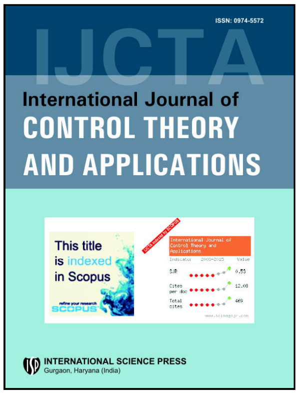

International Journal of Control Theory and Applications

ISSN : 0974-5572

© International Science Press

Volume 10 • Number 14 • 2017

Efficient Feature Subset Selection using Clustering and Correlation Measure

Smita Chormunge[1] and Sudarson Jena[2]

[1] *Department of Computer Science and Engineering GITAM University, Hyderabad, INDIA,*
Email: smita2728@rediffmail.com
[2] *Department of Information Technology GITAM University, Hyderabad, INDIA,*
Email: sudarsonjena@gitam.edu

Abstract: Finding useful data from the application, which contains ten to hundreds of attributes, is challenging issue in machine learning. Most of feature selection algorithms are fail to work with such high dimensional data. This paper addresses the dimensionality problem by proposing method where Clustering and filter measures work together to find relevant and non-redundant feature subset. Noisy and irrelevant features are eliminated by using k-means clustering algorithm and redundant features are remove from each cluster by applying correlation measure. Extensive experiments are carried out to compare proposed method with other well-known methods on UCI datasets with respect to k-nearest neighbor classifier. Empirical study demonstrate that proposed method is very efficient and effective in selecting good feature subset.

Keywords: High dimensional data, Clustering, feature selection, correlation, filter measure.

1. INTRODUCTION

Rapid development in today's technology, the applications such as image retrieval, information retrieval and text categorization contains vast amounts of multivariate data in terms of number of samples and attributes. To handle such high dimensional data is difficult task and it degrades the performance of learning algorithms. Data mining is a technique for extracting the useful knowledge from large amount of data. Finding useful information from the applications, which contains huge attributes, is challenging task. While performing operation and representing data, large number of features are collected because of unfamiliar relevant features. Irrelevant features do not make change in result and redundant features do not add anything to the target concept [1], but these irrelevant and redundant features significantly increase the computational cost of a learning process. When the dimension increases, it is computationally costly or practically prohibitive and diminish the accuracy of machine learning algorithms. Traditional feature selection methods fail to scale on such huge data with large attributes.

In machine learning, Feature selection is vital method to solve dimensionality problem by removing irrelevant, noisy and redundant features [2-3]. Selecting useful attributes with ease of feature selection, Machine-learning algorithms become more scalable, reliable and accurate. In feature subset selection to produce good feature subset, various evaluation measures and search techniques are used. Recent study report five categories evaluation measures such as uncertainty measures, distance measures, dependence measures, information entropy, consistency and classier error rate.

Numerous of feature selection method have been proposed for classification techniques [4-7]. Some of feature selection algorithms use statistical measures such as information gain measure, mutual information and correlation measure. Feature selection classified into main three approaches, which is based on evaluation measures that are Filters, Wrappers and Embedded methods [8]. According to recent study, clustering is also a one approach for feature selection. Cluster analysis group's elements in such way that elements in a group should be similar to one another and unrelated to the elements in other groups. One can consider that better the clustering when there is a greater the homogeneity within a group and greater the difference between groups [9]. Two popular clustering methods are K-means and hierarchical clustering methods.

To improve the performance of learning algorithms, feature selection method uses clustering methods, which demonstrated that application of cluster analysis has been more effective than traditional feature selection algorithms. To reduce the dimensionality C. Krier has presented a methodology by combining hierarchical constrained clustering of spectral variables and use of mutual information measures for selection of features [10]. [11] Author has proposed same methodology as Krier except that the former forces every cluster to contain consecutive features only. [12] In this paper Fast clustering based feature Selection algorithm (FAST) based on the MST method has been proposed. Clustering based strategy of FAST has a high probability of producing a subset of useful and independent features. Feature selection can simplify the calculation and help to get an accurate data model in data clustering [13]. Sotoca has proposed a feature selection supervised method based on feature clustering [14].

In this paper, new Feature Selection method is proposed to solve the dimensionality problem. K-means clustering method is use to group the similar features into clusters and removing irrelevant and noisy features. To find and eliminate redundant feature, correlation measure is apply on each cluster. Finally, features are ranked in decreasing order. Threshold value is apply to get representative feature subset. Empirical study performed on different UCI datasets with respect to K-nearest neighbor classifier. Proposed method is compare with ReliefF and CFS feature selection methods.

The section II of this paper presents the proposed feature selection method. The empirical study, experimental results and analysis are discussed in section III. Finally, section IV of this paper is discuss the concluding remarks.

2. PROPOSED FEATURE SELECTION METHOD

The application with huge number of features and samples is big issue to find the nature of features and select relevant and non-redundant features among such data. Performance of learning algorithms can be improve using by combining feature clustering and filter method than individual filter evaluation measures. To address the dimensionality issue, we proposed a feature selection method where clustering is integrating with correlation measure. K-means clustering method [15] find the nature of features and group the features based on distance function. In k-means, clustering algorithm user have to specify number of clusters. Then select arbitrarily the cluster center of each cluster. Find the closest features to center recalculate the center and mean value of cluster until no change in clustering data. Euclidean distance function is use to find similarity between features. After forming the clusters, the features, which not fit to any cluster, are eliminated in this step.

In second step, the redundant features of each cluster are identified and removed by using correlation measure. Correlation between features has been calculated by using equation (1) and (2). For feature X with values x and classes C with values c, where X, C are treated as random variables, Pearson's linear correlation coefficient is defined as:

$$(X,C)=\frac{E(XC)-E(X)E(C)}{\sqrt{\sigma^2(X)\sigma^2(C)}} \tag{1}$$

$$=\Sigma_i \frac{(xi-\overline{xi})-(ci-\overline{ci})}{\sqrt{\Sigma_i (xi-\overline{xi})^2\,\Sigma_j (ci-\overline{ci})^2}} \tag{2}$$

(X, C) value is -1 to +1 it is linearly dependent and zero if they are completely uncorrelated. Probability that two features are correlated is estimated using the error function. P(X ~ C) = erf(|(X,C)| $\sqrt{N/2}$). The feature list ordered by decreasing values of the P(X ~ C) may serve as feature ranking [16].

Threshold value is defined to cutoff the least value features. Correlation measure rank the relevant features, which are selected from the clusters. Features are ranked in decreasing order. These features are selected based on threshold, value of features, which are less than threshold value, is eliminated. Time complexity of proposed method is calculated based on k-means clustering algorithm and correlation measure is O (N) where T iterations performed on a sample size of number of instances, for number of attributes.

Algorithm

```
Input:    D ($f_1$, $f_2$, ……., $f_k$, $f_c$) // a training data set
n:        Number of Cluster
θ:        Threshold value
Output:   S //Representative feature subset
Procedure:
Step 1:   select arbitrarily initial cluster center k
Step 2:   calculate distance from each feature to each cluster k.
Step 3:   Based on cluster mean and similarity distance assign each feature to the closest cluster.
Step 4:   Compute the new mean value for each cluster
Step 5:   Repeat step 2 to 4 until there is no change in clusters.
Step 6:   delete the irrelevant features, which not fit to any clusters
Step 7:   for (i=0; i<=n; i++)
              {
                  $P=\Sigma_i \frac{(Fi-\overline{Fi})(Fci-\overline{Fci})}{\sqrt{\Sigma_i (Fi-\overline{Fi})^2\,\Sigma_j (Fci-\overline{Fci})^2}}$ //calculate correlation between features
              }
Step 8:   if p=1; delete one of feature.
Step 9:   Rank the relevant features {f1…. fn) based on decreasing order.
Step 10:  for (j=0; j<=n; j++)
              {
                 if ($f_j > \theta$)
                    S= $f_j$;
              }
Step 11:  stop process
```

3. EXPERIMENT AND RESULT

Empirical study performed on standard UCI datasets. Car, Vowel, Sponge, Lung Cancer, Sonar, Audiology and Arrhythmia datasets [20] are collected for experimental work. The summary of datasets is presented in table 1. Computational time and accuracy of proposed method is compared with CFS [19] and ReliefF [18] methods. Computational time is calculated in seconds. Accuracy of proposed method and other method is evaluated with respect to K-nearest neighbor classifier. Classification accuracy is analyzed using tenfold cross-validation strategy to have good estimations of the accuracy of the algorithms. Data mining tool Weka [17] is used for analyzing the results. Threshold value is defined as -1.79, which is default value of Weka.

Original datasets are first filter by applying k-means clustering algorithm on whole datasets. We have specified two number of clusters for grouping the features. By applying the correlation measure, calculated the number of clusters classified correctly. Execution time required to build up dataset is evaluated for all datasets using proposed method and other method. Computational time comparison of proposed method with other method is represented in table 2. Results shows that ReliefF method time complexity is more than proposed method and CFS method. Whereas proposed method takes less time to execute almost all datasets as compare to other both methods. Table 2 represent the Classifier accuracy comparison of proposed method with CFS and ReliefF methods. F-score based on precision and recall is used to estimate the classifier accuracy. The results display that the proposed method is very effective to select relevant and non-redundant feature subset than comparative methods for all datasets. ReliefF method perform well in accuracy but takes more time to execute data. CFS method selects only representative features whereas ReliefF and proposed method rank the features using ranker search.

Graphical representation of computational time and classifier accuracy of proposed method against other method is shown in fig.1and 2 respectively.

Table 1
Summary of datasets

Datasets	*Instances*	*Features*
Car	1728	7
Vowel	990	14
Sponge	76	46
Lung Cancer	32	57
Sonar	208	60
Audiology	226	70
Arrhythmia	452	280

Table 2
Computational time comparison of proposed method with other methods

Datasets	*Proposed Method*	*CFS*	*ReliefF*
Car	0.01	0.05	0.78
Vowel	0.03	0.03	0.7
Sponge	0.01	0.03	0.01
Lung Cancer	0.01	0.02	0.02
Sonar	0.01	0.05	0.14
Audiology	0.10	0.03	0.20
Arrhythmia	0.03	0.28	2.95

Table 3
Classifier accuracy comparison of proposed method with other methods

Datasets	*Proposed Method*	*CFS*	*ReliefF*
Car	98.3	92.5	92.5
Vowel	100.0	97.2	99.3
Sponge	100.0	91.5	91.4
Lung Cancer	96.8	67.4	67.4
Sonar	95.2	84.6	86.5
Audiology	95.6	68.9	75.5
Arrhythmia	95.1	59.3	50.5

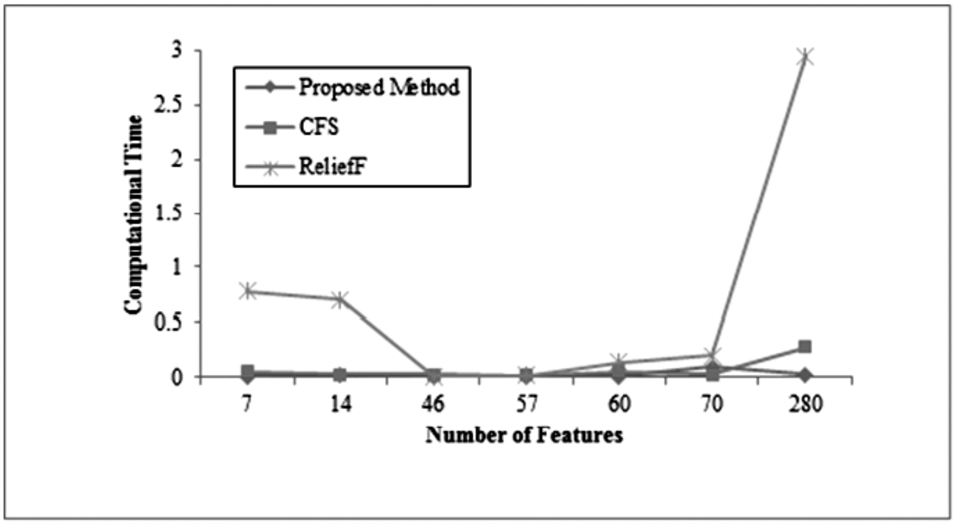

Figure 1: Comparison of computational time of proposed method against other methods

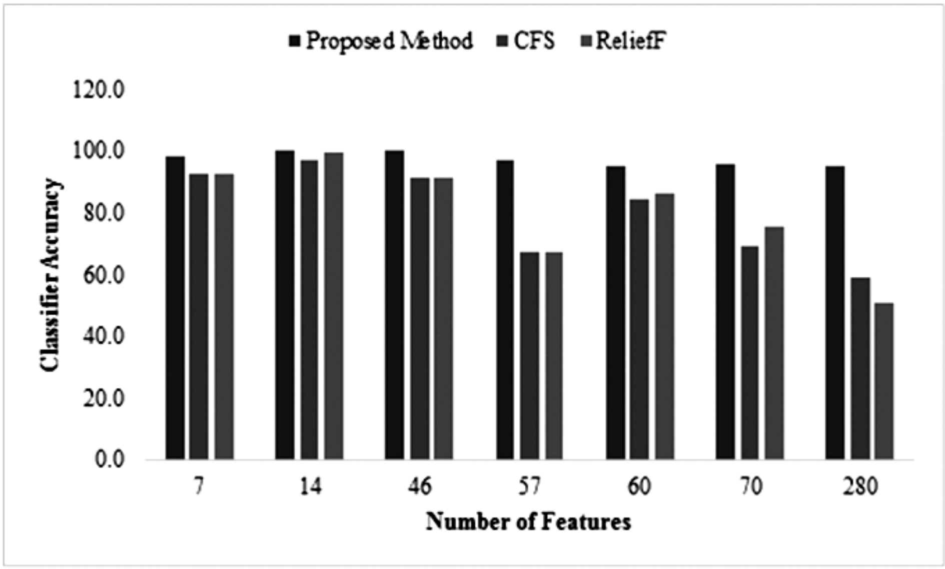

Figure 2: Comparison of classifier accuracy of proposed method against other methods

4. CONCLUSION

In this paper, new Feature Subset Selection method is proposed to solve the dimensionality problem. Clustering and correlation measure is integrating to find and eliminate irrelevant and redundant features. Computational time and accuracy is evaluated for UCI datasets using k-nearest neighbor classifier. Experimental results demonstrated that proposed method is reasonably efficient and effective to select good feature subset than other feature selection methods. In future, explore the method to select representative features directly without ranking the features to reduce the dimensionality and to increase the predictive accuracy of the model.

REFERENCES

[1] John, G. H., Kohavi, R. and Pfleger, K, "Irrelevant features and the subset selection problem", *Proc. the Eleventh International Conference on Machine Learning, 121-129,1994.*

[2] M. Dash and H. Liu, "Feature Selection for Classification", *Intelligent Data Analysis, vol. 1, no. 3, pp. 131-156,1997.*

[3] Liu, H. and Yu, L, "Toward Integrating Feature Selection Algorithms for Classification and Clustering", *IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 4, pp. 491-502,2005.*

[4] H. Frohlich, O. Chapelle, B. Scholkopf., "Feature selection for support vector machines by means of genetic algorithm, in: Tools with Artificial Intelligence", *Proceedings. 15th IEEE International Conference on, IEEE, pp. 142–148,2003.*

[5] S.-W. Lin, K.-C. Ying, C.-Y. Lee, Z.-J. Lee, "An intelligent algorithm with feature selection and decision rules applied to anomaly intrusion detection", *Applied Soft Computing 12 (10) 3285–3290,2012.*

[6] L. Yu, H. Liu, "Redundancy based feature selection for microarray data", *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp. 737–742,2004.*

[7] D. K. Bhattacharyya, J. K. Kalita, Network Anomaly Detection: *A Machine Learning Perspective, CRC Press,2013.*

[8] I. Guyon and A. Elisseeff., "An Introduction to Variable and Feature Selection", *J. Machine Learning Research, vol 3, pp. 1157-1182,2003.*

[9] Michael Steinbach, Levent Ertöz, Vipin Kumar, "The Challenges of Clustering High Dimensional Data.in New Vistas in Statistical Physics – Applications in Econophysic Bioinformatics, and Pattern Recognition*", Springer-Verlag, 2004.*

[10] C. Krier, D. Francois, F. Rossi, M. Verleysen, "Feature Clustering and Mutual Information for the Selection of Variables in Spectral Data", *Proc. European Symp. Artificial Neural Networks Advances in Computational Intelligence and Learning, pp. 157-162,2007.*

[11] G. Van Dijck, M.M. Van Hulle, "Speeding Up the Wrapper Feature Subset Selection in Regression by Mutual Information Relevance and Redundancy Analysis", *Proc. Int'l Conf.Artificial Neural Networks,2006.*

[12] Qinbao Song, Jingjie Ni, and Guangtao Wang A, "Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data", *IEEE Transaction on knowledge and Data Engineering, Vol. 25, No. 1,2013.*

[13] Yu-MengXu , Chang-DongWang , Jian-Huang, "Weighted Multi-view Clustering with Feature Selection", *Pattern Recognition,53, pp-25-35,2016.*

[14] Sotoca JM, Pla F," Supervised feature selection by clustering using conditional mutual information based distances",*Pattern Recogn 43(6):325–343,2010.*

[15] Onoda, T., Sakai, M., "Independent component analysis based seeding method for k-means clustering",*IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, doi:10.1109/WI-IAT.2011.29,2011.*

[16] Biesiada J., DuchW, "Feature selection for high-dimensional data Pearson redundancy based filter", *Advances in Soft Computing, 45, pp 242-249,2008.*

[17] Remco R. Bouckaert,Eibe Frank,Mark Hall, Richard Kirkby, Peter Reutemann, Alex Seewald, David Scuse*, WEKA Manual for Version 3-7-10,2013.*

[18] K. Kira and L.A. Rendell, "The Feature Selection Problem: Traditional Methods and a New Algorithm",*Proc. 10th Nat'l Conf. Artificial Intelligence, pp. 129-134,1992.*

[19] M. A. Hall. “Correlation-based Feature Subset Selection for Machine Learning”. PhD thesis, University of Waikato, Hamilton, New Zealand, 1998.

[20] Datasets can be downloaded from: *http://repository.seasr.org/Datasets/UCI/arff/*