Relevance Feature Analysis for High Dimensional Data using Feature Subset Selection Algorithm

[1]K.Chinnaiyan, [2]D.Vidya Bharathi

[1]PG Scholar, Dept. of CSE

Sona College of Technology, Salem – 636005, TN, India

cinnaa10@gmail.com

[2]Asst. Professor, Dept. of CSE

Sona College of Technology, Salem – 636005, TN, India

dvbharathi77@gmail.com

Abstract— **The problem of estimating the quality of attributes (features subset) and reducing the attribute space of a feature subset is an important issue in the machine learning. For high dimensional data, effective machine learning and data mining techniques are not available. Reduction in the attribute space of a feature set can be done using two techniques such as, feature subset selection and dimensionality reduction. Feature selection involves identifying a subset of original features. The attribute space of a feature subset that is obtained from the feature selection can be reduced to an extent by dimensionality reduction. The comparison is made between the subsets of the original attributes constructed by fast clustering based feature selection algorithm (FAST) and constructed the linear combinations of subsets of the original attributes by principle component analysis (PCA) in terms of the classification performance and runtime performance. Reducing the size of the attribute sets is achieved and the changes in the classification results are investigated. Moreover, we have a tendency to explore the relationship between the variance captured within the linear combinations among PCA and also the ensuing classification accuracy. The results show that the classification accuracy based on PCA is extremely sensitive to the kind of information which the variance captured the principal components is not essentially a significant indicator for the classification performance.**

Keywords—**Feature subset selection, feature extraction, dimensionality reduction, relevance, redundancy, high dimensionality, correlation coefficient**

I. INTRODUCTION

In data mining application and machine learning, feature selection may not be effective for high dimensional data that leads to the so-called problem of "curse of dimensionality". Query accuracy and efficiency is being degraded rapidly as there is an increase in dimensionality. In the field of machine learning and data mining, the rapid increase of data dimensionality such as genomic micro-array data , text categorization and digital images leads to a major challenge for feature selection which is still an intractable problem. The data representation in many applications is represented by a huge number of features (attributes), and the raw data often contain many uninformative (irrelevant and redundant) ones which paves way for the degradation of the learning performance and compromise the quality of clustering. High dimension increases leads to lack of understanding the dataset itself and applying the algorithm, since many algorithms are sensitive to largeness or high-dimensionality or both [2].

Therefore, to minimize the occurrence of irrelevant and redundant ones and to retain the precious features various robust and effective feature selection algorithms have been introduced recently [3].There are three major benefits of feature selection (FS): (1) improves the prediction performance of the predictors; (2) helps predictors do faster and more cost-effective prediction; and (3) improves the understanding of the process that generated data.

Choosing a better subset of good features with respect to the target, the feature subset selection is an effective way of reducing dimensions, removing irrelevant features, improving learning accuracy, and increasing result comprehensibility [4]. Many of the feature subset selection methods proposed and studied for machine learning applications can be divided into four broad categories: the Embedded, Wrapper, Filter, and Hybrid approaches. In these approaches, filter methods are good suited for feature selection methods.

Many applications need to use unsupervised techniques where there is no previous knowledge about patterns inside samples and its grouping, so clustering can be useful. Clustering is grouping samples base on their similarity as samples in different groups should be dissimilar. Both similarity and dissimilarity need to be elucidated in clear way. High dimensionality is one of the major causes in data complexity. Technology makes it possible to automatically obtain a huge amount of measurements. However, they often do not precisely identify the relevance of the measured features to the specific phenomena of interest. Data observations with thousands of features or more are now common, such as profiles clustering in recommender systems, personality similarity, genomic data, financial data, web document data and sensor data. However, high-dimensional data poses different challenges for clustering algorithms that require specialized solutions. Recently, some researchers have given solutions on high-dimensional problem. Our main objective is proposing a framework to combine relational definition of clustering with dimension reduction method to overcome aforesaid difficulties and improving efficiency and accuracy in feature subset selection algorithm to apply in high dimensional datasets. Feature subset selection algorithm is applied to reduced datasets which is done by cluster based sufficient dimension reduction method. The conventional data is simple when compared to the current data which is more complex because of multidimensional and high dimension data. Less meaningful clusters are produced by many newly proposed clustering algorithms. The use of multidimensional data will result in more noise, complex data, and the possibility of unconnected data entities. Many emerging dimension reduction techniques proposed, such as Local Dimensionality Reduction (LDR) tries to find local correlations in the data, and performs dimensionality reduction on the locally correlated clusters of data individually [3], where dimension reduction as a dynamic process adaptively adjusted and integrated with the clustering process [4]. Sufficient Dimensionality Reduction (SDR) is an iterative algorithm [8], which converges to a local minimum of and hence solves the Max-Min problem as well. A number of optimizations can solve this minimization problem, and reduction algorithm based on Bayesian inductive cognitive model used to decide which dimensions are advantageous [11].

Unsupervised techniques where used in many applications because there is no prior knowledge about patterns inside the samples and its grouping in which clustering can be useful. Grouping of similar sample bases are clustered together and samples in other clusters are dissimilar. Both similarity and dissimilarity need to be elucidated in clear way. High dimensionality is one of the major causes in data complexity. Technology makes it possible to automatically obtain a huge amount of measurements. However, they often do not precisely identify the relevance of the measured features to the specific phenomena of interest. Data observations with thousands of features or more are now common, such as profiles clustering in recommender systems, personality similarity, genomic data, financial data, web document data and sensor data. However, high-dimensional data poses different challenges for clustering algorithms that require specialized solutions. Recently, some researchers have given solutions on high-dimensional problem. Our main objective is proposing a framework to combine relational definition of clustering with dimension reduction method to overcome aforesaid difficulties and improving efficiency and accuracy in feature subset selection algorithm to apply in high dimensional datasets. Feature subset selection algorithm is applied to reduced datasets which is done by cluster based sufficient dimension reduction method.

II. RELATED WORKS

The major issue in considering the feature subset that is the backbone of the process. Most of the feature subset may contain irrelevant or redundant content which literally degrades the performance of the selection process. To minimize this lack of accuracy in learning process we should identify and eliminate the content which is the cause for the lack of good performance. A better feature subset is termed to have a highly correlated content needed for the specified class and yet uncorrelated with other class. The process of feature subset selection method is to identify and remove the irrelevant and redundant features in the dataset, because the irrelevant feature does not provide the target concepts accurately, and the redundant features are not to be a formal one to predict a better concept.

Usually, the research of feature subset selection has mainly looking for relevant features. A familiar example is Relief. Relief is the one which calculate the weighs the feature according to the ability of the feature on distance based criteria function that discriminate instances under different targets, but it is ineffective at removing redundant features[1]. Relief-F is the extension of Relief this method is enabled to work with noisy and incomplete data sets and also to deal with multi-class problems, but still cannot identify redundant features. The redundant features are also like the irrelevant features that affect the speed and accuracy of the learning algorithms, so it should also be removed from [1] redundant features are taken into account in the examples such as FCBF and CMIM. A good feature subset method will achieve CFS [22] that contains features highly correlated with the target, yet uncorrelated with each other [1]. FCBF ([10], [11]) is a fast filter technique which is used to identify the relevant features as well as the redundancy among those relevant features without any pair wise correlation analysis. The features which maximize their mutual data with the class to predict were iteratively picked by CMIM, which is conditionally response to any features that are already picked [1]. The FAST algorithm that we proposed is used to choose the features based on clustering method that is different from the above mentioned algorithms.

Recently, hierarchical clustering has been used in the context of text classification for word selection. Distributional clustering is used to cluster words into groups, which is based either on their participation in grammatical relations with other words by Pereira et al. or on the distribution of class labels associated with each word [9]. This distributional clustering of words are complicate in nature, and result in high computational cost, thus Dhillon et al. proposed a new algorithm for word clustering and applied it to text classification which is based on information divisive method. Another man named Butterworth et al. [1] uses a special metric of Barthelemy-Montjardet distance to identify the specific features cluster, and the dendrogram of the ensuing cluster hierarchy to choose the most relevant features, but the Barthelemy-Montjardet distance method for identifying cluster does not identify a feature subset that encourages the cluster to enhance the accuracy. Even though when compared with other feature selection methods its accuracy is somewhat lower.

III. METHODOLOGIES

A. Framework and definitions

The best feature selection algorithm (Qinbao song 2013) which can be also used to cluster faster in current trend is the FAST algorithm. It works as two steps as follows. The first step using graph-theoretic clustering technique the features are formed into clusters. The second step is to form the final subset of relevant features from each cluster that is the selected feature is strongly correlated with the target. The selected features from each cluster are relatively independent. The clustering based approach of FAST has a high prospect of generate a subset of helpful and independent features.

We have a tendency to build up a new algorithm which might with efficiently and effectively manage each irrelevant and redundant feature, and acquire a subset of good features. We attain this through a new framework (shown in Fig.1) that collected the two associated components of removal the irrelevant feature and eliminate the redundant feature. The former attains features relevant to the target classes by removing irrelevant ones, and then removes the redundant features from relevant ones, and construct the final subset of features. The cluster based sufficient dimension reduction algorithm is applied to the final subset of features. This algorithm involves reducing the space of final subset of features (i.e.) high dimensional data into a lower dimensional subspace such as the variance retained is maximized; least square reconstruction error is minimized. Finally obtain the set of new attributes as combination of original features. The removal of irrelevant feature is simple once the accurate relevance feature measure is clear or selected, at the same time as the redundant feature elimination is a bit of complicated.

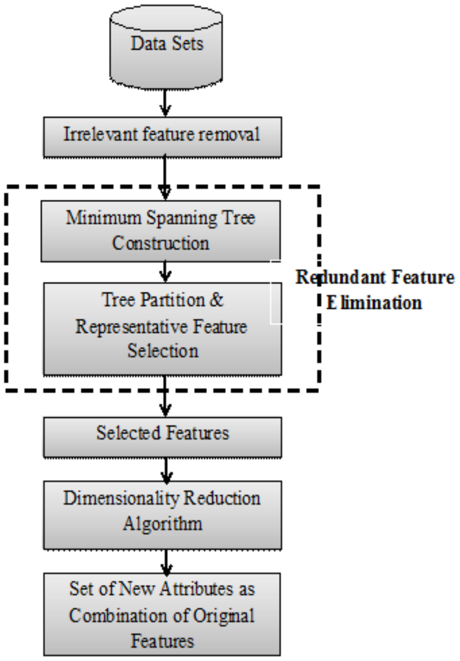
Data Sets
Irrelevant feature removal
Minimum Spanning Tree Construction
Tree Partition & Representative Feature Selection
Redundant Feature Elimination
Selected Features
Dimensionality Reduction Algorithm
Set of New Attributes as Combination of Original Features

Fig. 1: Framework of the feature selection and dimension reduction

Suppose F to be the full set of features, $Fi \in F$ be a feature, $Si = F - \{Fi\}$ and $S'i \subseteq Si$. Let $s'i$ be a value assignment of all features in $S'i$, fi a value-assignment of feature Fi, and c a value-assignment of the target concept C. The definition can be formalized as follows[1].

Definition 1: (Relevant feature) Fi is relevant to the target concept C if and only if there exists some $s'i$, fi and c, such that, for probability$p(S'i = s'i, Fi = fi) > 0$,$p(C = c \mid S'i = s'i, Fi = fi) \neq p(C = c \mid S'i = s'i)$.

Otherwise, feature Fi is an irrelevant feature.

Definition 2: (Markov blanket) given a feature (Yu L and Liu H 2004) $Fi \in F$, let $Mi \subset F$ $(Fi \not\in Mi)$, Mi is said to be a Markov blanket for Fi if and only if
$p(F - Mi - \{Fi\}, C \mid Fi, Mi) = p(F - Mi - \{Fi\}, C \mid Mi)$.

Definition 3: (Redundant feature) let S be a collection of features (Yu L and Liu H 2004), a feature in S is redundant if and providing it has a Markov Blanket among S. The relevant features have strong association with target conception necessary for a good subset of features, whereas redundant features aren't as a result of their values are completely associated with each other. Thus, ideas of feature relevance and feature redundancy are usually in terms of feature target idea correlation and feature correlation. The mutual information calculates what proportion the distribution of feature values and target concepts dissent from the statistical independence. It is a nonlinear evaluation of correlation between the feature values and target concepts. The Pearson correlation coefficient is used to generate rank weights for features by normalizing it to the correlation filter of features and target concepts, and it's used to calculate the goodness of each feature for classification by a number of researchers. Therefore, we choose Pearson correlation coefficient as the measure of correlation between either two features or a feature and the target concept.

B. *Correlation criteria*

Pearson correlation coefficient (r) (Qinbao song 2013) is used to measure the linear regression (dependence) between two variables X and Y. Based on a sample of paired data (X_i, Y_i), the sample Pearson correlation coefficient is defined as:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{s_X}\right)\left(\frac{Y_i-\bar{Y}}{s_Y}\right)$$

where

$$\frac{X_i-\bar{X}}{s_X},\ \bar{X} = \frac{1}{n}\sum_{i=1}^{n}X_i,\ \text{and } s_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2}$$

The first term represent the standard score; X_i and Yi are the number of values in the variable X and Y respectively, $\overline{X}$ represent the sample mean, and Sx represent the sample standard deviation, respectively.

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.

C. *Feature Extraction*

Feature extraction can be defined as follows: Given a set of features $S = \{v1, v2... vD\}$,find a new set of features S' derived from a linear or non-linear mapping of S. The cardinality of $|S'| = d$ and $J(S') \geq J(T)$ for all derived set of features T with $|T| = d$, where J is the evaluation function. Here d or some other parameter that can determine d (e.g., a threshold eigen value) is usually specified by the user. When all existing features are recombined to yield new features then we are dealing with feature extraction. Hence a mapping is defined that transforms any original D dimensional feature vector to a new d dimensional feature vector [17]. Ideally the mapping conserves or even enhances the discriminatory information while simultaneously reducing the dimensionality of the feature vector. Mapping can be linear or nonlinear. The following descriptions of a sample of classical feature extraction methods will bring out the methodical difference between feature transformations from selection.

Principal Component Analysis (PCA)

Principal component analysis (PCA) reduces the dimensionality of a data set by finding a new set of variables, smaller than the original set of variables. It retains most of the sample's information. The new variables, called principal components (PCs), are uncorrelated, and are ordered by the fraction of the total information each retains.

PCA is a dimension reduction technique that uses variance as a measure of interestingness and finds orthogonal vectors (principal components) in the feature space that accounts for the most variance in the data[16]. Principal component analysis is probably the oldest and best known of the techniques of multivariate analysis, first introduced by Pearson, and developed independently by Hotelling [13]. The advantages of PCA are identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. It is a powerful tool for analyzing data by finding these patterns in the data. Then compress them by dimensions reduction without much loss of information [15].

Apply PCA to reduce the dimension of the dataset

Step 1: Consider a dataset with X-dimensional Matrix.
Step 2: They are normalized by using Z-score.
Step 3: In the given matrix the single value decomposition data is calculated using X=UDVT
Step 4: Variance of the matrix is calculated using the diagonal elements.
Step 5: The obtained value is sorted in descending order.
Step 6: Largest variance represented as V helps in choosing the principle component p.
Step 7: Transformation matrix W is formed with the obtained p PCs.
Step 8: Finally a reduced projected dataset Y in a new co-ordinate axis.

IV. DATASET DESCRIPTION

Experiments are conducted on an EEG Eye state dataset which data is gathered from UCI repository web site. This web site is for finding suitable partners who are very similar from point of personality's view for a person. The dataset consists of 14 EEG values and each value indicating the eye state. All data is gathered from one continuous EEG measurement with the Emotive EEG Neuro headset. The measurement duration was 117 seconds. The eye state was detected during the EEG measurement via a camera and added later manually to the file and analysing the video frames. It indicates the state such as '1' the eye-closed and '0' the eye-open state. Data are organized in a table with 90 columns for attributes of people and 704 rows which are for samples. All attributes value in this table is ordinal and we arranged them with value from 1 to 15, therefore normalizing has not been done. All samples are included same number of attributes.

V. EXPERIMENTAL SETUP

In all experiments we use MATLAB software as a powerful tool to compute clusters and windows XP with Pentium 2.1 GHZ. Form the optimal subset from high dimensional datasets applied to FAST algorithm and reduced datasets done by principal component analysis.

A. *Experiment Results*

EEG Eye state original dataset is reduced using principal component analysis reduction method. Dataset consists of 14980 instances and 15 attributes. Here the Sum of Squared Error (SSE), representing distances between data points and their cluster centers have used to measure the clustering quality. The number of PCs obtained is same with the number of original variables. To eliminate the weaker components from this PC set we have calculated the corresponding variance, percentage of variance and cumulative variances in percentage, which is shown in Table 1. Then we have considered the PCs having variances less than the mean variance, ignoring the others. The variance in percentage is evaluated and the cumulative variance in percentage first value is same as percentage in variance, second value is summation of cumulative variance in percentage and variance in percentage. Similarly other value of cumulative variance is calculated.

	Variance	**Variance in percentage**	**Cumulative variance in percentage**
PC1	1.20173	4.0245	88.7532
PC2	0.41658	1.3869	93.9584
PC3	0.15248	0.5236	98.3335
PC4	0.02434	0.08236	99.825

Table 1.The Variances, Variances in Percentages, and Cumulative Variances in Percentages Corresponding to Pcs

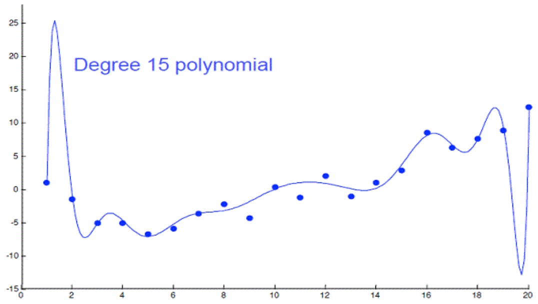

Fig1 Dataset performance graph

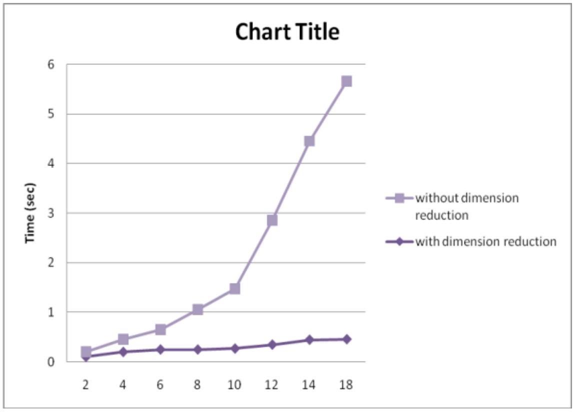

Fig 2 EEG Eye state dataset-Runtime (in ms)

Fig1 represents the EEG Eye state original dataset performance graph. Dataset consist of 14980 instances and 15 attributes. Fig2 represents the runtime (in ms) graph of EEG eye state original dataset without dimension reduction and with dimension reduction (PCA) respectively.

A. CONCLUSION

In this paper we have proposed the Relevance Feature Analysis for High Dimensional Data Using Feature Subset Selection Algorithm Dimensionality reduction is an efficient way of dealing data with high dimensionality. The purpose is to reduce the data so that computational load decreases and patterns of better quality can be extracted by pattern recognition and data mining algorithms. The estimating the quality of attributes (features subset) using the feature subset selection algorithm based on correlation filter and dimensionality reduction technique (PCA) is to reducing the attribute space of a feature set in the machine learning. This Feature subset selection and dimensionality reduction to obtain efficient processing time and performance of EEG eye state original dataset.

REFERENCES

[1] Qinbao Song, Jingjie Ni and Guangtao Wang, “A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data”,IEEE, 2013.

[2] Naijun Wu, Xiuyun Li, Jie Yang, Peng Liu, “Improved Clustering Approach based on Fuzzy Feature Selection”, IEEE, 2007.

[3] Shengyi Jiang, Lianxi Wang,” Unsupervised Feature Selection Based on Clustering”, IEEE, pp. 263, 2010.

[4] Liu H., Motoda H. and Yu L., Selective sampling approach to active feature selection, Artif. Intell., 159(1-2), pp 49-74 (2004).

[5] Mitchell T.M., Generalization as Search, Artificial Intelligence, 18(2), pp 203-226, 2000.

[6] Yu L. and Liu H., Efficiently handling feature redundancy in high dimensional data, in Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD ’03). ACM, New York, NY, USA, pp 685-690, 2003.

[7] Hall M.A. and Smith L.A., Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper, In Proceedings of the Twelfth international Florida Artificial intelligence Research Society Conference, pp 235-239, 1999.

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

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

[10] Yu L. and Liu H., Feature selection for high-dimensional data: a fast correlation-based filter solution, in Proceedings of 20th International Conference on Machine Leaning, 20(2), pp 856- 63, 2003.

[11] Yu L. and Liu H., Efficient feature selection via analysis of relevance and redundancy, Journal of Machine Learning Research, 10(5), pp 1205-1224, 2004.

[12] Dhillon I.S., Mallela S. and Kumar R., A divisive information theoretic feature clustering algorithm for text classification, J. Mach. Learn. Res.,3, pp 1265-1287, 2003.

[13] Chris Ding and Xiaofeng He, “K-Means Clustering via Principal Component Analysis”, In proceedings of the 21st International Conference on Machine Learning, Banff, Canada, 2004.

[14] Davy Michael and Luz Saturnine, 2007. Dimensionality reduction for active learning with nearest neighbor classifier in text categorization problems, Sixth International Conference on Machine Learning and Applications, pp. 292-297.

[15] Sembiring, Rahmat Widia, Jasni MohamadZain, Abdullah Embong: “Clustering High Dimensional Data Using Subspace And Projected Clustering Algorithm”, International Journal Of Computer Science &Information Technology (IJCSIT) Vol.2, No.4, pp.162-170 (2010).

[16] Sembiring, Rahmat Widia, Jasni Mohamad Zain, Abdullah Embong: “Alternative Model for Extracting Multidimensional Data Based- On Comparative Dimension Reduction”, ICSECS (2), pp. 28-42, (2011).

[17] Sembiring, Rahmat Widia, Jasni Mohamad Zain: “Cluster Evaluation Of Density Based Subspace Clustering”, Journal Of Computing, Volume 2, Issue 11, pp.14-19 (2010).

[18] Isabelle Guyon, Andr´e Elisseeff: “An Introduction to Variable and Feature Selection”, Journal of Machine Learning Research 3 (2003) 1157-1182.

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

[20] Hall M.A., Correlation-Based Feature Subset Selection for Machine Learning, Ph.D. dissertation Waikato, New Zealand: Univ. Waikato, 1999. Kononenko I., Estimating Attributes: Analysis and Extensions of RELIEF, In Proceedings of the 1994 European Conference on Machine Learning, pp 171-182, 1994.

[21] Kira K. and Rendell L.A., The feature selection problem: Traditional methods and a new algorithm, In Proceedings of Nineth National Conference on Artificial Intelligence, pp 129-134, 1992.

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