Clustering based Feature Selection from High Dimensional Data

Harshali D. Gangurde
Department of Computer Engineering
MET IOE-BKC, Savitribai Phule Pune University
Nasik, India
harshali.gang@gmail.com

Prof. K. V. Metre
Department of Computer Engineering
MET IOE-BKC, Savitribai Phule Pune University
Nasik, India
kvmetre@gmail.com

Abstract:- Data mining techniques have been widely applied to extract knowledge from large databases. Data mining searches for relationships and global patterns that exist in large databases that are 'hidden' among the huge data. Feature selection involves selecting the most useful features from the given data set and reduces dimensionality. Graph clustering method is used for feature selection. Features which are most relevant to the target class and independent of other are selected from the cluster.
The feature subset obtained is given to the various supervised learning algorithms to increase the learning accuracy and obtain best feature subset.
The feature selection can be efficient and effective using clustering approach. Based on the criteria of efficiency in terms of time complexity and effectiveness in terms of quality of data, useful features from the big data can be selected.

Keywords— *Feature selection,minimum spanning tree, clustering .*

*****

Introduction

Data mining is concerned with extraction of hidden predictive information from voluminous data. Different data mining functionalities are used for selecting the most relevant data from the big data set. Feature selection is a term that is used in data mining to determine the tools and techniques available for reducing data as input to a manageable size for processing and analysis. Clustering analysis is the task of grouping of objects (data) with similar features or attributes. Data processing is used to improve efficiency in mining process i.e.to extract data from the voluminous data with required features and removing irrelevant, redundant feature subset. Many feature subset selection methods are there like the Embedded, Wrapper, Filter and Hybrid to choose good subset of features.
Wrapper methods are feedback methods which incorporate the machine learning algorithm in the feature selection process, i. e. they depend on the performance of a specific classifier to evaluate the quality of a set of features. Wrapper methods search through the space of feature subsets using a learning algorithm to guide the search. A search algorithm is *wrapped* around the classification model to search for the space of different features. Filter methods are classifier agnostic, no-feedback, pre-selection methods that are independent of the machine learning algorithm. Just like wrapper methods, embedded approaches thus depend on a specific learning algorithm, but may be more efficient in several aspects. The hybrid methods are a mixture of filter and wrapper methods. It uses a filter method to reduce search space. The wrapper methods are computationally expensive and over fit on small training sets.

Feature subset selection is an approach of identifying subset of features that are mostly related to the target class. The main aim of feature selection is to remove irrelevant and redundant features which is also known as attribute or feature subset selection.
The purpose of feature selection from high dimensional data is to increase the level of accuracy and reduce dimensionality of the data.
Filter method is used in the clustering approach. A neighborhood graph of features or instances is computed. In the clustering approach minimum spanning tree (MST) based clustering algorithms is used. First, graph clustering method is used to divide features into clusters. After clustering, the most representative features which are strongly relevant to target classes are selected from each cluster to form the final feature subset. The clustering-based strategy gives best subset of features which are useful and independent.

I. Literature Survey

Feature selection identifies and eliminates as many irrelevant -and redundant features. Out of many available feature selection algorithms, some can eliminate irrelevant features but are incapable to handle redundant features [1] [2][3][4][5][6], but still some algorithms can eliminate the irrelevant while efficiently handling the redundant features [5][7][8][9]. Relief is an algorithm, which gives weight to each feature according to its ability to distinguish features under different targets based on distance-based criteria function [4]. Relief is not efficient in removing redundant features as two predictive highly correlated and weighted features [10]. Relief was proposed by Kira and Rendell in

ISSN: 2321-8169
3556 - 3560

1994. Relief is an easy to use, fast and accurate algorithm even with dependent features and noisy data. The algorithm is based on a simple principle. Relief works by measuring the ability of an attribute in separating similar instances. Relief-F (RFF) is an extension to relief algorithms which deals with multi-class problems and missing value. It is also improved to deal with noisy data and can be used for regression problems [3]. CFS,FCBF and CMIM are examples that take into consideration , along with irrelevant features, redundant features also affect the speed and accuracy of learning algorithms, and thus should be eliminated as well [7][9][10] [11][2].

CFS is achieved by the hypothesis that a good feature subset is one that contains features highly correlated with the target, yet uncorrelated with each other [7]. FCBF is a fast content based filter method which can identify relevant and redundant features, among relevant features without pair wise correlation analysis ([13] [14]). CMIM iteratively picks features which maximize their mutual information with the class to predict, conditionally to the response of any feature already picked [12]. Different from these algorithms, our proposed algorithm employs clustering based method to choose features. Recently, hierarchical clustering has been adopted in word selection in the context of text classification ([13], [15], and [16]). Distributional clustering clusters words into groups based on their grammatical relations with other words by Pereira et al. [13]. The distribution of class labels linked with each word by Baker and McCallum [15]. The Support Vector Machine (SVM) was originally designed for binary classification problems [18]. SVMs give good results for text categorization. The SVM is defined over a vector space where the classification problem is to find the decision surface that best separates the data points of one class from the other. In case of linearly separable data, the decision surface is a hyperplane that maximizes the margin between the two classes. Ensemble methods or classifier combination methods aggregate the predictions of multiple classifiers into a single learning model [19]. Several classifier models (called" or" learners) are trained and their results are usually combined through a voting or averaging process. The principal idea of bagging (short for bootstrap aggregating) is to aggregate predictions of several models of a given weak learner fitted to bootstrap samples of the original dataset by a majority vote. Some main advantages of bagging are its ability to reduce variance and to avoid model -over-fitting. It is an intuitive and easy to implement approach [20].

II. PROPOSED SYSTEM

The proposed system architecture processes the dataset in following stages:-

1. Clustering Algorithm: Clustering algorithm works in two steps as follows:-
• Removal of irrelevant features.
• Removing of redundant feature by constructing MST and selecting representatives from clusters relevant to target class
2. The representative feature selected from the clusters is given to the classifiers (supervised learners) for classification.
The classifiers used are: ANN, SVM. K-NN.
3. The Ensemble method Bagging applied aggregates the vote of the three classifiers to give the best feature subset.

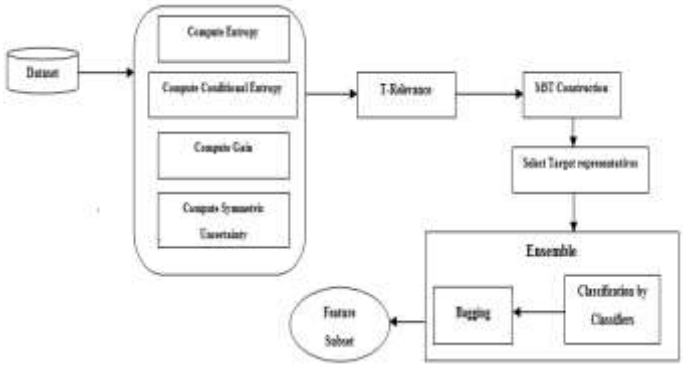

Fig. 1: System Architecture

III. IMPLEMENTATION DETAILS

A. *Symmetric Uncertainity*

Mutual information is considered to be a suitable criterion for feature selection. Symmetrical uncertainty measure is a normalization of mutual information. Symmetric Uncertainty has been used to determine the goodness of features for classification. The symmetric uncertainty is measured as follows:-

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

Here, H(X) and H(Y) are the entropies of discrete random variables X and Y.

Symmetric uncertainty treats a pair of variables symmetrically; it compensates for entropy partial towards variables with more values and normalizes its value to the range [0, 1]. A value 1 of *SU(X, Y)* indicates that knowledge of the value of either one completely predicts the value of the other and the value 0 implies that X and Y are independent.

B. *Entropy,Gain and Conditional Entropy*

H(X) i.e entropy is defined by p(x). Where p(x) is the prior probabilities for all values of X:-

-

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

Gain $(X|Y)$ reflects the information about Y provided by X. Gain is also called as information gain which is given by:-

$$Gain\ (X|Y) = H(X) - H\ (X|Y) = H(Y) - H(Y\ |X) \qquad (3)$$

H(X|Y) is the conditional entropy which reveals identity of random variable X given that the value random variable Y. Consider, 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:-

$$H\ (X|Y) = \Sigma_{y\epsilon Y}\ p(y)\ \Sigma_{x\epsilon Y}\ log_2\ p\ (x|y) \qquad (4)$$

Entropy 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. The order of two variables will not affect the value of entropy.

C. Defintions

1. Target-Relevance: The relevance between the features and the target concept C is referred to as the Target-Relevance of F_i and C, and denoted by $SU\ (F_i, C)$. If SU of any feature relevant to target class is greater than a predetermined threshold θ, we say that feature is a strong Target-Relevance feature.
2. Feature Correlation: The correlation between any pair of features is called Feature Correlation *and* denoted by $SU\ (F_i, F_j)$.
3. Feature Redundancy: Let $S=F_1, F_2 \ldots F_k$ $(K < F)$ be a cluster of features. If $\exists\ F_j \in S$, $SU(F_j,C) \geq SU(F_i,C) \wedge SU(F_i, F_j) > SU(F_i, C)$ is always corrected for each $F_i \in S (i \neq j)$ then F_i are redundant features with respect to the given F_j .

D. Algorithm

Input: Dataset Output:-Feature Subset.
1. Compute the Target-Relevance value for each feature. The features whose SU valueures with respect to target class are greater than a predefined threshold θ comprise the target-relevant feature subset.
2. Calculate the Feature-Correlation i.e $SU\ (F_i, F_j)$ value for each pair of features F_i and F_j The Feature Correlation $SU\ (F_i, F_j)$ is the weight of edges.
3. Build a minimum spanning tree(MST) using Prim's , then eliminate the edges whose weights are less than both of the Target-Relevance SU of F_i and F_j from the MST.
4. Check for redundant features in MST by the property that for each pair of nodes $SU\ (F_i, F_j)$, $SU\ (Fi, Fj) \geq SU\ (F_i,C) \wedge SU\ (F_i, F_j) \geq SU\ (F_i, F_j)$.
5. A forest is obtained after the removal of unwanted edges. Each tree $T \in$ Forest represents a cluster.
6. From each cluster we choose a representative feature whose Target-Relevance is the greatest.
7. The strong Target Relevance features selected from the clusters forms feature subset.
8. The feature subset obtained are given to the classifiers

IV. Experimental Setup & Result Analysis

The proposed system is tested on different high dimensional microarray and text dataset. The system works by taking input as one of the selected datasets. The selected dataset is processed under clustering algorithm the results obtained are the selected number of features or attributes relevant to target class. 15 different publicly available microarray and text datasets were used .

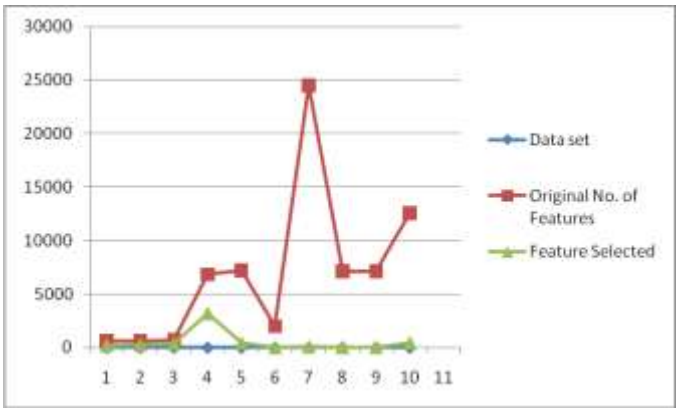

Fig2. Results: Feature Selected vs Original Features

The selected features are trained under different learning algorithms or classifiers to increase the learning accuracy of the selected high dimensional datasets. The training time needed by the classifiers is represented by graph for each classifier. The training time analysis includes training time for original high dimensional datasets and the selected feature subset from the algorithm are shown below in fig 3. and fig 4.

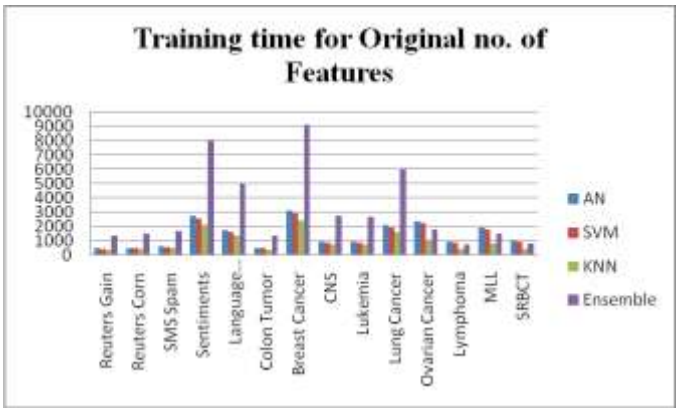

Fig 3. Training time for the Original no. of Features by the classifiers.

ISSN: 2321-8169
3556 - 3560

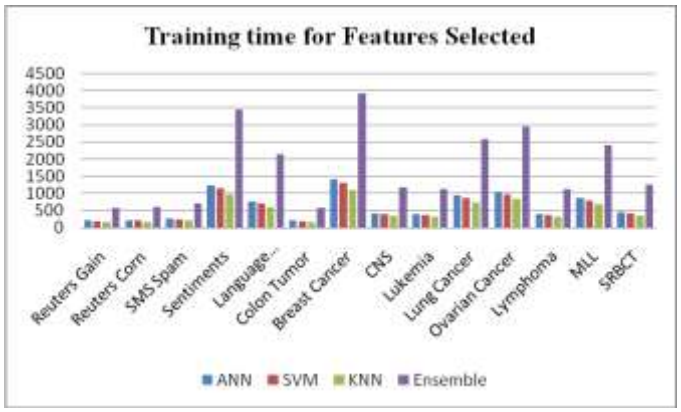

Fig 4. Training time for the selected feature subset by the classifiers. .

V. Conclusion

A novel clustering approach is proposed for feature selection from high dimensional data. The formation of clusters drastically reduces the dimensionality and helps in selection of relevant features for the concerned target class. The data pre processing removes the redundant and irrelevant features. The formation of clusters by constructing minimum spanning tree reduces the complexity for the computation of feature selection. The classification of features by the classifiers and the ensemble method gives better accuracy and provides best feature subset relevant to the target class. The benefit of feature selection is that the identity of the selected features can provide insights into the nature of the problem at hand. Therefore, the feature selection is an important step in efficient learning of large multi-featured data sets.

References

[1] Forman G., "An extensive empirical study of feature selection metrics for text classification", *Journal of Machine Learning Research, 3, pp 1289-1305, 2003"*..

[2] Hall M.A., "Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning" *Proceedings of 17th International Conference on Machine Learning, pp 359-366, 2000.*

[3] Kononenko I.,Estimating Attributes. "Analysis and Extensions of RELIEF", *Proceedings of the 1994 European Conference on Machine Learning, pp 171-182, 1994.,*

[4] Kira K. and Rendell L.A., "The feature selection problem: Traditional methods and a new algorithm", *Proceedings of Ninth National Conferenceon Arti_cial Intelligence, pp 129-134, 1992.*

[5] Modrzejewski M., "Feature selection using rough sets theory", *Proceedings of the European Conference on Machine Learning, pp 213-226, 1993.*

[6] Scherf M. and Brauer W., "FeatureSelection By Means of a Feature Weighting Approach" *Technical Report FKI-221 97, Institut fur Informatik, Technische Universitat Munchen, 1997.*

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

[8] Liu H. and Setiono R., "A Probabilistic Approach to Feature Selection: A Filter Solution" *Proceedings of the 13th International Conference on Machine Learning, pp 319-327, 1996.*

[9] Yu L. and Liu H.,.," Feature selection for high dimensional data: a fast correlation based filter solution" *Proceedings of 20 th International Conference on Machine Leaning, 20(2), pp 856-863, 2003.*

[10] Koller D. and Sahami M.., "Toward optimal feature selection" *Proceedings of International Conference on Machine Learning, pp 284-292, 1996*

[11] Kohavi R. and John G.H., "Wrappers for feature subset selection" *PArtif. Intell., 97(1-2), pp 273-324, 1999.*

[12] Fleuret F., "Fast binary feature selection with conditional mutual Information". *Journal of Machine Learning Research, 5, pp 1531-1555, 2004.*

[13] Pereira F.Tishby N. and Lee L., "Distributional c-lustering of English Words" *Proceedings of the 31st Annual Meeting on Association For Computational Linguistics, pp 183-190, 1993.*

[14] Dash M., Liu H. and Motoda H ., "Consistency based feature Selection" *Proceedings of the Fourth Paci_c Asia Conference on Knowledge Discovery and Data Mining, pp 98-109, 2000*

[15] Das S., "Filters, wrappers and a boosting-based hybrid for feature Selection" *Proceedings of the Eighteenth International Conference on Machine Learning, pp 74-81, 2001*

[16] Dash M. and Liu H., "Consistency-based search in feature selection" *Artificial Intelligence, 151(1-2), pp 155-176, 2003**

[17] Butterworth R., Piatetsky-Shapiro G. and Simovici D.A., "On Feature Selection through Clustering" *Proceedings of the Fifth IEEE internationalConference on Data Mining, pp 581-584, 2005.*

[18] Minh HoaiNguyen , FernandodelaTorre., " Optimal feature selection for support vector machines" *Elsevier,Pattern Recognition, pp 584591,2010*

[19] Alexey Tsymbal, Pdraig Cunningham, Mykola Pechenizkiy, Seppo Puuronen.,"Search Strategies for Ensemble Feature Selection in Medical Diagnostics" *Department of Computer Science, Trinity College Dublin, Ireland.*

[20] Yi-Wei Chen and Chih-Jen Lin., "Combining SVMs with Various Feature Selection Strategies", *Department of Computer Science, National Taiwan University, Taipei 106, Taiwan*

ISSN: 2321-8169
3556 - 3560