A Pragmatic Application of Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data

[1] N.Narendra Reddy , [2] K. Narayana

[1] Post-Graduate Student, Department of Computer Science and Engineering,
SIT, PUTTUR, India

[2] Head & Associate Professor, Department of computer Science and Engineering,
SIT, PUTTUR, India

Abstract - Using the rapid growth of computational biology and e-commerce applications, high-dimensional data becomes usual. Thus, mining high dimensional data is an urgent problem of great practical importance. Within the high dimensional data the dimensional reduction is a vital factor, to the purpose the clustering based feature subset selection algorithm is proposed in this particular paper. The characteristics are actually clustered Based on the class labels. The Relevance on the clustered features has become evaluated. The correlation on the relevant clustered feature will be evaluated. This technique improved by cluster based FAST Algorithm and Fuzzy Logic. FAST Algorithm can often Identify and taking out the irrelevant data set. This algorithm process implements using two different steps which are graph theoretic clustering methods and representative feature cluster is selected. Feature subset selection researchers have centered on in search of relevant features. The proposed fuzzy logic has focused on minimized redundant data set and improves the feature subset accuracy.

***Keywords* - Clustering, Fuzzy logic, Biology, E-commerce**

1. Introduction

The particular performance, robustness, along with usefulness of classification algorithms are improved when relatively few characteristics are likely to complete the classification. Thus, selecting relevant characteristics with all the construction of classifiers has taken lots of attention. With all the current intent behind choosing the subset of an good features Based on the target concepts, feature subset selection can be an efficient way for reducing dimensionality, removing inconsequential data, increasing learning reliability, and improving result comprehensibility Many feature subset selection methods are already proposed and studied for machine learning applications. Feature selection algorithms might be broadly classified into the filter model along with the wrapper model. The filter model is dependent upon general characteristics of the training data to select some features without involving any learning algorithm. The wrapper model requires one predetermined learning algorithm in feature selection and uses its performance to evaluate and determine which features are selected. With regards to each new subset related to features, the wrapper model would have to practice a hypothesis (or simply a classifier).

It is likely to give superior performance mainly because it finds features better suitable for the predetermined learning algorithm; it's also frequently more computationally expensive. After the number of features becomes very big, the filter model is usually choice due to computational efficiency. In filter feature selection methods FAST uses MST based clustering. Two key features about MST representation are:

(1) Simple structure of tree facilitates efficient implementation of clustering algorithms which otherwise are highly computationally challenging,

(2) It does definitely not depend on a detailed type of a cluster.

Classical clustering algorithms depend on either idea of grouping data around several centres or thought of separating data point using some regular geometric curve. They don’t behave well when the boundaries while using clusters are complex. MST based technique completely independent on the geometric changes through the boundaries of clusters. Hence shape complexity of clusters is incredibly rare in MST based clustering algorithms.

Feature as a group for suitability is evaluated by using a subset selection a subset of features. Feature subset selection methods are separated into Wrappers, Filters, Embedded and Hybrid methods. Embedded techniques are embedded in and particular in to a model. Wrappers Start using a search algorithm to get through the entire space of possible features and evaluate every subset simply by using a model inside the subsets. Wrappers are computationally expensive which has a risk greater than fitting towards model. Filters are similar to Wrappers from the search approach, but alternatively of evaluating a filter against a model, an easier filter is evaluated.

Two popular filter metrics for classification problems correlation and mutual information, although are both incorrect metrics. You will find, however, true metrics that happen to be functions within the mutual information. Other available filter metrics are: Correlation-based feature selection, Consistency-based feature selection, and Class separability, as well as Error probability, Inter class distance, probabilistic distance, and Entropy.

2. Related Works

2.1 Clustering Data

Cluster analysis groups data objects dependent upon only information within the information, that describes the object and their relationship .The goal is frequently which the objects having a group be just like the other and various from the objects in other groups. The higher the similarity that has a group as well as the greater distinction between groups, the higher quality or maybe more distinct the clustering.

2.2 Clustering High Dimensional Data

Clustering high-dimensional information is the cluster analysis expertise with anytime from various dozen to several a lots of dimensions. Such high-dimensional data spaces are generally encountered in areas as an example medicine, where DNA, microarray technology could certainly create large Number of measurements simultaneously, along with the clustering of text documents, where, when the word-frequency vector may be used, the amount of dimensions equals the size of the dictionary.

2.3 Irrelevant Features

Irrelevant features provide no useful information in virtually any context. The whole process of identifying and removing several irrelevant data features as they can be. That is caused by irrelevant features usually do not produce the predictive accuracy. By removing irrelevant data you have immediate gains for instance increased query performance and reduced storage requirements.

2.4 Redundant Features

Redundant features ordinarily do not redound that may receive an improved predictor towards the they provide mostly information which may be already contained in other feature(s). Then redundancy of information can be quite a known approach to obtain inconsistency, since customer might appear with some other values for given attribute.

2.5 Feature Selection

Feature selection in supervised learning have been well studied, the location where the absolute goal is to locate a feature subset that produces higher classification accuracy. Feature selections in unsupervised learning, learning algorithms are supposed to find natural grouping while using the examples within the feature space. Thus feature selection in unsupervised learning aims to name a fantastic subset of features that forms the best quality of clusters just for specific a number of clusters.

3. Fuzzy Based Feature Subset Selection Architecture

3.1 Architecture

Irrelevant features, together with redundant features, severely affect the truth with the learning machines. Thus, feature subset selection will be able to identify and take away because the irrelevant and redundant information as is possible. The cluster indexing and document assignments are repeated periodically to make up churn in order to maintain an up-to-date clustering solution. The k-means clustering technique and SPSSTool in order to develop an actual a serious amounts of online system to get a particular supermarket to calculate sales in a variety of annual seasonal cycles. The classification was based on nearest mean.

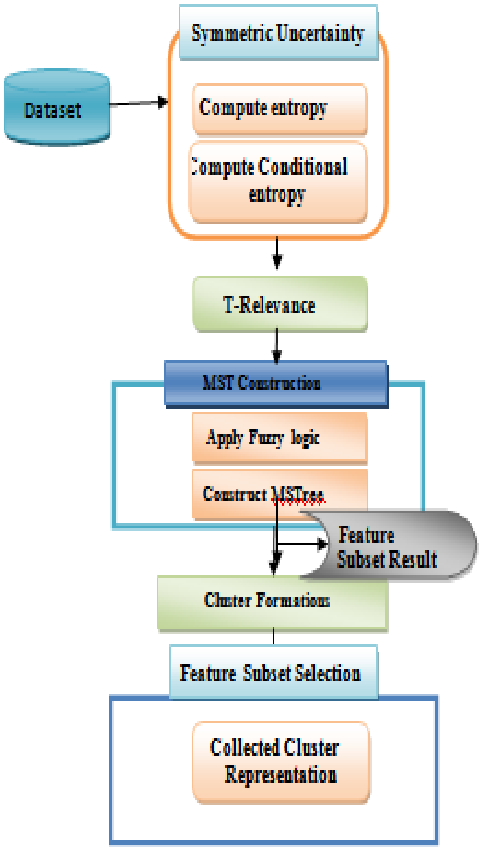

Fig: 3.1 Architecture

To be able to more precisely introduce the algorithm, and furthermore, as our proposed feature subset selection framework involves irrelevant feature removal and redundant feature elimination.

3.2 Feature Subset Selection Algorithm

Moreover, “good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) the other. Keeping these as the primary goal, we establish a novel algorithm which often can wisely manage both irrelevant and redundant features, and acquire an excellent feature subset.

We accomplish that by having a new feature selection framework which consisting of each connected components of irrelevant feature removal and redundant feature elimination. The previous obtains features highly relevant to the objective concept by reducing irrelevant ones, as well as the latter removes redundant features from relevant ones via choosing representatives from different feature clusters, and therefore produces the last subset. The irrelevant feature removal is straightforward when the right relevance is through defined or selected, even though the redundant feature elimination is of sophisticated.

In this proposed FAST algorithm, it involves

(a) The making of the minimum spanning tree (MST) from the weighted complete graph;
(b) The partitioning from the MST right into a forest with each tree representing a cluster; and
(c) Selecting representative features through the clusters.

So that you can more precisely introduce the algorithm, and since our proposed feature subset selection framework involves irrelevant feature removal and redundant feature elimination, we firstly present the original definitions of relevant and redundant features, and then provide our definitions dependant on variable correlation as follows. John et al. presented a specification of relevant features.

3.3 Fast Algorithm

The proposed FAST algorithm logically consists of three steps:

(i) Removing irrelevant features,
(ii) Constructing a MST from relative ones,
(iii) Partitioning the MST and selecting

Inputs: D ($F1$, $F2$, Fm, C) - the given data set
θ- the T-Relevance threshold.

Output: S - selected feature subset.

Part 1: Irrelevant Feature Removal

```
Step 1: for i = 1 to m do
Step 2: T-Relevance = SU ($Fi$, $C$)
Step 3 if T-Relevance >$\theta$then
```

Part 2: Minimum Spanning Tree Construction

```
Step4 : S = S ∪ {$Fi4$};
Step 5: G = NULL; //G is a complete graph
Step 6: for each pair of features {$Fi$,} ⊂ S do
Step 7: F-Correlation = SU (,$j$)
Step8:$AddF'iand/orF'jtoG$ $with$ F-Correlation $as$ $the$ $weight$ $of$ $the$ $corresponding$ $edge$;
Step 9: minSpanTree = Prim (G); //Using Prim Algorithm to generate the minimum spanning tree Part 3: Tree Partition and Representative Feature Selection
Step 10: Forest = minSpanTree
Step 11: for each edge $Eij$11 ∈Forest do
Step 12: if SU ($F'i$, $F'j$) <SU ($F'i$, $C$) ∧SU ($F'i$, $F'j$) <SU ($F'j$,$C$) then
Step 13: Forest = Forest –$Eij$
Step 14: S = $\phi$ 15 for each tree $Ti$
Step 15:∈Forest do
Step 16:$FjR$= argmax$F'k$∈$Ti$SU ($F'k$,$C$) 17 S = S ∪ {$FjR$}; 18 return S FAST Algorithm
```

FAST executes quite well about the microarray data. The reason is based on both the top features of the information set itself and the property with the projected algorithm. Microarray data has got the environment with the large numbers of characteristics other than small sample size, which could cause curse of dimensionality. Within the presence of numerous features, researchers become aware of which it is general which a high number of characteristics usually are not instructive because they are moreover inappropriate or superfluous according to the class concept.

Consequently, selecting a small Variety of discriminative genes from numerous genes is essential for booming sample categorization. Our projected FAST efficiently filters out quite a few inappropriate features within the initial step which reduces the likelihood of inappropriately bringing the inappropriate features in to the succeeding analysis. Then, within the subsequent step, FAST eliminate a lot of outmoded Features by way of selecting a single representative characteristic from each cluster of outmoded features. Consequently, simply a very small amount of discriminative characteristics are selected.

4. Conclusions

In this particular paper, we have proposed a clustering-based feature subset selection algorithm for high dimensional data. The algorithm involves 1) removing irrelevant features, 2) constructing the absolute minimum spanning tree from relative ones, and 3) partitioning the MST and selecting representative features.

The projected feature subset selection algorithm FAST was tested as well as the investigational results demonstrate that, evaluated along with other various kinds of feature subset selection algorithms, the projected algorithm not only decrease the quantity of features, and also advances the performances with the renowned various kinds of classifiers.

From the proposed algorithm, a cluster contains features. Each cluster is treated as being a single feature and therefore dimensionality is drastically reduced.

5. Future Work

In the future work, we decide to explore a variety of correlation measures, and study some formal properties of feature space.

References

[1] L. Kaufman, and P.J. Rousseeuw (1990) Finding groups in data: An introduction to cluster analysis. John Wiley and Sons, New York.

[2] J. Daxin, C. Tang and A. Zhang (2004) Cluster analysis for Gene expression data: A survey, IEEE Transaction on Knowledge and Data Engineering, Vol. 16 Issue 11, pp. 1370-1386.

[3] R. Agrawal, J. Gehrke, D. Gunopulos and Raghavan (1998) Automatic subspace clustering of high dimensional data for data mining applications, In Proceedings of the SIGMOD, Vol. 27 Issue 2, pp. 94-105.

[4] M. Steinbach, L. Ertöz and V. Kumar, “The challenges of clustering high dimensional data”, [online] available : http://www.users.cs.umn.edu/~kumar/papers/high_dim_clustering_19.pdf

[5] J. Gao, P. W. Kwan and Y. Guo (2009) Robust multivariate L1 principal component analysis and dimensionality reduction, Neurocomputing, Vol. 72: 1242-1249.

[6] A.Jain and R. Dubes (1988) Algorithms for clustering data, Prentice Hall, Englewood Cliffs, NJ.

[7] K. Fukunaga, (1990) Introduction to statistical pattern recognition, Academic Press, New York.

[8] G. Strang (1986) Linear algebra and its applications. Harcourt Brace Jovanovich, third edition.

[9] A.Blum and P. Langley (1997) Selection of relevant features and examples in machine learning, Artificial Intelligence, Vol. 97:245–271.

[10] H. Liu and H. Motoda (1998), Feature selection for knowledge discovery & data mining, Boston: Kluwer Academic Publishers.

[11] J. M. Pena, J. A. Lozano, P. Larranaga and Inza, I. (2001) Dimensionality reduction in unsupervised learning of conditional gaussian networks, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23(6):590 - 603.

[12] L. Yu and H. Liu, (2003), Feature selection for high dimensional data: A fast correlation based filter solution, In Proceedings of the Twentieth International Conference on Machine Learning, pp. 856-863.

[13] J. Friedman (1994) An overview of computational learning and function approximation, In: From Statistics to Neural Networks. Theory and Pattern

Recognition Applications. (Cherkassky, Friedman, Wechsler, eds.) Springer-Verlag 1

[14] M. Ester, H.-P. Kriegel, J. Sander and X. Xu (1996) A Density-based algorithm for discovering clusters in large spatial databases with noise, In Proceedings of the 2nd ACM International Conference on Knowledge Discovery and Data Mining (KDD), Portland, OR., pp. 226-231.

[15] G. Sheikholeslami, S. Chatterjee and A. Zhang "Wavecluster: A multi-resolution clustering approach for very large spatial databases," In Proceedings of the 24th VLDB Conference (1998).

[16] A. Hinneburg and D. A. Keim, "An efficient approach to clustering in large multimedia databases with noise," Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining, New York, pp. 58-65 (1998).