Dimension Reduction in Microarray Dataset using Feature Selection

A.K.Shafreen Banu[1] Dr.S.Hari Ganesh [2]

[1]Dept. of IT, Bishop Heber College, Tiruchirapalli, 620017, India

[2] Dept. of CSE, H.H The Rajah's College (Autonomous), Pudukottai – 622 001, India

Abstract

In data mining dimension reduction is widely used in Medicine, Bioinformatics etc. Selection of features from huge dataset is tedious. Usually high dimension data contains noise , irrelevant information and small amount of relevant information. Reduction of dimensionality is very important to extract the important features, which is useful for predicting the results. This proposal layouts the high dimension data reduction using three ways.(i)Feature Selection (ii)Linear Dimensionality Reduction (iii)Non-Linear Dimensionality Reduction. In this work Feature Selection based on mutual information for feature filtering to select the relevant features with minimal redundancy. Linear Dimension Reduction is used in high dimension dataset for extracting the latent variables. The Non-linear dimension reduction is used to reduce the dimension for visalizing. Results are presented to show the efficiency of this work.

Keywords: Microarray Dataset, Dimension Reduction, Feature Selection.

I. INTRODUCTION

Dimension reduction has become important research area. The aim of dimension reduction is to visualize, understand and reduce the complex dataset structure. Gene expression microarray dataset usually has high dimension data. Analysis and visualization is tedious with this kind of dataset and it is critical in the field of medicine for diagnosis. This paper is adopted by two technologies namely feature selection and dimension reduction. Selecting the relevant features from the data set is feature selection, for feature selection and dimension reduction many algorithms are proposed like f-statistics, t-statistics etc.

Dimension reduction is classified into(i)Linear Dimension Reduction(LDR) namely PCA.(ii)Non-Linear Dimension Reduction(NLDR) namely MDS,ISOMAP[1] and local linear embedding[3].The above techniques are only effective for high dimensional data if it is used alone. Whereas feature selection is used to select the relevant features not reduce dimension.PCA also behave in same manner like linear method used for complex dataset of high dimension. Local Linear Embedding(LLE) is powerful and efficient for dimension reduction for NLDR.LLE is time consuming , requires more memory and complex computation which will be $O(dn^3)$, $O(dnk^3)$ and $O(rn^2)$.Many papers are published for dimension reduction and feature selection and found that no single algorithm works well only combination of algorithm produce effective result. The aim of this work is meant for dimensionality reduction .This approach involves dimension reduction for tumour microarray dataset. The purpose of this approach is to reduce the data attributes with minimum noise without losing original information. This approach starts with feature selection by using linear and followed by non-linear dimension reduction. Aim of feature selection is to remove useless and noisy data and gain only useful and beneficial information. In next step LDR with PCA to find maximum variance space. In order to keep more interesting data for final understanding of data it is necessary to follow non-linear dimensionality reduction. The remaining part of paper is divided into Section 2 with feature selection, Section 3 with Dimension reduction, Section 4 with Methods, Section 5 with experiments and Section 6 with conclusions and future enhancements.

II. FEATURE SELECTION

Microarray data has very high dimension dataset and has many number of features. Eventhough many features are there only small number of features are relevant for disease diagnosis [4][5].Microarray has large number of noisy redundant features, so the high dimension algorithm affects the result data accuracy and quality. hats why feature selection is applied to data to remove the irrelevant and redundant data. Feature selection facilitates many benefits like feature understanding, visualization, reducing time, dimension reduction, similar measures[14].Generally feature selection categorized into two approaches. The feature selection regardless of classifier known as feature ranking or filtering [15].Next feature selection approach regarding classifier for prediction, method called wrapper method which is useful to build good predictor[16].

III. DIMENSIONALITY REDUCTION

Dimensionality reduction provides understand and visualize the structure if high dimension and complex datasets. Here the proposed work is going to carry out two types of dimension reduction (i) Linear dimension reduction(PCA) (ii)Non-linear dimension reduction. LLE, ISOMAP, KPCA.

1. Linear Dimension Reduction:

One of the oldest methods for data analysis is PCA introduced by Pearson [2], which is used to extract the latent variables from high dimensional dataset. The method limits its effectiveness by global linearity and covariance matrix. To reduce the dimension PCA acts as first step converting high dimension into lower number of dimension and make the structure simplified.PCA finds the components with maximum variance.

2. Non-Linear Dimension Reduction:

Latent variables in non-linear approaches acts rich compared to linear method. So non-linear are powerful than linear method. Lee and Verleysen[8] describe the framework for NLDR. Distance measure is done with Euclidian Distance like MDS, ISOMAP.

- *Basic Approaches*
- *Identify the neighbours in i/p space.*
- *Develop a square matrix*
- *Compute embedding for matrix using eigenvector*

IV. FEATURE SELECTION, LINEAR DMENSIONAL REDUCTION AND NON-LINEAR DMENSIONAL REDUCTION

Feature Selection, Linear Dimensional Reduction and Nonlinear Dimensional Reduction Framework is presented in this paper for high dimensional data reduction, time computations and better visualization. This framework is composed of three steps as illustrated in Figure .a. As previously said, the initial step in this framework is feature selection.. Maximum Relevance Minimum Redundancy feature selection based mutual information has been used to rank the features and then select the top ranking ones which represent the most significant and correlated features. The next step is to apply LKPCA in order to find the maximum variance of the data and select the principal components with help of factor analysis technique. A non-linear dimension reduction algorithm LLE is finally used on the obtained data from the output of LKPCA. This algorithm is chosen because it is powerful in dimensional reduction.

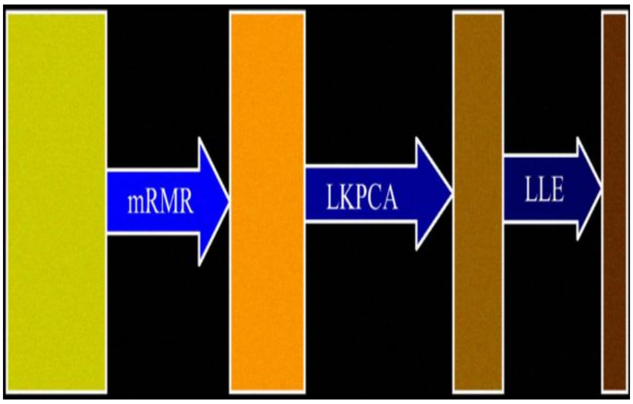

Figure.(a).Structure of the method

V. EXPERIMENTS

A Tumor dataset used here for demonstration. Totally it contains 255 features and observations of 72.These observations are classified into two where (1)means healthy and (0) means affected with disease. Among 255 features 198 is selected after applying feature selection. The below graphs represents the results of the data obtained from LKPCA, PCA and LLE

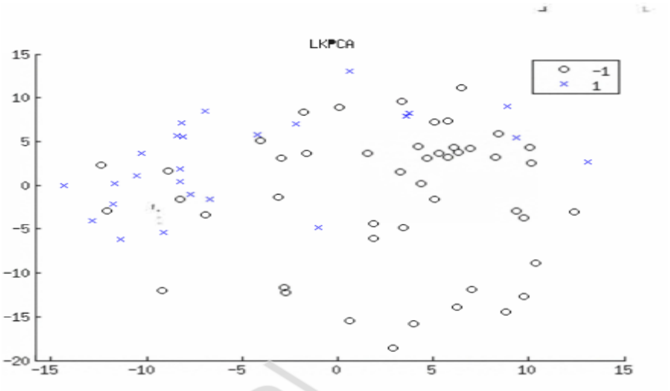

Figure. (b) Output of LKPCA algorithm

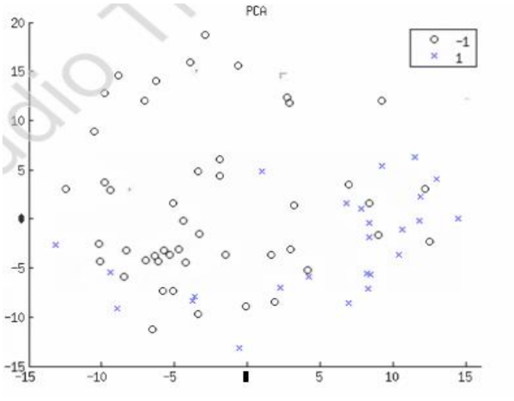

Figure. (c) Output of PCA algorithm

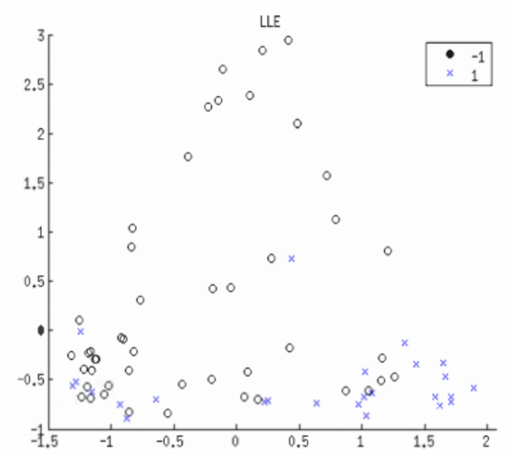

Figure. (d) Output of LLE algorithm

VI. CONCLUSION

In this work dimension reduction in microarray dataset using feature selection on LKPCA,LLE and mRMR are shown, the usage of feature selection is shown here. Both PCA and LKPCA provides same results according to time performance, whereas LLE with LKPCA provides better results provided good dimension reduction in least time. The output of this work helps us to see the condition of the patient.LLE operates on dataset without applying feature selection or after applying feature selection. In future it is able to apply in clinical tumor dataset for prediction. It is also planned that developing an algorithm for feature selection to handle new incoming changes in dataset.

REFERENCES

[1]. Tenenbaum, V. de Silva, and IC. Langford, A global geometric framework for nonlinear dimensionality reduction. Science. 290(5500):2319-2323,2009.

[2]. Pearson, K., On lines and planes of closest it to systems of points in space . Philosophical Magazine,2 :559-572,1 90 l.

[3]. Roweis S.T. and Saul L.K. Nonlinear dimensionality reduction locally linear embedding. Science. 290(5500):2323-2326, 2000.

[4]. Li W, Yang Y, How many genes are needed for a discriminate microarray data analysis?, in Critical Assessment of Techniques for Microarray Data Mining Workshop,pp. 137-150,2000.

[5]. Xiong M, Fang Z, Zhao J, Biomarker identiication by feature wrappers. Genome Res 11:1 878-1887,2 00l .

[6]. Wang JY, Almasri I, and Gao X. Adaptive graph regularized nonnegative matrix factorization via feature selection. The 21st International Conference on Pattern Recognition (ICPR2012). Tsukuba, Japan. November 2012.

[7]. Barrett T, Troup D, Wilhite S, Ledoux P, and Rudnev D. NCBI GEO: archive for high throughput functional genomic data. Nucleic Acids Research 2009;37:885.

[8]. Wang JY, Bensmail H, and Gao X. Joint learning and weighting of visual vocabulary for bag-of-feature based tissue classification.Pattern Recognition 2013 http://dx.doi.org/10.1016/j.patcog.2013.05.001.

[9]. Joshi D.M., Rana N.K., Misra V.M., "Classification of Brain Cancer Using Artificial Neural Network", International Conference on Electronic Computer Technology (ICECT), IEEE 2010.

[10]. Kohavi R. and John G. Wrappers for feature subset selection. Artificial Intelligence 1997; 97(1-2):273-324.

[11]. Li T, Zhang C, and Ogihara M. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics 2004;20(15):2429-2437.

[12]. Xing E, Jordan M, and Karp R. Feature selection for high dimensional genomic microarray data. In Proceedings of the 18th International Conference on Machine Learning. 2001. p. 601-608.

[13]. Watanabe A, Mabuchi T, Satoh E, Furuya K, Zhang L, Maeda S, and Naganuma H. Expression of syndecans, a heparan sulfate proteoglycan, in malignant gliomas: participation of nuclear factor-kappaB in upregulation of syndecan-1 expression. Neuro Oncology 2006;77:25-32 deskPDF

[14]. Guyon.! and Elisseef..A, An Introduction to Variable and Feature Selection. Jounal of Machine Leaning Research 3 (2003) 1157-1182,2002.

[15]. Langley P, Selection of relevant features in machine leaning, in AAAI Fall Symposium on Relevance, 1994

[16]. Arakeri M.P., Reddy G.R.M., "An intelligent content-based image retrieval system for clinical decision support in brain tumor diagnosis", International Journal of Multimedia and Information Retrieval, Springer 2013.

[17]. Shafaati M, Solomon A, Kivipelto M, Björkhem I, Leoni L. Levels of ApoE in cerebrospinal fluid are correlated with

Tau and 24S hydroxycholesterol in patients with cognitive disorders. Neurosci Lett 2007; 425:78-82.

[18]. Yang C, Sudderth J, Dang T, Bachoo R, McDonald J. DeBerardinis RJ Glioblastoma cells require glutamate dehydrogenise to survive impairments of glucose metabolism or Akt signaling. Cancer Research 2009;69:7986 7993.

[19]. A.K.Shafreen Banu,"A Study of Feature Selection Approaches for Classification",. IEEE Explore Digital Library: ISBN: 978-1-4799-6816-9, Vol. 2, pp. 223 - 228).

[20]. J. James Manoharan ,Dr. S. Hari Ganesh, M. Lovelin Ponn Felciah and A.K. Shafreen Banu, "Discovering Students' Academic Performance Based on GPA using K-Means Clustering Algorithm", Feb. 27 2014-March-12014,ISBN:978-1-4799-2876-7, IEEE World Congress on Computing and Communication Technology.