International Journal of Computer Science
Oddity...Probe...Reviste...

http://www.ijcsjournal.com Volume 2, Issue 1, No 4, 2014. ISSN: 2348-6600.
Reference ID: IJCS-050. PAGE NO: 287-292.

An Improved Feature Subset Selection with Correlation Measures using Clustering Technique

Ms.S.Sasikala

PG Student
Department of CSE
Bharathiyar Institute of Engineering for Women
Salem, Tamil Nadu, India
Sasibtech05@gmail.com

***Abstract*— Clustering techniques are used to partition the transaction data values. Vector based similarity models are suitable for low dimensional data values. High dimensional data values are clustered using subspace clustering method. Feature selection involves identifying a subset of the most useful features that produces compatible results as the original set of features. In this paper the clustering technique are used to improve the feature subset selection with correlation measure. Based on these criteria, a fast clustering based feature selection algorithm, FAST, is proposed and experimentally evaluated in this paper. A feature selection algorithm is constructed with the consideration of efficiency and effectiveness factor. The efficiency concerns the time required to find a subset of features. The effectiveness is related to the quality of the subset of features. Fast clustering-based feature selection algorithm (FAST) is used to cluster the high dimensional data. The feature selection process is improved with correlation measures. Redundant feature filtering mechanism is used to filter the similar features and also the custom threshold is used to improve the clustering accuracy.*Index Terms*—Component, formatting, style, styling, insert.**

Keywords-Feature clustering, Feature subset selection, redundant filtering, filter method.

I. Introduction

Feature subset selection is an effective way for reducing dimensionality, increasing learning accuracy, removing irrelevant and redundant data, and improving result comprehensibility [5],[11]. The main aim of feature selection (FS) is to determine the relevant feature in the high dimensional data. Usually FS algorithms involve heuristic or random search strategies in an attempt to avoid this prohibitive complexity. They can be divided into four broad categories: the embedded, filter, and hybrid approaches.

The wrapper methods use the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subsets. The embedded methods incorporate feature selection as a part of the training process and are usually specific to given leaning algorithms.[23].The wrapper methods are computationally expensive and tend to over fit on training sets[13],[14].The filter methods are independent of learning algorithm with good generality. The filter methods, in addition to the generality, are usually a good choice when the number of features is very large.[20]. The general graph theoretic clustering is simple: compute a neighborhood graph of instances, then delete any edge in the graph that is much longer/shorter (according to some criterion) than its neighbors. In cluster analysis, graph-theoretic methods have been well studied and used in many applications. In our study, we apply graph-theoretic clustering methods to the features. Based on the minimum spanning tree (MST) method, we propose a Fast clustering- based feature Selection algorithm (FAST).

We adopt the 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. 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 and to form the final subset of features.

II. RELATED WORK

Feature selection is frequently used as a preprocessing step to machine learning. Feature subset selection can be viewed as the process of identifying and removing many irrelevant and redundant features. This is because: (i) irrelevant features do not contribute to the predictive accuracy [13], and (ii) redundant features do not rebound to getting a better predictor for that they provide mostly information which is already present in other features.

Hierarchical clustering has been adopted in word selection in the context of text classification [19],[23].Distributed clustering has been used to cluster words into groups based either on their participation in particular grammatical relations with other words by Lee et al[24].Distributed clustering of words are in nature, and result in sub optimal word clusters and high dimensional cost Kumar et al [18].

III. Feature Subset Selection Algorithm

In this section we discuss about how to find the relevant features from the high dimensional data. The irrelevant features and redundant features are severely affect the accuracy of the learning a machines.[11][21].Feature subset selection algorithm are able to identify and remove the irrelevant and redundant information. Moreover, "good feature subsets contains features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other."

Feature subset selection can be viewed as the process of classifying and removing as many unrelated and completed features as possible. Feature selection methods are used to identify the key features from a data collection. Irrelevant feature identification and redundant feature filtering methods are used in the feature selection process. Correlation similarity measures are used for the relationship analysis. Fast clustering method is used to perform data clustering with feature selection process. Minimum Spanning Tree is used to construct the tree with transaction values.

The feature selection process is improved with a set of correlation measures. Dynamic feature intervals can be used to distinguish features. Redundant feature filtering mechanism is used to filter the similar features. Custom threshold is improved with clustering accuracy. The graph based clustering algorithm is designed with MST. Feature selection process is enhanced with dynamic threshold values. Mutual information measures how much the distribution of the feature values and target classes differ from statistical independence.

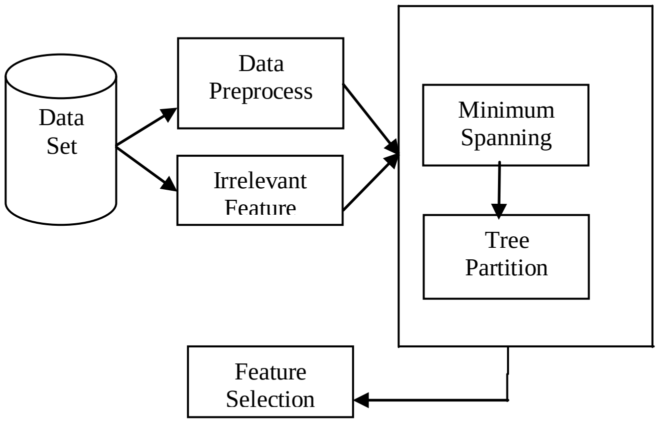

Fig 1: Framework of the proposed feature subset selection algorithm.

The symmetric uncertainty is defined as follows

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

Where,

1) H(X) is the entropy of a discrete random variable X. Suppose p(x) is the prior probabilities for all values of X,H(X) is defined by

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \quad (2)$$

2) Gain (X|Y) is the amount by which the entropy of Y decreases. It reflects the additional information about Y provided by X and is called the information gain which is given by

$$Gain(X|Y) = \frac{H(X) - H(X|Y)}{H(Y) - H(Y|X)} \quad (3)$$

Where H (X|Y) is the conditional entropy which quantifies the remaining entropy (i.e. uncertainty) a random variable X given that the value of another random variable Y is known. Suppose p(x) is the prior probabilities for all values of X and p (x|y) is the posterior probabilities of X given the values of Y, H (X|Y) is defined by

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

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

Definition1(T-Relevance) The relevance between any pair of feature Fi F and the target concept C is referred to as the T-Relevance of Fi and C denoted by SU(F_i, C).

Definition 2(F-Correlation) The correlation between any pair of feature F_i and F_j(F_i, $F_j \in F \wedge I \neq j$) is called the F-Correlation of F_i and F_j, and denoted by SU (F_i, F_j).

Definition3 (F-Redundancy) Let S= {F_1,F_2,…F_i, …, F_k< |F|} be a cluster of features. If ∃ Fj ∈ S,SU(F_j, C)> SU(F_i, C) SU(F_i, F_j)> SU(F_i, C) is always corrected for each F_i S(i j),then Fi are redundant features with respect to the given F_j(i.e each F_i is a F-Redundancy)

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 PROPOSED SYSTEM

In this section, we discuss how to evaluate the goodness of features for classification. In general, a feature is good if it is relevant to the class concept but is not redundant to any of the other relevant features. If we adopt the correlation between two variables as a goodness measure, the above definition becomes that a feature is good if it is highly correlated to the class but not highly correlated to any other features. In other words, if the correlation between a feature and the class is high enough to make it relevant to (or predictive of) the class and the correlation between it and any other relevant features does not reach a level so that it can be predicted by any of the other relevant features, it will be regarded as a good feature for the classification task.

There exist broadly two approaches to measure the correlation between two random variables. One is based on classical linear correlation and the other is based on information theory. Under the first approach the most well-known measure is linear correlation coefficient. There are several benefits of choosing linear correlation as a feature goodness measure for classification.

First, it helps removes feature with near zero linear correlation to the class. Second, it helps to reduce redundancy among selected features. It is known that if data is linearly separable if all but one of a group of linearly dependent features is removed. However it is not safe to always assume linear correlations that are not linear in nature. Another limitation is that the calculation requires all features contain numerical values.

To overcome these shortcomings, in our solution we adopt the other approach and choose a correlation measure based on the information theoretical concept of entropy, a measure of the uncertainty of a random variable. Using symmetrical uncertainty (SU) as the goodness measure, we are now ready to develop a procedure to select good features for classification based on correlation analysis of features for (including the class). This involves two aspects (1) how to decide whether a feature is relevant to the class or not; and (2) how to decide whether such a relevant feature is redundant or not when considering it with other relevant features.

The answer to the first question can be using a user defined threshold SU value, as the method used by many other feature weighting algorithms (e.g., Relief). More specifically, suppose a dataset S contains N features and C class. Let SU_{ic} denote the SU value that measures the correlation between a feature Fi and class C (named C-correlation), then a subset S' of relevant features can be decided by a threshold SU value δ.The answer to the second question is more complicated because it may involve analysis of pairwise correlation between all features (named F-correlation), which results in time complexity of $0(N^2)$ associated with the number of features N for most existing algorithms.

Since F-correlation is also captured by SU values, in order to decide whether a relevant feature is redundant or not, we need to find a reasonable way to decide the threshold level for F-correlations as well. In other words, we need to decide whether the level of correlation between two features in S' is high enough to cause redundancy so that one of them may be removed from S'. In the context of a set of relevant features S' already identified for the class concept, when we try to determine the highly correlated features for a given feature Fi and the class concept, SU $_{I,c}$ as a reference. Therefore, even the correlation between this feature and the class concept is larger than some threshold value and therefore making this Feature relevant to the class concept, this correlation is by no means predominant.

Definition4: (Predominant Feature). A feature is predominant to the class, if its correlation to the class is predominant or can become predominant after removing its redundant peers. According to the above definitions, a feature is good if it is predominant in predicting the class concept, and feature selection for classification s a process that identifies all predominant features and remove redundant ones among all relevant features, without having to identify all the redundant peers for every feature in S', and thus avoids pairwise analysis of F-correlation between all relevant features.

V RESULT AND ANALYSIS

In this section we present the experimental results in terms of the proportion of selected features, the time to obtain the feature subset, the classification accuracy. 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.

In order to further explore feature selection algorithms whose reduction rates have statistically significant differences, we performed a Nemenyi test.The results indicate that the proportion of selected features of FAST is statistically smaller than those of Relief-F, CFS andFCBF, and there is no consistent evidence to indicate statistical difference between FAST, Consist, and FOCUS-SF, respectively. For real world data, we often do not have such prior knowledge about the optimal subset, so we use the predictive accuracy on the selected subset of feature.

Generally the individual evaluation based feature selection algorithms of FAST, FCBF and Relief are much faster than the subset evaluation based algorithms of CFS, Consist and FOCUS-SF.FAST is consistently faster than all other algorithms. The runtime of FAST is only 0.1 % of that of CFS 2.4 % of that of Consist, 2.8 % of that of FOCUS-SF, 7.8% of that of Relief, and 76.5% of that of FCBF, respectively. The Win/Draw/Loss records show that FAST outperforms other algorithms as well.

Table 1: Runtime (in ms) of the six feature selection algorithms

Dataset	FAST	FCBF	CFS	Relief	Consist	FOCUS-SF
Chess	105	60	345	12660	1990	653
Coil	1472	716	938	13918	3227	660
Elephant	866	875	1483	30416	53845	1282
Arrhy	783	312	905	1072	3492	1098
Colon	166	115	821	744	1360	2940
Ar10p	706	458	736	7945	1624	1032
Pie10p	678	1223	1224	3874	57934	960
Oh0.wc	5283	5990	6650	7636	3568	4689
Oh10.wc	5549	6033	1034	4898	4149	8446
B-cell1	160	248	9345	5652	4882	3421
b-cell2	626	1618	4524	4166	5102	1273
b-cell3	635	2168	3415	5102	2914	8446
Average	3573	4671	5456	4580	7490	5227
Win/draw/ loss	-	22/0/3	31/0/1	20/0/6	35/0/0	34/0/1

From the analysis above we can know that FAST performs very well on the microarray data. The reason lies in both the characteristics of the dataset itself and the property of the proposed algorithm. Microarray data has the nature of the large number of features (genes) but small sample size, which can cause "curse of dimensionality" and over-fitting of the training data [22]. For the purpose of exploring the relationship between feature selection algorithms and data types, i.e. which algorithm are more suitable for which types of data, we rank the six feature selection algorithms according to the classification accuracy of the feature selection method. Therefore, selecting a small number of discriminative genes from thousands of genes is essential for successful sample classification [21], [24].

Our proposed FAST effectively filters out a mass of irrelevant features in the first step. This reduces the possibility of improperly bringing the irrelevant features into the subsequent analysis. Then in the second step FAST removes redundant features by using the redundant filtering mechanism. Our proposed FAST also requires a parameter θ that is the threshold of feature relevance. Different θ values might end with different classification results.

VI CONCLUSION

In this paper, we propose a novel concept of predominant correlation, introduce an efficient way of analyzing feature redundancy, and design a fast correlation based filter approach and 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. We compared the performance of the proposed algorithm with those of the five feature selection algorithm FCBF, CFS, Relief, Consist, and FOCUS-SF. The above experimental results suggest that the feature selection for classification on high dimensional data are efficiently achieving the high degree of dimensionality reduction and enhance classification accuracy with predominant correlation features.

For the future work, we plan to enhance the correlation measure to improves the searching efficiency will modify the Fast algorithm.

REFERENCES

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

[2] D.A. Bell and H. Wang., "Formalism for Relevance and its Application in Feature Subset Selection," Machine Learning, 41(2), pp 175-195,2000.

[3] M. Dash, H. Liu and H. Motoda ., "Consistency based Feature Selection", In Proceeding of the fourth Pacific Asia Conference on Knowledge Discovery and Data Mining, pp 98-109,2000.

[4] S. Das., "Filters, Wrappers And A Boosting Based Hybrid For Feature Selection", In Proceeding Of The 18th International Conference On Machine Learning, pp 74-81, 2001.

[5] I.S Dhillon., S .Mallela. and R .Kumar., "A Divise Information Theoretic Feature Clustering Algorithm For Text Classification," J. Mach. Learn. Res., 3,, pp 1265-1287, 2003.

[6] F .Fleuret., "Fast Binary Feature Selection with Conditional Mutual Information, Journal of Machine" Learning Research, 5,Pp 1531-1555, 2004.

[7] G. Forman., "An Extensive Empirical Study of Feature Selection Metrics for Text Classification". Journal of Machine Learning Research, 3, pp 1289-1305, 2003.

[8] F. Herrera, and S. Garcia., "An Extension on Statistical Comparison of Classifiers over Multiple Datasets for All Pairwise Comparisons". J.Mach. Learn. Res., 9,pp 2677-2694, 2008.

[9]I. Guyn . and A. Elisseeff ., "An Introduction to Variable and Feature Selection". Journal of Machine Learning Research, 3 pp 1157-1182, 2003.

[10] M.A Hall ., "Correlation Based Feature Subset Selection for Machine Learning", Ph.D Dissertation Waikato, New Zealand: Univ. Waikato,1999.

[11] C. Krier C. D. Francois., F. Rossi, and M. Verleysen., "Feature Clustering And Mutual Information For The Selection Of Variables In Spectral Data". In Proc European Symposium On Artificial Neural Networks Advances In Computational Intelligence And Learning, pp 157-162, 2007.

[12] M. Last., A. Kandel. And O. Mainmom., "Information Theoretic Algorithm For Feature Selection, Pattern Recongnitition Letters", 22(6-7),pp 799-811,2001.

[13] L.C Monila, L. Belanche. And A. Nebot., "Feature Selection Algorithms: A Survey and Experimental Evaluation", In Proc. IEEE Int.Conf. Data Mining. ,pp 306-313, 2002.

[14] H. Park, and H. Kwon., "Extended Relief Algorithm In Instance Based Feature Filtering", In Proceeding Of The Sixth International Conference On Advanced Language Processing And Web Information Technology (ALPTI 2007), pp 123-128, 2007.

[15]B. Raman. And T.R Loerger,, "Instance Based Filter for Feature Selection Journal of Machine Learning Research, 1, pp 1-23, 2002.

[16] M. Robnik-Sikonja. and I. Kononenko., "Theoritic and Empirical Analysis of Relief and Relief", Machine Learning Research, 53, pp 23-69, 2003.

[17] P. Scanlon., G. Potamiano., "Mutual information based visual feature selection for lip-reading, in int. conf. on spoken language processing, 2004.

[18] Scherf M. and Brauer W., Feature Selection By Means Of A Feature Weighting Approach, Technical Report FKI-221-97, Institute Fur Informatics, and Technics Universidad Munched 1997.

[19] C. Sha, X.Quiu. And A. Zhou., "Feature Selection Based On A New Dependency Measure," 2008 Fifth International Conference On Fuzzy Systems And Knowledge Discovery, 1, pp 266-270, 2008.

[20] J. Souza., "Feature Selection with A General Hybrid Algorithm, Ph.D, University Of Ottawa, Ottawa, Ontario, Canada, 2004.

[21] G. Van Dijk. and M.M Van Hulle .," Speeding Up the Wrapper Feature Subset Selection in Regression by Mutual Information Relevance and Redundancy Analysis, "International Conference on Artificial Neural Networks, 2006.

[22] G.I Webb., "Multi boosting for Combining Boosting and Wagging, Machine Learning, 40(2), pp 159-196, 2000.

[23] E. Xing., M. Jordan. And R. Karp, "Feature Selection for High Dimensional Genomic Microarray Data, "In Proceedings of the Eighteenth International Conference on Machine Learning, pp 601-608, 2008.

[24] J. Yu., S.S.R. Adipi. And P.H Artes., "A Hybrid Feature Selection Strategy For Image Defining Features: "Towards Interpretation Of Optic Nerve Images, In. Proceedings Of 2005 International Conference On Machine Learning And Cybernetics, 8, pp 5127-5132, 2005.

[25] L. Yu. And H. Liu H., "Feature Selection For High Dimensional Data: Fast Correlation Based Filter Solution, "In Proceedings of 20th International Conferences On Machine Learning, 20(2).