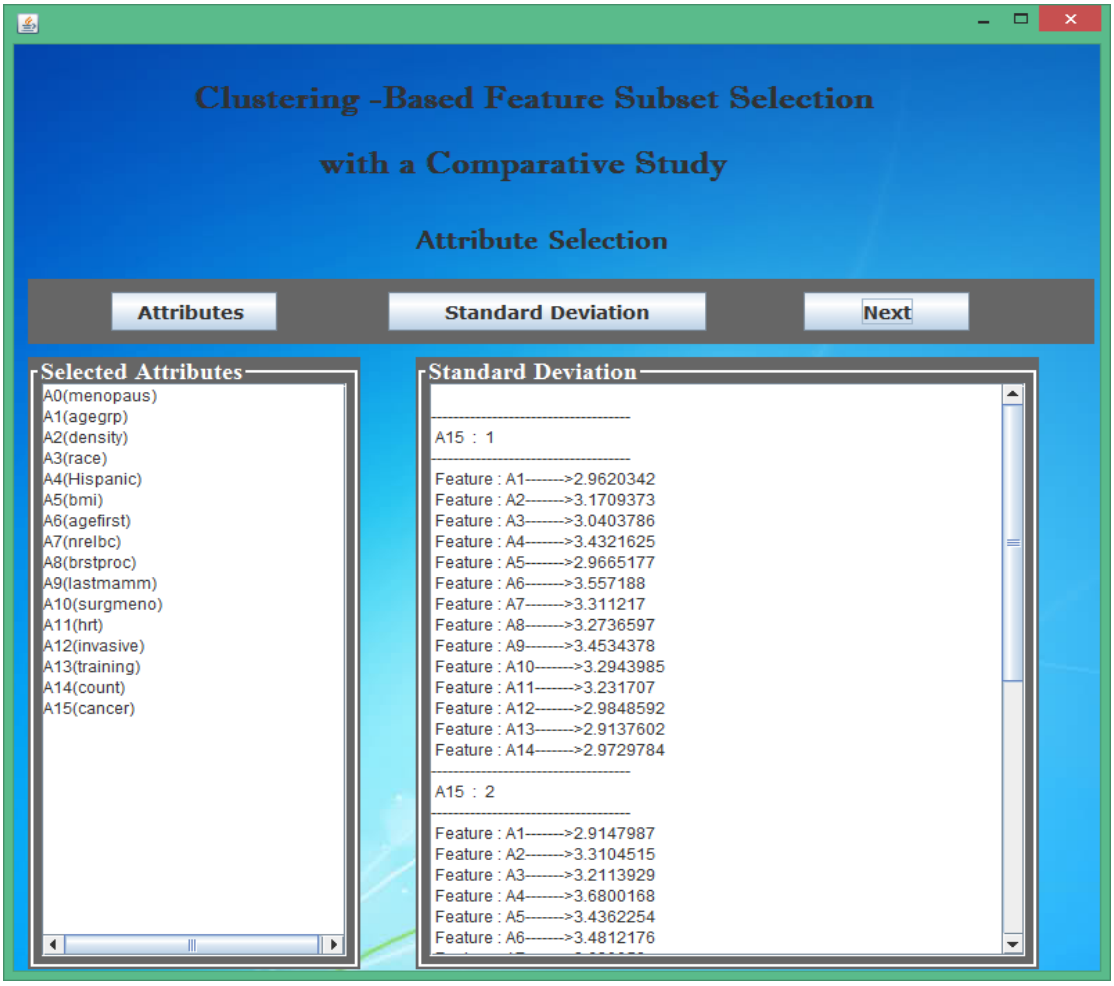


Fig. 2 Attributes before Feature Selection

The first step is step of the algorithm is to remove irrelevant features by calculating t relevance .To calculate T relevance Entropy, Information gain is computed.

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 3, pp: (366-372), Month: July - September 2015, Available at: www.researchpublish.com

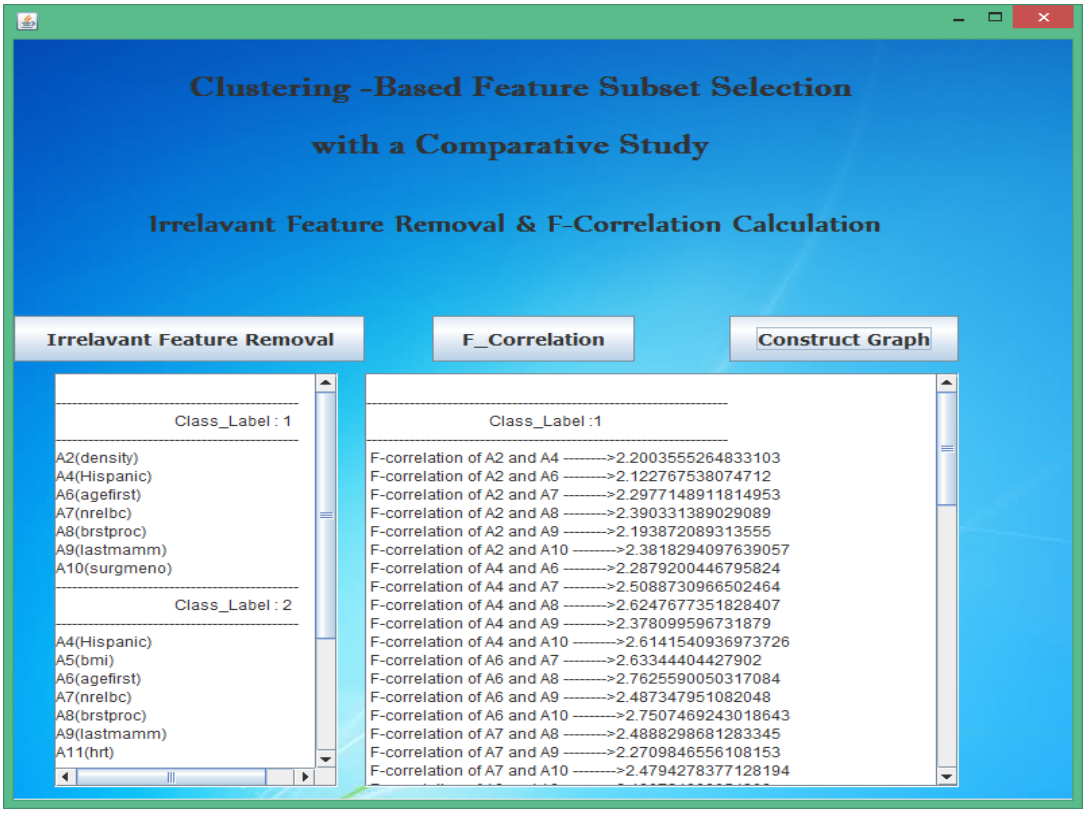


Fig. 3 Irrelevant Feature Removal

All the features with T-relevance value less than the threshold value are considered irrelevant and hence removed to get a relevant feature set.

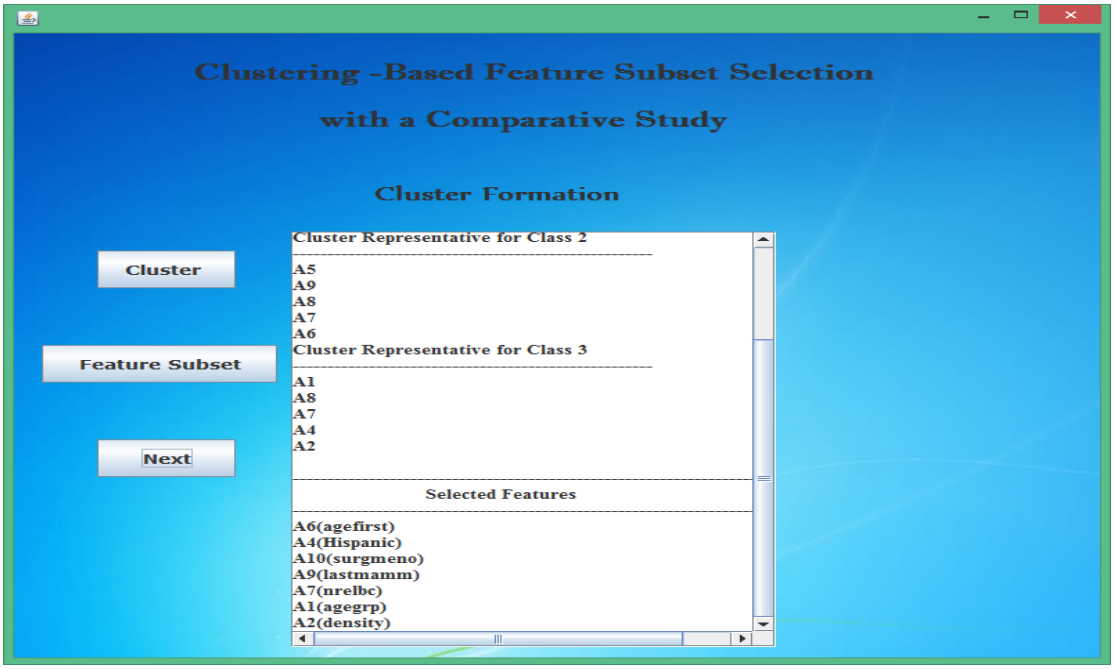
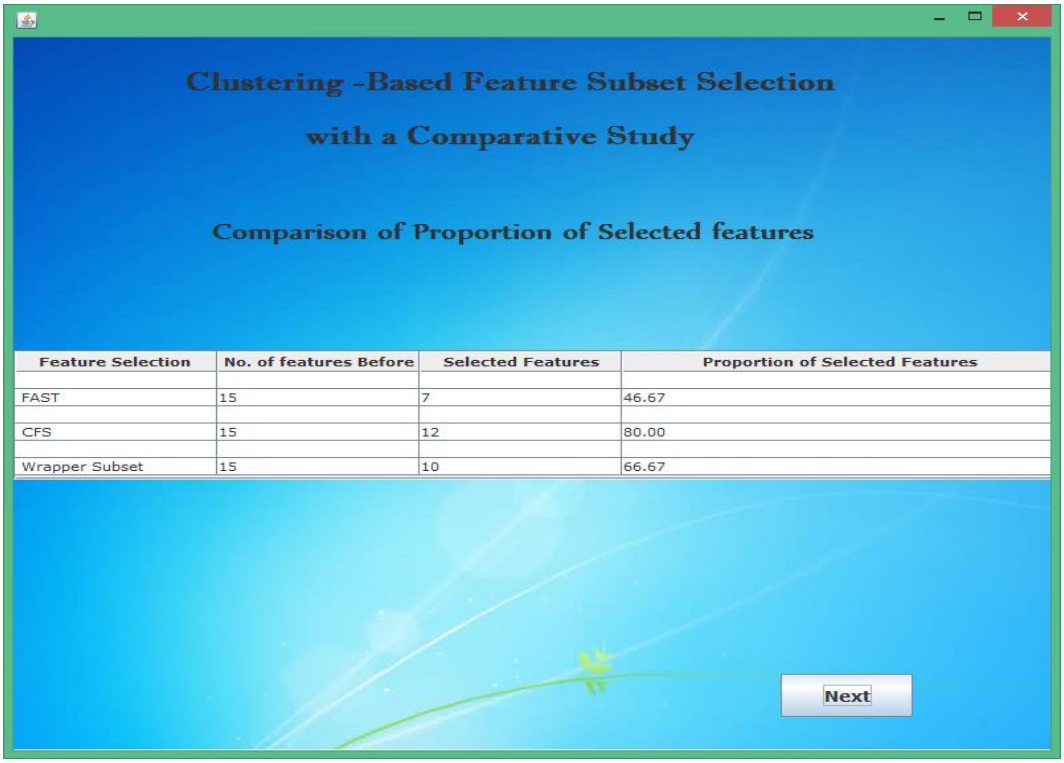


Fig. 4 Redundant Feature Removal and Final Feature Set

F-correlation is calculated between every pair of relevant features and a Graph is constructed. MST is constructed from the complete graph using Kruskal algorithm. Unnecessary edges are removed to partition the MST and obtain a forest, where each subtree is considered as a cluster. Representative features are selected from the clusters to obtain final feature set.

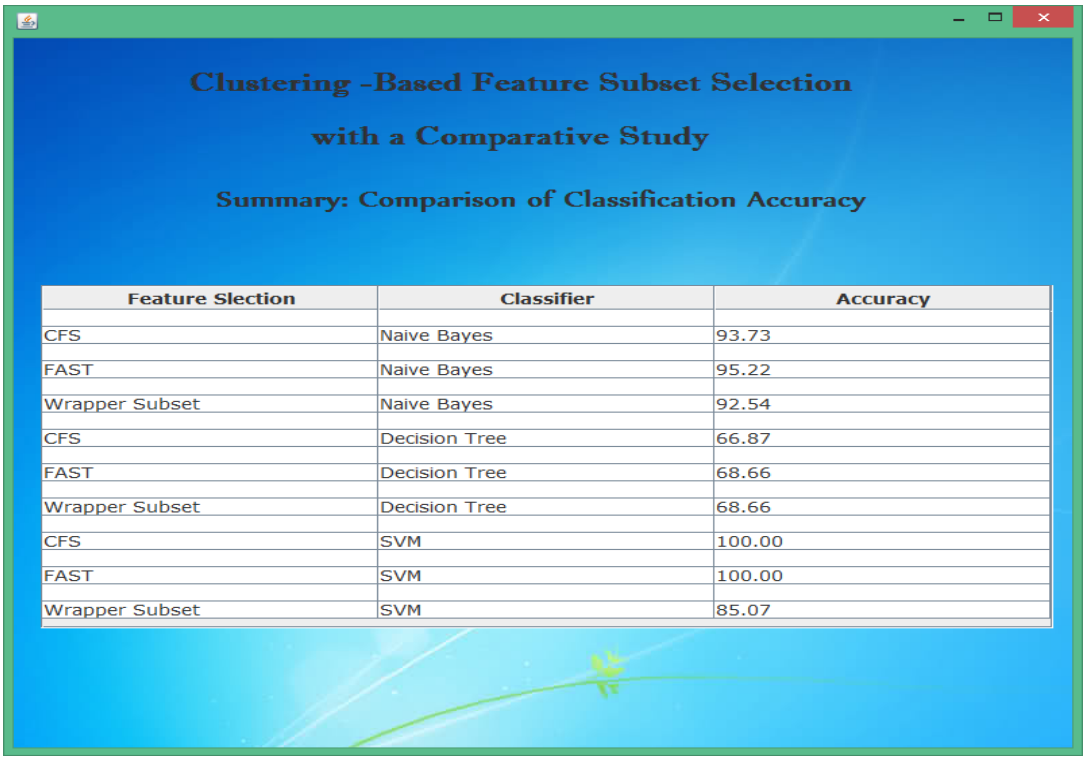
ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 3, pp: (366-372), Month: July - September 2015, Available at: www.researchpublish.com



Feature Selection	No. of features Before	Selected Features	Proportion of Selected Features
FAST	15	7	46.67
CFS	15	12	80.00
Wrapper Subset	15	10	66.67

Fig.5 Proportion of Selected Features

The implemented FAST algorithm is compared with other existing feature selection algorithms like CFS and Wrapper subset in terms of proportion of selected features.



Feature Slection	Classifier	Accuracy
CFS	Naive Bayes	93.73
FAST	Naive Bayes	95.22
Wrapper Subset	Naive Bayes	92.54
CFS	Decision Tree	66.87
FAST	Decision Tree	68.66
Wrapper Subset	Decision Tree	68.66
CFS	SVM	100.00
FAST	SVM	100.00
Wrapper Subset	SVM	85.07

Fig. 6 Comparison of Classification Accuracy

The Classification accuracy is compared using different classifiers. With three feature selection algorithms and three classifiers we obtain nine combinations of Feature Selection-Classification pairs. The classification accuracy for each combination is computed.

6. CONCLUSION

We have implemented a novel clustering-based feature subset selection algorithm for high dimensional data based on the algorithm specified in the paper [1]. The algorithm involves 1) removing irrelevant features, 2) constructing a minimum spanning tree from relative ones, and 3) partitioning the MST and selecting representative features. In the proposed algorithm, a cluster consists of features.

Each cluster is treated as a single feature and thus dimensionality is drastically reduced. The selected features were cross-checked for correctness with the help of other existing feature subset selection algorithms and were found to be correct. This implementation gives users flexibility like removing irrelevant features and constructing minimum spanning tree from the relative subset present in the dataset. The proposed work has characterized the associations between the irrelevant and redundant features and how they drastically affect the performance of the classifier.

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, January 2013.

[2] Laura Maria Cannas, "A Framework for Feature Selection in High-Dimensional Domains" 2012.

[3] L.C. Molina, L. Belanche, and A. Nebot, "Feature Selection Algorithms: A Survey and Experimental Evaluation," Proc. IEEE Int'l Conf. Data Mining, pp. 306-313, 2002.

[4] J. Demsar, "Statistical Comparison of Classifiers over Multiple Data Sets," J. Machine Learning Res., vol. 7, pp. 1-30, 2006.

[5] G. H. John, R. Kohavi and K. Pfleger, "Irrelevant Features and the Subset Selection Problem," Proceedings of the Eleventh International Conference on Machine Learning, pp121-129, 1994.

[6] Yun Zheng and Chee Keong Kwoh," A Feature Subset Selection Method Based On High-Dimensional Mutual Information" Entropy 2011.

[7] Padraig Cunningham,"Dimension Reduction", University College Dublin, Technical Report UCD-CSI-2007-7 August 8th, 2007.

[8] L.C. Molina, L. Belanche, and A. Nebot, "Feature Selection Algorithms: A Survey and Experimental Evaluation," Proc. IEEE Int'l Conf. Data Mining, pp. 306-313, 2002.

[9] D.A. Bell and H. Wang, "A Formalism for Relevance and Its Application in Feature Subset Selection," Machine Learning.

[10] R. Butterworth, G. Piatetsky-Shapiro, and D.A. Simovici, "On Feature Selection through Clustering," Proc. IEEE Fifth Int'l Conf. Data Mining, pp. 581-584, 2005.

[11] Leo Breiman, "Random Forests", Machine Learning, Volume 45, Issue 1, pp. 5-32, Kluwer Academic Publishers, 2001.

[12] Wang Kay Ngai, Ben Kao, Chun Kit Chui, "Efficient Clustering of Uncertain Data"ICDM'06 IEEE Computer Society, 2006.

[13] M.A. Hall, "Correlation-Based Feature Subset Selection for Machine Learning," PhD dissertation, Univ. of Waikato, 1999.

[14] F. Fleuret, "Fast Binary Feature Selection with Conditional Mutual Information", J. Machine Learning Research, vol. 5, pp. 1531-1555, 2004.

[15] L. Yu and H.Liu, "Feature Selection for High-Dimensional Data:A Fast Correlation-Based Filter Solution," Proc. 20th Int'l Conf. Machine Leaning, vol. 20, no. 2, pp. 856-863, 2003.

[16] E.R. Dougherty,"Small Sample Issues for Microarray-Based Classification," Comparative and Functional Genomics, vol. 2, no. 1, pp. 28-34, 2001.