High Dimensional Data Classification through Attribute Reduction using RPCA Approach

Rajashree Dash[1], Rasmita Dash[2]

[1,2]Department of Computer Science and Engineering,
ITER, Siksha O Anusandhan University,
Bhubaneswar, Orissa, India
[1]rajashree_dash@yahoo.co.in, [2]rasmita02@yahoo.co.in

***Abstract*:** Classification is one of the most commonly encountered decision making tasks of human activity. A classification problem occurs when an object needs to be assigned to a predefined group or class based on a number of observed attributes related to that object. For many classification problems, a higher number of attributes used do not necessarily translate into higher classification accuracy. Hence attribute reduction can serve as a pre-processing tool of great importance before solving the classification problems. The main purpose is to reduce the maximum number of irrelevant features while maintaining acceptable classification accuracy. In this paper we proposed to use a hybridized RPCA approach of attribute reduction, which initially apply PCA to obtain reduced uncorrelated attributes specifying maximal variances in the data with minimum loss of information. Then we proposed to use Rough set theory on the PCA reduced data to discover discriminative features that will be the most adequate ones for classification. Lastly neural network has been applied for comparing the classification accuracy of some biological data sets with original attributes and reduced attributes.

***Keywords*:** Data Classification, Feature Reduction, Feature Selection, Principal Component Analysis, Rough Set Theory.

1. Introduction

The ever increasing demand for a knowledge based system has focussed much of attention of researchers on knowledge acquisition. The task of extracting general knowledge from databases is known to be the most difficult part of creating a knowledge-based system. Data mining is a convenient way of knowledge extraction from large data sets and focusing on issues relating to their feasibility, usefulness, effectiveness and scalability. Classification is one of the most frequently encountered data mining tasks, used to assign an object into a predefined group or class based on a number of observed attributes related to that object. For many classification problems, a higher number of attributes used do not necessarily translate into higher classification accuracy. In some cases the performance of algorithms devoted to speed and predictive accuracy of the data characterization can even decrease. Therefore, attribute reduction can serve as a pre-processing tool of great importance before solving the classification problems. The main purpose is to reduce the maximum number of irrelevant features while maintaining acceptable classification accuracy.

Attribute reduction is the transformation of high-dimensional data into a meaningful representation of reduced dimensionality that corresponds to the intrinsic dimensionality of the data [1], [2]. The intrinsic dimensionality of data is the minimum number of parameters needed to account for the observed properties of the data. Dimensionality reduction approaches fall into two categories i.e. Feature Selection (FS) and Feature Reduction (FR). Feature Selection algorithm aims at finding out a subset of the most representative features according to some objective function in discrete space. Feature Extraction/ Feature Reduction algorithms aim to extract features by projecting the original high-dimensional data into a lower-dimensional space through algebraic transformations. It finds the optimal solution of a problem in a continuous space.

In this research, an approach of high dimensional data classification using neural network through attribute reduction using RPCA method has been proposed, in which initially PCA has been applied to obtain reduced uncorrelated attributes specifying maximal variances in the data and then the Rough set theory has been applied to generate the reduced set of necessary attributes or to construct the core of the attribute set by finding the upper and lower approximation of the reduced data set. This is a combination of feature selection approach with feature reduction to obtain a minimal set attributes retaining a suitably high accuracy in representing the original features. This approach will produce a reduced set of attributes which specify the maximal variances in the data as well as the discriminative features most adequate for classification, with minimum loss of information. Lastly the classification accuracy of some biological data sets with original and reduced attributes has compared using neural network.

2. Related Work

The problem of finding a reduced set of relevant features retaining a suitably high accuracy in representing the original features has been the subject of much research. Feature Selection using Rough sets theory is a way to identify relevant features, which has been validated by the improvement on the performance of the KNN classifier [3]. The classification accuracy of the classifiers intended for use in high dimensional domains can be increased by applying Principal Component Analysis which also increases its computational efficiency [4]. In [5] a new face recognition method based on PCA, LDA and neural network has been proposed specifying a high recognition rate. Rough set has also used for feature selection in medical data bases like Mammograms, HIV etc. without decision attribute with the

application of clustering [6]. A novel method for dimensionality reduction of a feature set by choosing a subset of the original features that contains most of the essential information, using the same criteria as the ACO hybridized with Rough Set Theory has proposed in [8]. RST can only be applied on discretized data. A survey of discretization technique has been proposed in [13]. It has been validated that, the unsupervised methods like k-means clustering can perform equally well to that of supervised methods as it uses minimum square error partitioning to generate an arbitrary number k of partitions reflecting the original distribution of the partition attribute.

3. Preliminaries

3.1 Principal Component Analysis

Principal Component Analysis [11],[12] is an unsupervised Feature Reduction method for projecting high dimensional data into a new lower dimensional representation of the data that describes maximum variances in the data with minimum reconstruction error. It transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components. Hence PCA is a statistical technique for determining key variables in a high dimensional data set that explain the differences in the observations and can be used to simplify the analysis and visualization of high dimensional data set, without much loss of information.

PCs are calculated using the eigen value decomposition of a data covariance / correlation matrix or singular value decomposition matrix, usually after mean centering the data for each attribute. Covariance matrix is preferred when the variances of variables are very high compared to correlation. It would be better to choose the type correlation when the variables are of different types. Similarly the SVD method is used for numerical accuracy.

The transformation of the dataset to the new principal component axis produces the number of PCs equivalent to the no. of original variables. But for many datasets, the 1st several PCs explain the most of the variances, so the rest can be eliminated with minimal loss of information. The various criteria used to determine how many PCs should be retained for the interpretation are as follows:

- Using Scree Diagram plots the variances in percentage corresponding to the PCs which will automatically eliminate the PCs with very low variances.
- Fixing a threshold value of variance, so that PCs having variance more than the given threshold value will be retained rejecting others.
- Eliminate PCs whose eigen values are smaller than a fraction of the mean eigen value.

3.2 Rough Set Theory

Rough set theory is a new mathematical approach to imprecision, vagueness and uncertainty. It can be used for reduction of data sets, finding hidden data patterns and generation of decision rules. RST can be used as a tool to discover data dependencies and to reduce the no. of attributes contained in the data set using the data alone, requiring no additional information [9],[10]. Given a dataset with discretized attributes, it is possible to find a reduct of original attributes that are most predictive of the class attribute. Rough set reducts can be found by using degree of dependency or using discernibility matrix. In [7] a detailed concept of Rough set theory and decision tables for data analysis has provided.

4. Proposed Model

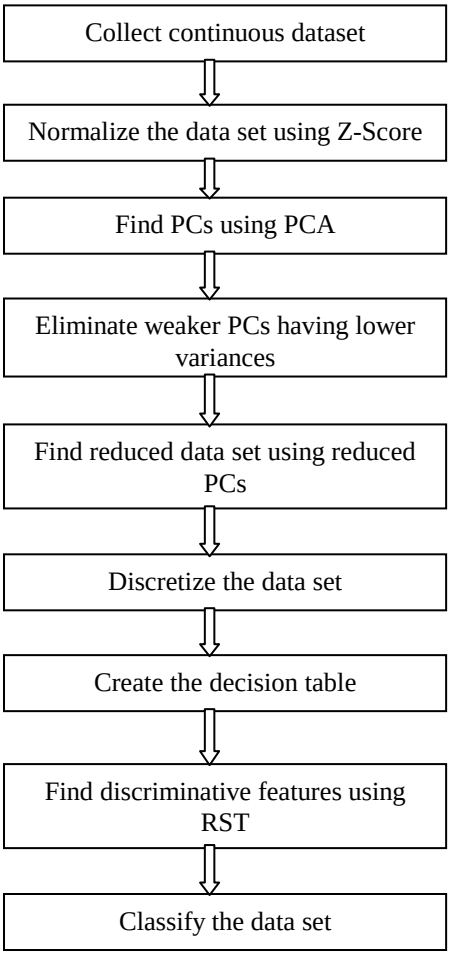

Figure1. Data Classification through Attribute Reduction using RPCA Approach

Classification accuracy for high dimensional data may not be accurate most of the time due to noisy and outliers associated with original data. Also for some data the computational complexity increases rapidly as the dimension increases. Hence to improve the accuracy of classification, we proposed a method to apply PCA on original data set, so that the correlated variables exist in the original data set will be transformed to possibly uncorrelated variables, which are reduced in size, and then to apply the rough set theory on that reduced data set, which may contain some redundancy and to get the discriminative features. Before applying RST, we proposed to discretize the data set using a suitable unsupervised discretization technique. Lastly a suitable classification technique will be applied to the test data considering its reduced attributes to find its class value.

5. Experimental Analysis

The proposed method has been implemented on four biological data sets containing initially the continuous attributes i.e. Pima Indian Diabetes data set, Lung Cancer data set, Breast Cancer data set and SPECTF Heart data set, taken from the UCI machine learning repository using met lab. The data set details are given in the table 1.

Table 1. Data set Details

Data set	No. of Instances	No. of Attributes	No. of Class values
Pima Indian Diabetes	156	8	2
Breast Cancer	286	9	2
Spectf Heart	267	44	2
Lung Cancer	32	56	3

The experimental details are given in following steps.

Step 5.1 Data Normalization

Using the Normalization process, the initial data values are scaled so as to fall within a small specified range; so that any attribute having higher domain value will not dominate the attribute having lower domain value.

Step 5.2 Attribute reduction using RPCA

A set of PCs are calculated using Singular value decomposition of the normalized data. Then a transformation matrix is created containing the PCs having variances more than the mean variance, ignoring the other PCs and this transformation matrix is applied to the normalized data set to produce the new reduced projected dataset.

The reduced data set is discretized using an unsupervised discretization method. Here we preferred to use discretization using k-means clustering as it uses minimum square error partitioning to generate an arbitrary number k of partitions reflecting the original distribution of the partition attribute and also it can perform equally well to that of supervised methods.

To apply RST, first a decision table containing object ids, the discretized attributes and the decision attribute is created. The class attribute of the data set has been considered as the decision attribute. Rough Set methods for finding reduct of attributes mainly categorized into two distinct approaches: those that incorporate the degree of dependency measure (or extensions), and those that apply heuristic methods to generate discernibility matrices. Although it is guaranteed to discover all minimal subsets using discernibility matrix method, it is a costly operation. Again simplifying discernibility function for reduct is a NP-hard problem. Hence here it is preferred to use the dependency based approach. Using Rough Set Theory the reduction of attributes is achieved by comparing equivalence relations generated by sets of attributes. Attributes are removed so that the reduced set provides the same predictive capability of the decision feature as the original. Here we have first calculated the dependency of each attribute and then the best candidate has chosen. This process has continued till the dependency of the reduct equals the consistency of the data set. The reduced set of attributes obtained by applying RST on the discretized data set, with different discretization intervals has shown in the table 2.

Table 2. Reduced Attributes obtained through RPCA Approach

Data set	No. of original attributes	No. of reduced attributes obtained by PCA	No of reduced attributes obtained by RPCA with different discretization intervals K=2	K=3	K=4	K=5
Pima Indian Diabetes	8	3	3	3	3	3
Breast Cancer	9	3	2	2	2	2
Spectf Heart	44	9	8	5	5	5
Lung Cancer	56	17	7	4	4	4

Step 5.3 Data Classification

The biological data sets have been classified using a feed forward back propagation network with two layers. One third of the data set has used as test data and the remaining as training data. Classifying the data set with original attributes, the PCA reduced attributes and with the reduced attributes obtained through RPCA approach, the classification accuracy obtained has shown in the table 3. In all cases the classification accuracy obtained by the reduced data set through RPCA approach is more than the original data set.

Table 3. Comparison of Classification Accuracy of Datasets

Data Set	Classification accuracy with original attributes.	Classification accuracy with PCA reduced attributes.	Classification accuracy with reduced attributes obtained by RPCA model .
Pima Indian Diabetes	58	64	64
Breast Cancer	76	81	86
Spectf Heart	85	71	85
Lung Cancer	65	68	74

6. Conclusion

In this paper some biological datasets with large no. of attributes have been classified through attribute reduction by RPCA approach. The attribute reduction through RPCA approach is a suitable combination of feature selection method with the feature reduction method. It has been implemented on the continuous data set, by applying an efficient PCA method, an efficient unsupervised

discretization technique and a reduction algorithm of RST. As a result of which a no of uncorrelated and discriminative attributes, more adequate for classification has been obtained. These attributes also specifies the maximal variances among the dataset by retaining the original property of the data set. Again comparing the classification accuracy of the data sets using neural network, it was observed that the data classification through attribute reduction using RPCA approach provides more accuracy compared to the PCA reduced data and the original dataset, by retaining the original property of data set.

References

[1] Maaten L.J.P., Postma E.O., Herik H.J. van Den., "Dimensionality Reduction: A Comparative Review.," Tech. rep. University of Maastricht, 2007.

[2] Yan J., Zhang B., Liu N., Yan S., Cheng Q., Fan W., Yang Q., Xi W. and Chen Z., "Effective and Efficient Dimensionality Reduction for Large Scale and Streaming Data Preprocessing," IEEE Transactions on Knowledge and Data Engineering, Vol. 18, No. 3, pp. 320-333, 2006.

[3] Frida Coaquira and Edgar Acuna, "Applications of Rough Sets Theory in Data Preprocessing for Knowledge Discovery," Proceedings of the World Congress on Engineering and Computer Science, pp.707-712, 2007.

[4] Changjing Shang and Qiang Shen, "Aiding Classification of Gene Expression Data with Feature Selection: A Comparative Study," International Journal of Computational Intelligence Research, Vol. 1, No. 1, pp. 68-76, 2005.

[5] Sahoolizadeh A. H., Heidari B. Z. and Dehghani C. H, " A New Face Recognition Method using PCA, LDA and Neural Network," World Academy of Science, Engineering and Technology, Vol. 41, pp. 7-12, 2008.

[6] Thangavel k. and Pethalakshmi A, "Feature Selection for Medical Database Using Rough System," AIML Journal, Vol. 6, No. 1, pp. 11-17, 2006.

[7] Nasiri J. H. and Mashinichi M, "Rough Set and Data Analysis in Decision Tables," Journal of Uncertain Systems, Vol. 3, No. 3, pp. 232-240, 2009.

[8] Mishra Debahuti, Rath Amiya Kumar and Acharya Milu, "Rough ACO: A Hybridized Model for Feature Selection in Gene Expression Data," International Journal of Computer Communication and Technology, Vol. 1, No. 1, pp. 85-98, 2009.

[9] Srivastava D. K, "Data Classification: A Rough - SVM Approach," Contemporary Engineering Sciences, Vol. 3, No. 2, pp. 77- 86, 2010.

[10] Lee-Chuan Lin, Zhu Jing, Junzo Watada, Tomoko Kashima and Hiroaki Ishii, " A Rough Set Approach to Classification and its Application for the Creative City Development," International Journal of Innovative Computing, Information and Control, Vol. 5, No. 12, pp. 4859-4866, 2009.

[11] Sampath Deegalla and Henrik Bostro, "Classification of Microarrays with kNN: Comparison of Dimensionality Reduction Methods," Proceedings of the 8th International Conference on Intelligent data Engineering and Automated Learning, pp. 800-809, 2007.

[12] Valarmathie P., Srinath M. and Dinakaran K, "An Increased Performance of Clustering High Dimensional Data Through Dimensionality Reduction Technique," Journal of Theoretical and Applied Information Technology, Vol. 13, pp. 271-273, 2009.

[13] Sellappan Palaniappan and Tan Kim Hong, " Discretization of Continuous Valued Dimensions in OLAP Data Cubes," International Journal of Computer Science and Network Security, Vol. 8, No. 11, pp. 116-126, 2008.

Author Biographies

Rajashree Dash She is an assistant professor in computer science and engineering department of ITER, SOA University, Bhubaneswar, India. She has completed her MTECH from SOA University. Her area of research is data mining, software engineering.

Rasmita Dash She is an assistant professor in computer science and engineering department of ITER, SOA University, Bhubaneswar, India. She has completed her MTECH from SOA University. Her area of research is data mining, software engineering.