Scientific Journal Impact Factor (SJIF): 1.711
e-ISSN: 2349-9745
p-ISSN: 2393-8161
International Journal of Modern Trends in Engineering and Research
www.ijmter.com

Clustering Based Feature Subset Algorithm in High Dimensional Data

M.Ravichandran[1], Dr.A. Shanmugam[*2]

[1]*Department of Information Technology, Asst.Prof(Sr.Grade), Bannari Amman Institute of Technology, S a thyamangalam, Erode(Dt), Tamilnadu, India*

[2]*Professor, Dept of Electronics and Communication Engineering, SNS College of Technology, Coimbatore, India.*

Abstract— In the high dimensional data the dimensional reduction is an important factor, for that purpose the clustering based feature subset selection algorithm is proposed in this paper. The features have been clustered according to the class labels. The Relevance of the clustered features has been evaluated. The correlation of the relevant clustered feature is then evaluated. Based on the correlation evaluation the Minimum Spanning Tree (MST) has been generated. The representatives of each class have been identified by the MST. The effectiveness is determined in terms of time required to find the subset of feature and the efficiency is determined terms of quality of the subset. By comparing the proposed algorithm with the existing feature selection algorithms like FCBF, reliefF, CFS etc with respect to the four classification algorithms namely Naive Bayer, the tree based c4.5, the instance based IB1 and rule based RIPPER the proposed algorithm is better in terms of efficiency and accuracy. The results are computed with various types of data set.

Keywords: Minimum Spanning Tree (MST), FCBF, ReliefF, CFS, IBI, Naïve Bayes

I. INTRODUCTION

Feature selection involves identifying a subset of the most useful features that produces compatible results as the original entire set of features. This is since 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 several feature subset selection algorithms, some can effectively remove irrelevant features but fail to handle redundant features yet some of others can eliminate the irrelevant while taking care of the redundant features. The proposed algorithm focuses on the feature subset selection to perform the searching relevant features.

A famous sample is Relief which weighs each feature according to its ability to discriminate instances under different targets based on distance-based criteria function.

II. RELATED WORK

1 Fast Correlation-Based Filter

In FCBF [2], from the dataset with N features and a class C the algorithm finds a set of predominant features S_{best} for the class concept. Initially it calculates the Symmetric uncertainty (SU) for each feature, selects relevant features into S'_{list} based on the predefined threshold and orders them in descending order according to their SU values. Then the redundant feature has been removed from the $S'_{list.}$.

According to Heuristic 1 [2], the iteration starts from the first element in S'_{list} and continues as follows. For all the remaining features (from the one right next to fp to the last one in S'_{list}), if f_p

happens to be a redundant peer to a feature f_q, it will be removed from S'_{list}. After one round of filtering features based on f_p, the algorithm will take the currently remaining feature right next to f_p as the new reference to repeat the filtering process. The algorithm stops until there is no more feature to be removed from
S'_{list}.

2 Minimum Redundancy Maximum Relevance Feature

The MRMR (minimum redundancy maximum relevance) method [3] selects features that have the highest relevance with the target class and are also minimally redundant, i.e., selects features that are maximally dissimilar to each other.

3 Correlation-based Feature

In this paper the results of the experiments in which have compared the correlation-based feature selection strategy with the unmodified pair wise approach. The experiments were performed using neural network classifiers on commonly used benchmark data sets.

4. ReliefF

The quality of attributes in the problems with the strong dependencies between the attributes can be estimated with the efficient procedure namely ReliefF [3]. In ReliefF, the feature subset selection is made by means of data preprocessing method. The quality of genes according to their well distinguished values between the instances that are nearer to each other. ReliefF is capable of dealing with multi-class datasets and is an efficient method to deal with noisy and incomplete datasets. It can be used to estimate the quality and identify the existence of conditional dependencies between attributes effectively.

5 Support Vector Machine Recursive Feature Elimination

SVM-RFE was introduced by Guyon et. al., for ranking genes from gene expression data for cancer classification [5]. It is now being widely used for gene selection and several improvements have been recently suggested. SVM-RFE, starting with all the genes, removes the genetic factor that is least significant for classification recursively in a backward elimination manner.

6 Fast Binary Feature Selection with Conditional Mutual Information

The Fast Binary Feature Selection [3] is the very fast feature selection technique based on conditional mutual information. The proposed algorithm uses the conditional mutual information to select a family of binary features which are individually discriminating and weekly dependent.

In FBFS, the naïve Bayesian classifier is used in the process of feature selection by means of Conditional Mutual Information Maximization (CMIM).

III. FEATURE SUBSET SELECTION ALGORITHM

1 Relevance analysis

The proposed analysis removes irrelevant features by ranking correlation between feature and class SU(X,C) and between feature and feature SU(X,Y) by calculating Symmetrical Uncertainty (SU), given as below,

$$SU(X,Y) = 2 * \frac{Gain(X \mid Y)}{H(X) + H(Y)} \quad (1)$$

Where,

X and Y are the features,

$$Gain(X \mid Y) = H(X) - H(X \mid Y) \quad (2)$$

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \quad (3)$$

$$H(X \mid Y) = -\sum_{y \in Y} p(y) \sum_{x \in X} p(x \mid y) \log_2 p(x \mid y). \quad (4)$$

Here, p (x|y) denotes the posterior probabilities of X given the values of Y and p(x) denotes the prior probabilities for all the values of X.

SU is the personalized version of Information Gain that balances the bias and the SU has the ranges between 0 and 1. If SU between the feature and the class is equal to 1, means that this feature is completely related to the corresponding class. Conversely, if SU between the feature and the class is equal to 0, then the features are irrelevant to this corresponding class.

2 Threshold -Relevance feature analysis

As far as the Relevancy analysis is performed, if SU(X, C) is greater than a user determi ned threshold, we

Algorithm: CHD

Input:	S(f1,f2,…, fn, C) //training dataset	
	Th	//threshold value
Output:	I	//Feature subset

Begin:

I=empty;

for i=1 to N begin

SUic =calculateSU(fi ,C); if(SUic >Th)

addto(I, fi);

end

I=Desc (I) based on SUic; count=0;

endcount=-1; flag=0;

while endcount<>count begin count=endcount;

fp =firstelement(I);

while fp <>NULL AND flag<>1 begin fq =lastelement (I);

pass=0;

while fq <>NULL AND pass <>1 begin if(fp == fq) break;

tempSUpq = calculateSU(fp , fq) if(tempSUpq>SUp &&

tempSUpq>SUq) delete(I, fq); pass=1;

count=count+1; else fq =previous(I, fq);

end

fp =next(I, fp); end

end MinimumSpanTree(SU(f_i,C)); for each C in S

repc=Max(SU(f_i,C); end

end:

say that X is a strong Relevance feature and it is given as
$X \in F$.

3 Redundancy feature analysis

Once the irrelevant features has been removed based on the threshold value$^{\theta}$, the redundancy between the features can be determined by the following condition,

$$SU(X, Y) \leq SU(X, C) \,||\, SU(X, Y) \leq SU(Y, C) \quad (5)$$

In this above equation (5), we can remove the all redundancy feature set and finally we get the reduced dimension.

3.4 Tree Generation and Representative Feature Selection

Generate the tree by using the reduced dimension feature, and select the representative of each class label as,

$$\text{Max } (SU(X, C)) \quad (6)$$

IV.EXPERIMENTS AND RESULTS

In order to evaluate the performance of CHD algorithm, 8 different datasets were used. The dataset were taken as image and text type to perform the evaluation. The proposed algorithm CHD is compared with the FCBF.

Data	Data Name	F	I	T	Domain
1	Chess	37	3196	2	Text
2	mfeat-fourier	77	2000	10	Image
3	fbis.wc	2001	2463	17	Text
4	tr12.wc	5805	313	8	Text
5	tr23.wc	5833	204	6	Text
6	tr11.wc	6430	414	9	Text
7	PIX10P	10001	100	10	Image
8	ORL10P	10305	100	10	Image

1 Classification Accuracy

The 10-fold cross-validation precisions of the four different types of classifiers on the 8 data sets before and after each feature selection algorithm is performed, respectively.

1.1 Naive Bayes

The Naive Bayes classifier, the classification decision may often be perfect even if its probability estimates are inaccurate. Even though, Naive Bayes classifier is simple, it can often outperform more complex classification methods.

Data Name	Classification precision of Naive Bayes with	
	CHD	FCBF
Chess	92.98	92.09
mfeat-fourier	80.25	79.20
fbis.wc	70.21	52.25
tr12.wc	84.25	57.95
tr23.wc	94.11	53.98
tr11.wc	82.47	58.72
PIX10P	98.00	98.40
ORL10P	99.20	98.80

1.2 C4.5

In C4.5 classifier,the user defined threshold can splits the attribute values into two partitions. The values above the threshold can be partitioned into one child and the remaining as another child. The missing attribute values can also be handled by this algorithm.

In pseudo code the algorithm is:

1. Check for base cases
2. For each attribute a (Find the normalized information gain from splitting on a)
3. Let a_best be the attribute with the highest normalized information gain
4. Create a decision node that splits on a_best
5. Recur on the sublists obtained by splitting on a_best, and add those nodes as children of node

Data Name	Classification precision of C4.5 with	
	CHD	FCBF
Chess	94.12	94.12
mfeat-fourier	71.25	75.74
fbis.wc	81.68	85.43
tr12.wc	89.87	94.80
tr23.wc	80.22	82.04
tr11.wc	82.47	58.72
PIX10P	97.00	95.40
ORL10P	90.33	82.60

1.3 IB1

IBL Streams is an instance-based learning algorithm for performing the classification and regression on data streams. The method is capable to handle large streams through low requirements in terms of memory and computational control. Moreover, it disposes of apparatuses for adapting to concept drift and concept shift. The implementation of IBL Streams is supposed to be used as an extension to the MOA (Massive Online Analysis) framework for data stream mining.

Data Name	Classification precision of IB1 with	
	CHD	FCBF
Chess	90.18	91.47
mfeat-fourier	77.87	81.69
fbis.wc	60.09	61.91
tr12.wc	82.11	83.43
tr23.wc	90.18	86.55
tr11.wc	78.43	79.65
PIX10P	99.00	99.00
ORL10P	100.00	97.60

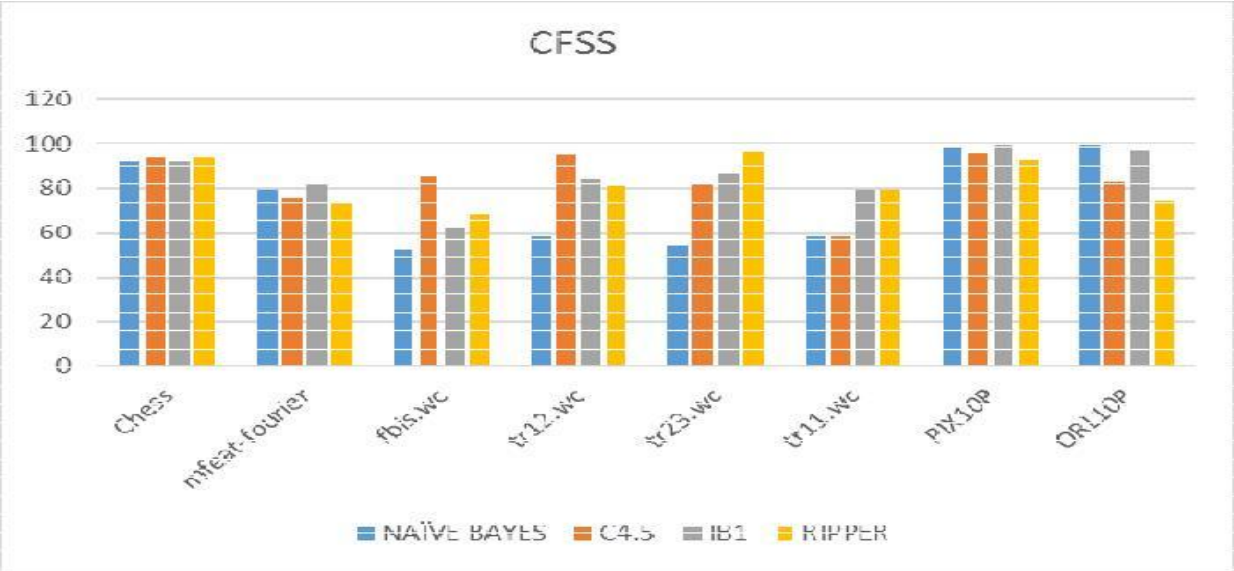
1.4 RIPPER

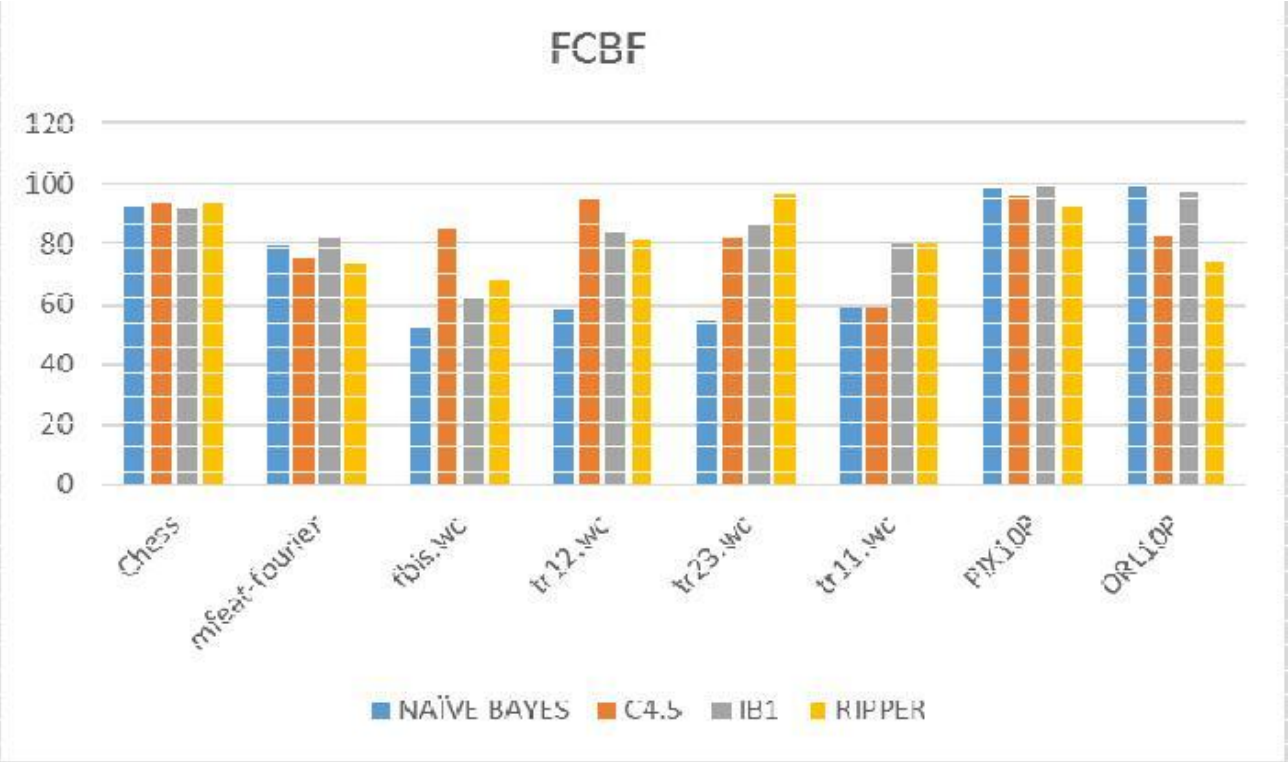
RIPPER, is an inductive rule apprentice. This algorithm generated a detection perfect composed of resource rules that was built to detect future examples of malicious executable. This algorithm used libBFD evidence as features. By RIPPER, this is a rule-based learner. The RIPPER builds a set of rules that identify the classes though it minimizing the amount of error on it. The error is distinct by the number of exercise examples misclassified by the rules.

Data Name	Classification precision of RIPPER with	
	CFSS	FCBF
Chess	94.09	94.09
mfeat-fourier	70.40	73.46
fbis.wc	65.58	68.18
tr12.wc	82.53	81.13
tr23.wc	91.15	95.96
tr11.wc	80.13	79.52
PIX10P	96.67	93.00
ORL10P	85.33	73.80

2 Sensitivity Analysis

Like many other feature selection algorithms, our proposed CHD also requires a parameter $^{\theta}$ that is the threshold of feature relevance. Different θ 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 certain classifier, a 10 fold cross-validation strategy stood employed to reveal how the classification accuracy is changing with value of the parameterθ .





V. CONCLUSION

In this paper, we have presented a novel clustering-based feature subset selection algorithm for high dimensional data. The algorithm contains 1) removing irrelevant features, 2) partitioning the MST and 3) selecting representative features. In the proposed algorithm, a cluster contains of features. Each cluster is preserved as a single feature and thus dimensionality is extremely reduced.

The results had illustrated that the proposed method is very effective and has great potential for relevant selection.

ACKNOWLEDGMENT

I would be grateful to my guide for her cooperation in studying various clustering techniques used in high dimensional data.

REFERENCES

[1] E. Bonilla-Huerta, et al., "Hybrid Filter- Wrapper with a Specialized Random Multi- Parent Crossover Operator for Gene Selection and Classification Problems," Bio-Inspired Computing and Applications, pp. 453-461, 2012.

[2] Lei Yu, Huan Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution", (ICML-2003), Washington DC, pp.1-8, 2003

[3] Yi Zhang, Chris Ding, Tao Li, "A Two-Stage Gene Selection Algorithm by Combining ReliefF and mRMR", IEEE Trans. Pattern Analysis and Machine Intelligence, 27, 2009.

[4] Huan Liu and Lei Yu, "Toward Integrating Feature Selection Algorithms for Classification and Clustering", IEEE Transactions on Knowledge and Data Engineering, Volume 17, Issue 4, Pages: 491 - 502 , 2005

[5] Li Zhuo , Jing Zheng, Fang Wang, Xia Li, Bin Ai and Junping Qian,"A Genetic Algorithm based Wrapper Feature selection method for Classification of Hyperspectral Images using Support Vector Machine", The International Archives

of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. XXXVII. Part B7,pp 397-402.
[6] H.Almuallim and T.G. Dietterich, "Algorithms for Identifying Relevant Features," Proc. Ninth Canadian Conf. Artificial Intelligence, pp. 38-45, 1992.
[7] H.Almuallim and T.G.Dietterich, "Learning Boolean Concepts in the Presence Irrelevant Features," Artificial Intelligence,vol.69, nos. 1/2, pp. 279-305,1994.
[8] A.Arauzo-Azofra J.M.Benitez, and J.L.Castro, "A Feature Set Measure Based on Relief," Proc.Fifth Int'l Conf.Recent Advances in Soft Computing, pp.104-109,2004.
[9] L.D. Baker and A.K. McCallum, "Distributional Clustering of Words for Text Classification," Proc. 21st Ann. Int'l ACM SIGIR Conf. Research and Development in information Retrieval, pp. 96-103, 1998.
[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] T. Cover. The best two independent measurements are not the two best. IEEE Trans. Systems, and Cybernetics, 4:116.117, 1974.

