Feature Selection Algorithm for High Dimensional Data using Fuzzy Logic

T.Jaga Priya Vathana[1], C.Saravanabhavan[2] and Dr. J.Vellingiri[3]

[1] Kongunadu College of Engineering and Technology/Anna University Chennai

Received: 7 December 2012 Accepted: 31 December 2012 Published: 15 January 2013

Abstract

Feature subset selection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy and improving results comprehensibility. This process improved by cluster based FAST Algorithm and Fuzzy Logic. FAST Algorithm can be used to Identify and removing the irrelevant data set. This algorithm process implements using two different steps that is graph theoretic clustering methods and representative feature cluster is selected. Feature subset selection research has focused on searching for relevant features. The proposed fuzzy logic has focused on minimized redundant data set and improves the feature subset accuracy.

Index terms—

1 Introduction

he performance, robustness, and usefulness of classification algorithms are improved when relatively few features are involved in the classification. Thus, selecting relevant features for the construction of classifiers has received a great deal of attention.

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 examples of embedded approaches. The wrapper methods use the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected sub-sets, 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.

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. Pereira et al., **??**aker et al., and Dillon et al. employed the distributional clustering of words to reduce the dimensionality of text data. 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 in-stances, 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. In our study, we apply graphtheoretic 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 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. The proposed feature subset se-lection algorithm FAST was tested upon 35 publicly available image, microarray, and text data sets. 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.

2 II.

3 Literature Review a) Statistical Comparisons of Classifiers over Multiple Data Sets

In this method introduce some new pre-or post processing step has been proposed, and the implicit hypothesis is made that such an enhancement yields an improved performance over the existing classification(D D D D D D D D)

algorithm. Alternatively, various solutions to a problem are proposed and the goal is to tell the successful from the failed. A number of test data sets is selected for testing, the algorithms are run and the quality of the resulting models is evaluated using an appropriate measure, most commonly classification accuracy. The remaining step, and the topic of this paper, is to statistically verify the hypothesis of improved performance. Various re-searchers have addressed the problem of comparing two classifiers on a single data set and proposed several solutions. The core of the paper is the study of the statistical tests that could be (or already are) used for comparing two or more classifiers on multiple data sets. Learning algorithms is used for the Classification purpose. The main disadvantage of this process is the problems with the multiple data set tests are quite different, even in a sense complementary.

4 b) A Features Set Measure Based on Relief

It used six real world dataset from the UCI repository have been used. Three of them have classification Problem with discrete features, the next two classifications with discrete and continuous features, and the last one is approximation problem. The learning algorithm is used to check the quality of feature selected are a classification and regression tree layer with pruning. This process and algorithms is implemented by the orange data mining System. Overall, the non-parametric tests, namely the Wilcox on and Friedman test are suitable for our problems. They are appropriate since they assume some, but limited commensurability. They are safer than parametric tests since they do not assume normal distributions or homogeneity of variance. There is an alternative opinion among statisticians that significance tests should not be per-formed at all since they are often misused, either due to misinterpretation or by putting too much stress on their results. The main disadvantage of the system is it measure to low accuracy of the search process.

5 c) Feature Clustering and Mutual Information for the Selection of Variables In Spectral Data

It face many problems in spectrometry require predicting a quantitative value from measured spectra. The major issue with spectrometric data is their functional nature; they are functions discredited with a high resolution. This leads to a large number of highlycorrelated features; many of which are irrelevant for the prediction. The approach for the features is to describe the spectra in a functional basis whose basis functions are local in the sense that they correspond to welldefined portions of the spectra. This process has clustering algorithm that algorithm recursively merges at each step the two most similar consecutive clusters. This algorithm return the output value associated with each cluster, its representative, is chosen to be the mean of the spectra over the range of features defined by the cluster. The main disadvantage of the problem is low number of clusters identified by the method allows the interpretation of the selected variables: several of the selected clusters include the spectral variables identified on these benchmarks as meaningful in the literature.

6 d) On Feature Selection through Clustering

This paper introduce an algorithm for feature selection that clusters attributes using a special metric and, then uses a hierarchical clustering for feature selection. Hierarchical algorithms generate clusters that are placed in a cluster tree, which is commonly known as a dendrogram. Clustering's are obtained by extracting those clusters that are situated at a given height in this tree. It use several data sets from the UCI dataset repository and, due to space limitations we discuss only the results obtained with the votes and zoo datasets, Bayes algorithms of the WEKA package were used for constructing classifiers on data sets obtained by projecting the initial data sets on the sets of representative attributes. Approach to attribute selection is the possibility of the supervision of the process allowing the user to opt between quasi-equivalent attributes It face classification problems that involve thousands of features and relatively few examples came to the fore. We intend to apply our techniques to this type of data.

7 III.

8 Fuzzy based Feature Subset Selection Algorithms

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. The cluster indexing and document assignments are repeated periodically to compensate churn and to maintain an up-to-date clustering solution. The k-means clustering technique and SPSS Tool to develop a real time and online system for a particular supermarket to predict sales in various annual seasonal cycles. The classification was based on nearest mean.

In order to more precisely introduce the algorithm, and because our proposed feature subset selection framework involves irrelevant feature removal and redundant feature elimination. Feature subset selection algorithm 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. We achieve this through a new feature selection framework which composed of the two connected components of 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.

9 Global Journal of Computer Science and Technology

The irrelevant feature removal is straightforward once the right relevance measure is defined or selected, while the redundant feature elimination is a bit of sophisticated. In our proposed FAST algorithm, it involves (a) the construction of the minimum spanning tree (MST) from a weighted complete graph; (b) the partitioning of the MST into a forest with each tree representing a cluster; and (c) the selection of representative features from the clusters.

In order to more precisely introduce the algorithm, and because our proposed feature subset selection framework involves irrelevant feature removal and redundant feature elimination, we firstly present the traditional definitions of relevant and redundant features, then provide our definitions based on variable correlation as follows.

John et al. presented a definition of relevant features. Suppose to be the full set of features, be a feature, = { } and . Let ' be a valueassignment of all features in ' , a value-assignment of feature , and a value-assignment of the target concept . The definition can be formalized as follows.

Definition: (Relevant feature) is relevant to the target concept if and only if there exists some , and , such that, for probability (' = , =)>0, (= ' = , =) (= =). Otherwise, feature is an irrelevant feature. Definition 1 indicates that there are two kinds of relevant features due to different : (i) when = , from the definition we can know that is directly relevant to the target concept; (ii) when '

, from the definition we may obtain that (,)= ( ). It seems that is irrelevant to the target concept. However, the definition shows that feature is relevant when using { }to describe the target concept.

10 Feature Subset Selection

11 Collected Cluster Representation

12 Feature Subset Result

Mutual information measures how much the distribution of the feature values and target classes differ from statistical independence. This is a nonlinear estimation of correlation between feature values or feature values and target classes. The symmetric uncertainty ( ) is derived from the mutual information by normalizing it to the entropies of feature values or feature values and target classes, and has been used to evaluate the goodness of features for classification by a number of researchers (e.g., Hall], Hall and Smith, Yu and Liu,, Zhao and Liu,). Therefore, we choose symmetric uncertainty as the measure of correlation between either two features or a feature and the target concept.

The symmetric uncertainty is defined as follows (,)=2× ( ) ( )+ ( ).

Where, 1. ( )is the entropy of a discrete random variable . Suppose ( ) is the prior probabilities for all values of , ( )is defined by ( )= ( )log2 ( ).

13 Gain (

) is the amount by which the entropy of decreases. It reflects the additional information about provided by and is called the information gain which is given by( )= ( ) ( ) = ( ) ( ).

Where ( ) is the conditional entropy which Quantifies the remaining entropy (i.e. uncertainty) of a random variable given that the value of another random variable is known. Suppose ( ) is the prior probabilities for all values of and ( )is the posterior probabilities of given the values of , (

)is defined by ( )= ( ) ( )log2 ( ). (4) Information gain is a symmetrical measure. That is the amount of information gained about after observing is equal to the amount of information gained about after observing . This ensures that the order of two variables (e.g., (,) or (,)) will not affect the value of the measure.

Symmetric uncertainty treats a pair of variables sym-metrically, it compensates for information gain's bias toward variables with more values and normalizes its value to the range [0,1]. A value 1 of (,) indicates. That knowledge of the value of either one completely predicts the value of the other and the value 0 reveals that and are independent. Although the entropy-based measure handles nominal or discrete variables, they can deal with continuous features as well, if the values are discredited properly in advance. Given (,) the symmetric uncertainty of variables and , the relevance T-Relevance between a feature and the target concept , the correlation F-Correlation between a pair of features, the feature Redundancy F-Redundancy and the representative feature R-Feature of a feature cluster can be defined as follows.

Definition:

(T-Relevance) The relevance between the feature and the target concept is referred to as The T-Relevance of and , and denoted by (,). If (,) is greater than a predetermined threshold , we say that is a strong T-Relevance feature.

Definition: (F-Correlation) The correlation between any pair of features and (,) is called the F-Correlation of and , and denoted by (,).

< } be a cluster of features. if , (,) (,) (,) > (,) is always corrected for each ( ), then are redundant features with respect to the given (i.e. each is a F-Redundancy).

Definition: (R-Feature) A feature is a representative feature of the cluster (i.e. is a R-Feature) if and only if, = argmax This means the feature, which has the strongest T-Relevance, can act as a R-Feature for all the features in the cluster. According to the above definitions, feature subset selection can be the process that identifies and retains the strong T-Relevance features and selects R-Features from feature clusters. The behind heuristics are that 1. Irrelevant features have no/weak correlation with Target concept; 2. Redundant features are assembled in a cluster and a representative feature can be taken out of the Cluster.

IV.

14 Algorithm and Analysis

The proposed FAST algorithm logically consists of three steps: 1. removing irrelevant features, 2. constructing a MST from relative ones, 3. Partitioning the MST and selecting Representative features. After removing all the unnecessary edges, a forest is obtained. Each tree Forest represents a cluster hat is denoted as which is the vertex set of as well. As illustrated above, the features in each cluster are redundant. **??**) features are selected as relevant ones in the first part, when =1, only one feature is selected. Thus, there is no need to continue the rest parts of the algorithm, and the complexity is( ). When 1<

15 Global

, the second part of the algorithm firstly constructs a complete graph from relevant features and the complexity is (2), and then generates a MST from the graph using Prim algorithm whose time complexity. The third part partitions the MST and chooses the representative features with the complexity. Thus when the complexity of the algorithm. This means when FAST has linear complexity while obtains the worst complexity when. However is heuristically set to be in the implementation of FAST. So the complexity, which is typically less than since. This can be explained as follows.

V.

16 Data Source

For the purposes of evaluating the performance and effectiveness of our proposed FAST algorithm, verifying whether or not the method is potentially useful in practice, and allowing other researchers to confirm our results, 35 publicly available data sets were used. The numbers of features of the 35 data sets vary from 37 to 49152 with a mean of 7874. The dimensionality of the 54.3% data sets exceed 5000, of which 28.6% data sets have more than 10000 features. The 35 data sets cover a range of application domains such as text, image and bio microarray data classification.

17 VI.

18 Experiment Setup

To evaluate the performance of our proposed FAST algorithm and compare it with other feature selection. Algorithms in a fair and reasonable way, we set up our experimental study as follows. 1) The proposed algorithm is compared with five different types of representative feature selection algorithms. They are (i) FCBF, (ii) Relief, (iii) CFS, (iv) Consist and (v) FOCUS SF [2], respectively. FCBF and Relief evaluate features individually. Relief searches for nearest neighbors of instances of different classes and weights features according to how well they differentiate instances of different classes. The other three feature selection algorithms are based on subset evaluation. CFS exploits best-first search based on the evaluation of a subset that contains features highly correlated with the tar-get concept, yet uncorrelated with each other. The Consist method searches for the minimal subset that separates classes as consistently as the full set can under bestfirst search strategy. FOCUS-SF is a variation of FOCUS [2]. FOCUS has the same evaluation strategy as Consist, but it examines all subsets of features. Considering the time efficiency, FOUCS-SF replaces exhaustive search in FOCUS with sequential forward selection.

Four different types of classification algorithms are employed to classify data sets before and after feature selection. They are (i) the probability-based Naive Bayes (NB), (ii) the tree-based C4.5, (iii) the instance-based lazy learning algorithm IB1, and (iv) the rule-based RIPPER, respectively. Naive Bayes utilizes a probabilistic method for classification by multiplying the individual probabilities of every feature-value pair. This algorithm assumes independence among the features and even then provides excellent classification results. Decision tree learning algorithm C4.5 is an extension of ID3 that accounts for unavailable values, continuous attribute value ranges, pruning of decision trees, rule derivation, and so on. The tree comprises of nodes (features) that are selected by information entropy. Instance-based learner IB1 is a single-nearest-neighbor algorithm, and it classifies entities taking the class of the closest associated vectors in the training set via 3) When evaluating the performance of the feature subset selection algorithms, four metrics, (i) the proportion of selected features (ii) the time to obtain the feature subset, (iii) the classification accuracy, and (iv) the Win/Draw/Loss record, are used. The proportion of selected features is the ratio of the number of features selected by a feature selection algorithm to the original number of features of a data set. The Win/Draw/Loss record presents three values on a given measure, i.e. the numbers of data sets for which our proposed algorithm FAST obtains better, equal, and worse performance than other five feature selection algorithms, respectively. The measure can be the proportion of selected features, the runtime to obtain a feature subset, and the classification accuracy, respectively.

19 VII.

20 Results and Analysis

In this paper present the experimental results in terms of the proportion of selected features, the time to obtain the feature subset, the classification accuracy, and the Win/Draw/Loss record. For the purpose of exploring the statistical significance of the results, we performed a nonparametric Friedman test followed by Nemenyi post-hoc test, as advised by Demsar and Garcia and Herrerato to statistically compare algorithms on multiple data sets. Thus the Friedman and the Nemenyi test results are reported as well a) Proportion of selected features Records the proportion of selected features of the six feature selection algorithms for each data set. From it we observe that) generally all the six algorithms achieve significant reduction of dimensionality by selecting only a small portion of the original features. FAST on average obtains the best proportion of selected features of 1.82%. The Win/Draw/Loss records show FAST wins other algorithms as well. 2) For image data, the proportion of selected features of each algorithm has an increment compared with the corresponding average proportion of selected features on the given data sets except Consist has an improvement. This reveals that the five algorithms are not very suitable to choose features for image data compared with for microarray and text data. FAST ranks 3 with the proportion of selected features of 3.59% that has a tiny margin of 0.11% to the first and second best proportion of selected features 3.48% of Consist and FOCUS-SF, and a margin of 76.59% to the worst proportion of selected features 79.85% of Relief. 3) For microarray data, the proportion of selected features has been improved by each of the six algorithms compared with that on the given data sets. This indicates that the six algorithms work well with microarray data. FAST ranks 1 again with the proportion of selected features of 0.71%. Of the six algorithms, only CFS cannot choose features for two data sets whose dimensionalities are 19994 and 49152, respectively. 4) For text data, FAST ranks 1 again with a margin of 0.48% to the second best algorithm FOCUS-SF. TABLE 2: Proportion of selected features of the six feature selection algorithms.

The Friedman test can be used to compare k algorithms over Ndata sets by ranking each algorithm on each data set separately. The algorithm obtained the best performance gets the rank of 1, the second best ranks 2, and so on. In case of ties, average ranks are assigned. Then the average ranks of all algorithms on all data sets are calculated and compared. If the null hypothesis, which is all algorithms are performing equivalently, is rejected under the Friedman test statistic, post-hoc tests such as the Nemenyi test can be used to determine which algorithms perform statistically different. The Nemenyi test compares classifiers in a pairwise manner. In order to further explore whether the reduction rates are significantly different we performed a Friedman test followed by a Nemenyi post-hoc test. The null hypothesis of the Friedman test is that all the feature selection algorithms are equivalent in terms of proportion of selected features. The test result isp=0. This means that at = 0.1, there is evidence to reject the null hypothesis and all the six feature selection algorithms are different in terms of proportion of selected features In order to further explore feature selection algorithms whose reduction rates have statistically significant differences, we performed a Nemenyi test. Fig. 3 shows the results with = 0.1 on the 35 data sets. The results indicate that the proportion of selected features of FAST is statistically smaller than those of Relief, CFS and FCBF, and there is no consistent evidence to indicate statistical differences between FAST, Consist, and FOCUS-SF, respectively. 1. Compared with original data, the classification accuracy of Naive Bayes has been improved by FAST, CFS, and FCBF by 12.86%, 6.62%, and 4.32%, respectively. Unfortunately, Relief, Consist, and FOCUS-SF have decreased the classification accuracy by 0.32%, 1.35%, and 0.86%, respectively. FAST ranks 1 with a margin of 6.24% to the second best accuracy 80.60% of CFS. At the same time, the Win/Draw/Loss records show that FAST outer forms all other five algorithms. 2. For image data, the classification accuracy of Naïve Bayes has been improved by FCBF, CFS, FAST, and Relief by 6.13%, 5.39%, 4.29%, and 3.78%, respectively. However, Consist and FOCUS-SF have decreased the classification accuracy by 4.69% and 4.69%, respectively. This time FAST ranks 3 with a margin of 1.83% to the best accuracy 87.32% of FCBF. 3. For microarray data, the classification accuracy of Naive Bayes has been improved by all the six algorithms FAST, CFS, FCBF, ReliefF, Consist, and FOCUS-SF by 16.24%, 12.09%, 9.16%, 4.08%, 4.45%, and 4.45%, respectively. FAST ranks 1 with a mar-gin of 4.16% to the second best accuracy 87.22% of CFS. This indicates that FAST is more effective than others when using Naive Bayes to classify microarray data. 4. For text data, FAST and CFS have improved the classification accuracy of Naive Bayes by 13.83% and 1.33%, respectively. Other four algorithms Re-liefF, Consist, FOCUS-SF, and FCBF have decreased the accuracy by 7.36%, 5.87%, 4.57%, and 1.96%, respectively. FAST ranks 1 with a margin of 12.50% to the second best accuracy 70.12% of CFS.

Selection algorithms FAST, FCBF, CFS, Relief, Consist, and FOCUS-SF by 5.31%, 4.54%, 7.20%, 0.73%, 0.60%, and 0.60%, respectively. This time FAST ranks 2 with a margin of 1.89% to the best accuracy 83.6% of CFS and a margin of 4.71% to the worst accuracy 76.99% of Consist and FOCUS-SF. 3) For microarray data, the classification accuracy of C4.5 has been improved by all the six algorithms FAST, FCBF, CFS, Relief, Consist, and FOCUS-SF by 11.42%, 7.14%, 7.51%, 2.00%, 6.34%, and 6.34%, respectively. FAST ranks 1 with a margin of 3.92% to the second best accuracy 79.85% of CFS. 4) For text data, the classification accuracy of C4.5 has been decreased by algorithms FAST, FCBF, CFS, ReliefF, Consist and FOCUS-SF by 4.46%, 2.70%, 19.68%, 13.25%, 16.75%, and 1.90% respectively. FAST ranks 3 with a margin of 2.56% to the best accuracy 83.94% of FOCUS-SF and a margin of 15.22% to the worst accuracy 66.16% of CFS. This means that at = 0.1, there are evidences to reject the null hypotheses and the accuracies are different further differences exist in the six feature selection algorithms. From Fig. ?? we observe that the accuracy of Naïve Bayes with FAST is statistically better than those with Relief, Consist, and FOCUS-SF. But there is no consistent evidence to indicate statistical accuracy differences between Naive Bayes with FAST and with CFS, which also holds for Naive Bayes with FAST and with FCBF. From Fig. ?? we observe that the accuracy of C4.5 with FAST is statistically better than those with Relief, Con-sist, and FOCUS-SF. But there is no consistent evidence to indicate statistical accuracy differences between C4.5 with FAST and with FCBF, which also holds for C4.5 with FAST and with CFS. From Fig. ?? we observe that the accuracy of IB1 with FAST is statistically better than those with Relief. But there is no consistent evidence to indicate statistical accuracy differences between IB1 with FAST and with FCBF, Consist, and FOCUS-SF, respectively, which also holds for IB1 with FAST and with CFS. From Fig. ?? we observe that the accuracy of RIPPER with FAST is statistically better than those with Relief. But there is no consistent evidence to indicate statistical accuracy differences between RIPPER with FAST and with FCBF, CFS, Consist, and FOCUS-SF, respectively. For the purpose of exploring the relationship between feature selection algorithms and data types, i.e. which algorithms are more suitable for which types of data, we rank the six feature selection algorithms according to the classification accuracy of a given classifier on a specific type of data after the feature selection algorithms are performed. Then we summarize the ranks of the feature selection algorithms under the four different classifiers, and give the final ranks of the feature selection algorithms on different types of data. Table ?? shows the results. From Table ?? we observe that (i) for image data, CFS obtains the rank of 1, and FAST ranks 3; (ii) for microarray data, FAST ranks 1 and should be the undisputed first choice, and CFS is a good alternative; (iii) for text data, CFS obtains the rank of 1, and FAST and FCBF are alternatives; and (iv) for all data, FAST ranks 1 and should be the undisputed first choice, and FCBF, CFS are good alternatives.

21 VIII.

22 Sensitivity Analysis

Like many other feature selection algorithms, our pro-posed FAST also requires a parameter that is the threshold of feature relevance. Different values might end with different classification results. In order to explore which parameter value results in the best classification accuracy for a specific classification problem with a given classifier, a 10 fold cross-validation strategy was employed to reveal how the classification accuracy is changing with value of the parameter. Just like the default values used for FAST in the experiments are often not the optimal in terms of classification accuracy, the default threshold values used for FCBF and Relief (CFS, Consist, and FOCUS-SF do not require any input parameter) could be so. In order to explore whether or not FAST still outperforms when optimal threshold values are used for the comparing algorithms, 10-fold cross-validation methods were firstly used to determine the optimal threshold values and then were employed to conduct classification for each of the four classification methods with the different feature subset selection algorithms upon the 35 data sets. The results reveal that FAST still outperforms both FCBF and Relief for all the four classification methods, Fig. 10 shows the full details. signed ranks tests with = 0.05 were performed to values are smaller than 0.05, this indicates that the FAST is significantly better than both FCBF and Relief.

23 Global Journal of Computer Science and Technology

IX.

24 Conclusion

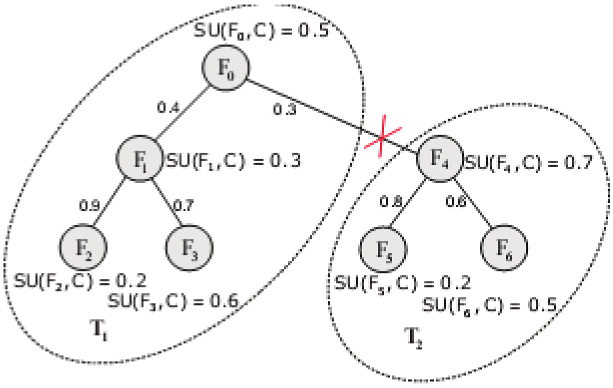
In this paper, we have presented a novel clustering-based feature subset selection algorithm for high dimensional data. The algorithm involves (i) removing irrelevant features, (ii) constructing a minimum spanning tree from relative ones, and (iii) partitioning the MST and selecting representative features. In the proposed algorithm, a cluster consists of features. Each cluster is treated as a single feature and thus dimensionality is drastically reduced. We have compared the performance of the proposed algorithm with those of the five well-known feature selection algorithms FCBF, Relief, CFS, Consist, and FOCUS-SF on the 35 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 Bayes, C4.5, 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 different types of correlation measures, and study some formal properties of feature space.

25 Global

1

Figure 1: VolumeFigure 1 :

[1] [2]



2

Figure 2: Figure 2 :

Algorithm 1: FAST

inputs: $D(F_1, F_2, ..., F_m, C)$ - the given data set
θ - the *T-Relevance* threshold.
output: S - selected feature subset .

//==== Part 1 : Irrelevant Feature Removal ====
for i = *1 to* m **do**
  T-Relevance = SU (F_i, C)
  if *T-Relevance* $> \theta$ **then**
    $S = S \cup \{F_i\}$;

//==== Part 2 : Minimum Spanning Tree Construction ====
G = NULL; *//G is a complete graph*
for *each pair of features* $\{F'_i, F'_j\} \subset S$ **do**
  F-Correlation = SU (F'_i, F'_j)
  Add F'_i *and/or* F'_j *to* G *with F-Correlation as the weight of the corresponding edge;*
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* $E_{ij} \in$ *Forest* **do**
  if $\mathrm{SU}(F'_i, F'_j) < \mathrm{SU}(F'_i, C) \wedge \mathrm{SU}(F'_i, F'_j) < \mathrm{SU}(F'_j, C)$ **then**
    Forest = *Forest* $- E_{ij}$
$S = \phi$
for *each tree* $T_i \in$ *Forest* **do**
  $F^j_R = \mathrm{argmax}_{F'_k \in T_i} \mathrm{SU}(F'_k, C)$
  $S = S \cup \{F^j_R\}$;
return S

Figure 3: C

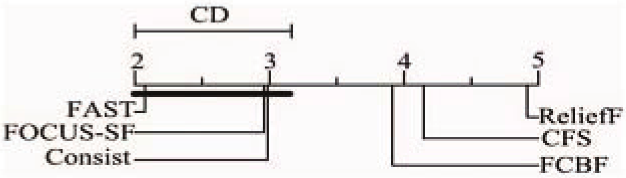


Figure 4:

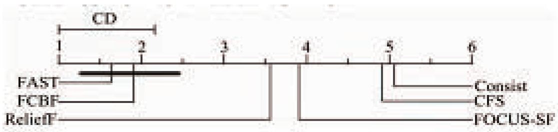
3 

Figure 5: Figure 3 :

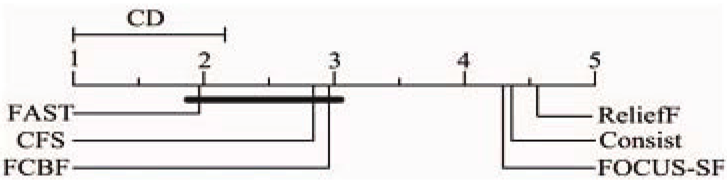


Figure 6:

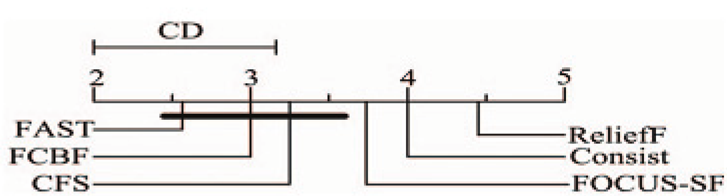
4 

Figure 7: Figure 4 :

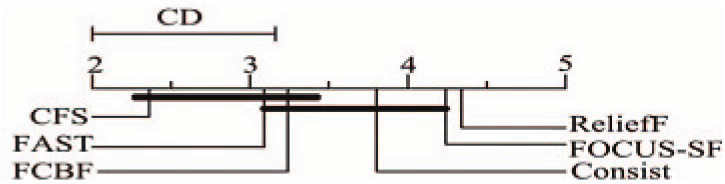
5678 

Figure 8: Figure 5 :CFigure 6 :Figure 7 :Figure 8 :

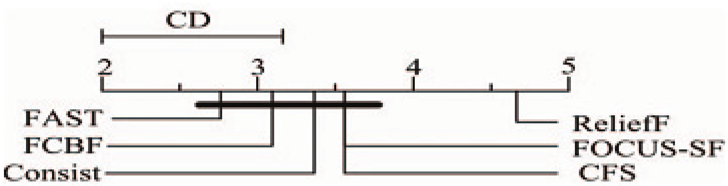
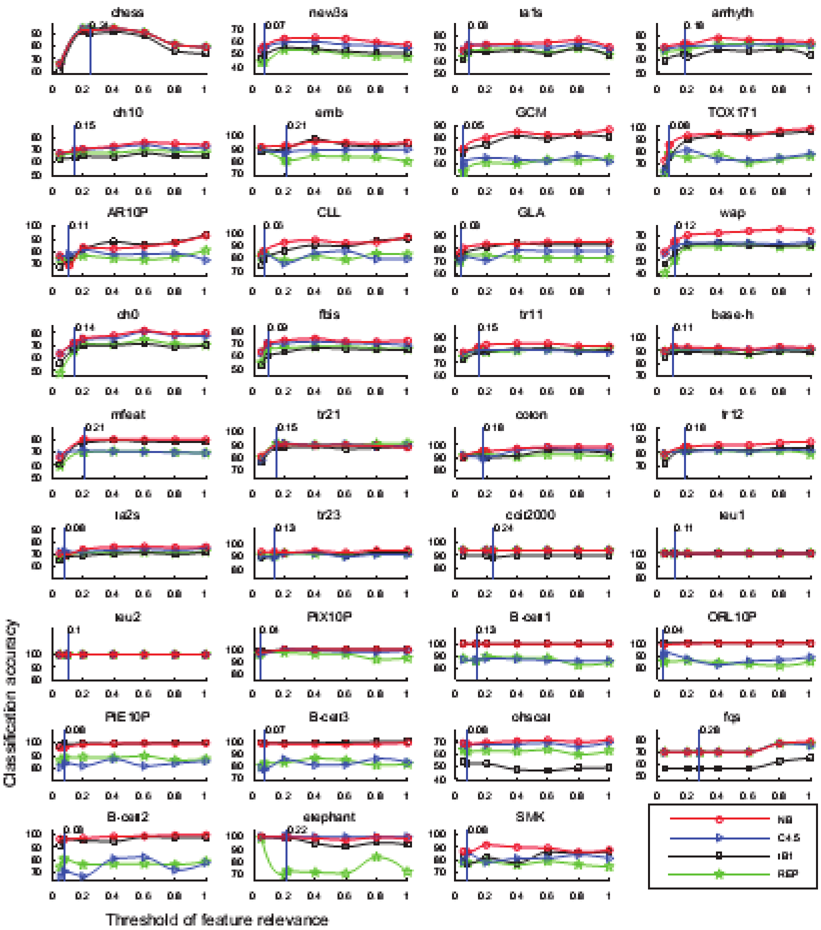
9 

Figure 9: VolumeFigure 9 :



10

Figure 10: Figure 10 :

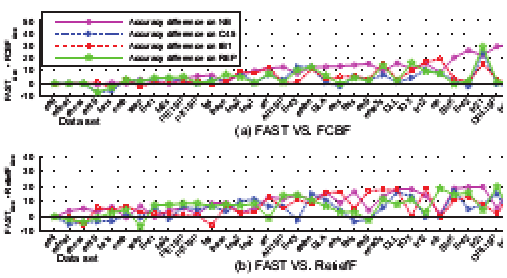


Figure 11:

The reason behind is that

either is interactive with or is re- dun- dant with

- . In this case, we say is indirectly relevant to

the target concept. Let ( ), is said to be a Markov

blanket for if and only if ({ },

,)= ({ },). Definition: (Redundant

feature). Let be a set of features, a feature in is

redundant if and only if it has a

Figure 12:

Figure 13:

[Dhillon et al. ()] ‘A divisive information theoretic feature clustering algorithm for text classification’. I S Dhillon , S Mallela , R Kumar . *J. Mach. Learn. Res* 2003. 3 p. .

[Arauzo-Azofra et al. ()] ‘A feature set measure based on relief’. A Arauzo-Azofra , J M Benitez , J L Castro . *Proceedings of the fifth international conference on Recent Advances in Soft Computing*, (the fifth international conference on Recent Advances in Soft Computing) 2004. p. .

[Bell and Wang ()] ‘A formalism for relevance and its application in feature subset selection’. D A Bell , H Wang . *Machine Learning*, 2000. 41 p. .

[Almuallim and Dietterich ()] ‘Algorithms for Identifying Relevant Features’. H Almuallim , T G Dietterich . *Proceedings of the 9th Canadian Conference on AI*, (the 9th Canadian Conference on AI) 1992. p. .

[Dash et al. ()] ‘Consistency based feature Selection’. M Dash , H Liu , H Motoda . *Proceedings of the Fourth Pacific Asia Conference on Knowledge Discovery and Data Mining*, (the Fourth Pacific Asia Conference on Knowledge Discovery and Data Mining) 2000. p. .

[Dash and Liu ()] ‘Consistency-based search in feature selection’. M Dash , H Liu . *Artificial Intelligence* 2003. 151 (1-2)) p. .

[Baker and Mccallum ()] ‘Distributional clustering of words for text classification’. L D Baker , A K Mccallum . *Proceedings of the 21st Annual international ACM SIGIR Conference on Research and Development in information Retrieval*, (the 21st Annual international ACM SIGIR Conference on Research and Development in information Retrieval) 1998. p. .

[Cohen ()] ‘Fast Effective Rule Induction’. W Cohen . *Proc. 12th international Conf. Machine Learning (ICML’95)*, (12th international Conf. Machine Learning (ICML’95)) 1995. p. .

[Dash and Liu ()] ‘Feature Selection for Classification’. M Dash , H Liu . *Intelligent Data Analysis* 1997. 1 (3) p. .

[Biesiada and Duch] ‘Features election for highdimensionaldata -son redundancy based filter’. J Biesiada , W Duch . *Advancesin Soft Computing* 45 p. 8.

[Das ()] ‘Filters, wrappers and a boosting-based hybrid for feature Selection’. S Das . *Proceedings of the Eighteenth International Conference on MachineLearning*, (the Eighteenth International Conference on MachineLearning) 2001. p. .

[Almuallim and Dietterich ()] ‘Learning boolean concepts in the presence of many irrelevant features’. H Almuallim , T G Dietterich . *Artificial Intelligence* 1994. 69 (1-2) p. .

[Chanda et al. ()] ‘Mining of Attribute Interactions Using Information Theoretic Metrics’. P Chanda , Y Cho , A Zhang , M Ramanathan . *Proceedings of IEEE international Conference on Data Mining Workshops*, (IEEE international Conference on Data Mining Workshops) 2009. p. .

[Fayyad and Irani ()] ‘Multi-interval discretization of continuous-valued attributes for classification learning’. U Fayyad , K Irani . *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, (the Thirteenth International Joint Conference on Artificial Intelligence) 1993. p. .

[Butterworth et al. ()] ‘On Feature Se-lection through Clustering’. R Butterworth , G Piatetsky-Shapiro , D A Simovici . *Proceedings of the Fifth IEEE international Conference on Data Mining*, (the Fifth IEEE international Conference on Data Mining) 2005. p. .

[Chikhi and Benhammada ()] ‘ReliefMSS: a variation on a feature ranking Relief algorithm’. S Chikhi , S Benhammada . *Int. J. Bus. Intell. Data Min* 2009. 4 p. .

[Dougherty ()] ‘Small sample issues for microarray-based classification’. E R Dougherty . *Comparative and Functional Genomics* 2001. 2 (1) p. .

[Demsar ()] ‘Statistical comparison of classifiers over multiple data sets’. J Demsar . *J. Mach. Learn. Res* 2006. 7 p. .

[Cardie ()] ‘Using decision trees to improve casebased learning’. C Cardie . *Pro-ceedings of Tenth International Conference on Machine Learning*, 1993. p. .

[Battiti ()] ‘Using mutual information for selecting features in supervised neural net learning’. R Battiti . *IEEE Transactions on Neural Networks* 1994. 5 (4) p. . (confirm the results as advised by Demsar. All the ??)