Efficient Retrieval Technique for Microarray Gene Expression

J Jacinth Salome [1], Dr. R.M Suresh [2]

[1.] Department of Computer Science, L.N Government College, Ponneri-601204, Tamil Nadu, India
[2.] Department of Computer Science and Engineering, Jerusalem Engineering College, Chennai-600100, India
jsalomej@gmail.com

Abstract: The DNA mciroarray gene data is in the expression levels of thousands of genes for a small amount of samples. From the microarray gene data, the process of extracting the required knowledge remains an open challenge. Acquiring knowledge is the intricacy in such types of gene data, though number of researches is arising in order to acquire information from these gene data. In order to retrieve the required information, gene classification is vital; however, the task is complex because of the data characteristics, high dimensionality and smaller sample size. Initially, the dimensionality diminution process is carried out in order to shrink the microarray data without losing information with the aid of LPP and PCA techniques and utilized for information retrieval. In this paper, we propose an effective gene retrieval technique based on LPP and PCA called LPCA. The technique like LPP and PCA is chosen for the dimensionality reduction for efficient retrieval of microarray gene data. An application of microarray gene data is included with classification by SVM. SVM is trained by the dimensionality reduced gene data for effective classification. A comparative study is made with these dimensionality reduction techniques.

[J Jacinth Salome, R.M Suresh. **Efficient Retrieval Technique For Microarray Gene Expression.** *Life Sci J* 2013;10(3):604-609] (ISSN:1097-8135). http://www.lifesciencesite.com. 89

Keywords: Information Retrieval, Gene classification, microarray gene expression, dimensionality reduction, Locality Preserving Projection (LPP), Principal Component Analysis (PCA), Support Vector Machine (SVM)

1. Introduction

Data mining plays a vital role in twenty-first century – the information age, (Ramamohanarao, 1989). They contribute in various fields of research such as 1) information sharing and collaboration, 2) security association mining, 3) classification and clustering, 4) intelligence text mining, 5) spatial and temporal crime pattern mining, 6) criminal/terrorist network analysis and more, (Chen, 2008). Recent advances in microarray technology have enabled the measurement of the simultaneous expression of thousands of genes under multiple experimental conditions. DNA microarray technology is also the field that exploits the data mining techniques. The molecular biologists face the challenges in discovering the essential knowledge from this kind of enormous volume of data, (Slavkov, 2005). In this type of knowledge seeking applications; information retrieval is one of the most crucial technologies (Lee, 2007) to mine the required information from the enormous amount of data. Microarray techniques have been successfully used to investigate useful information for cancer diagnosis at the gene expression level, the true integration of existing methods into day-to-day clinical practice is very challenging.

In the process of information retrieval in DNA microarray technology, gene classification is quite tough task, because of the characteristics of the data, which contain high dimensionality and small sample size (Leung, 2009; Alireza, 2009). While the DNA micro array technology considerably expedite the procedure of discovering the utility of genes like cancer classification. In the process of mining gene expressions under multi-conditions microarray experiments, gene clustering is another interesting task (Wai-Ho, 2005). The tools may be used for the identification of new tumor classes using gene expression profiles.

Microarray experiments normally produce a large amount of datasets with expression values for thousands of genes but still not more than a few dozens of samples, thus very exact arrangement of tissue samples in such high dimensional problems is a tricky task (Zhang, 2007). For the purpose of retrieving information from a microarray gene expression, we propose an effective retrieval technique based on LPP and PCA. Within this context, efficient retrieval emerges as a suitable paradigm specially intended for the development of biomedical informatics applications and decision support systems. As a first process in the proposed gene retrieval, the high dimensionality of the microarray gene data is reduced using dimensionality reduction technique (Changjing, 2005; Jian, 2006).

The LPP is chosen for the dimensionality reduction because of its ability of preserving locality of neighborhood relationship. The SVM is trained by the dimensionality reduced gene data for effective classification. SVM has the ability to learn with very few samples and so it is selected for the proposed technique. Hence, the classification is developed with the blending of dimensionality reduced technique and SVM results in effectual and powerful classification of gene expression data. Moreover, a comparative study is made with the LPP and PCA-based gene retrieval techniques (Gunanidhi, 2009).

2. The Proposed Gene Retrieval Technique LPCA

It is known that the microarray gene expression data is quite difficult because of the characteristics of the data, high dimensionality and small sample size (Satchidananda, 2008). We propose a technique LPCA for efficient gene retrieval for the application of different dimensionality reduction algorithms to experiment on gene expression data. This technique may improve the quality of the data analysis results, and may support the prediction of the number of relevant clusters in the microarray datasets. The techniques are described here. The proposed technique is comprised of two stages:

- Dimensionality reduction by two techniques LPP and PCA.
- Classification of gene data

Let, the microarray gene expression data is X_{jk}; $0 \le j \le n_g$, $0 \le k \le n_s$, where, n_g represents the number of genes from which the data is taken and n_s represents the number of samples. The gene data is of higher dimension and so it is subjected to dimensionality reduction. In the dimensionality reduction, the high dimensional gene data X_{ij} is converted to low dimensional data.

2.1 Dimensionality Reduction by LPP

Firstly, dimensionality reduction, of the proposed gene data is performed using LPP. From the gene data of different classes Y_{ijk} , a concatenated matrix is obtained as given in the Eq. (1). In the concatenated matrix, the gene data of all the classes are combined and it is given as a single matrix (Kitsana, 2004). The matrix Y_{conc} is given as:

$$Y_{conc\,jl} = \sum_{i=0}^{n_c - 1} Y'_{ijl} \qquad (1)$$

where,

$$Y'_{ijl} = \begin{cases} Y_{ijl} & ;\ if\ \ l \in (i, n_s(i+1) - 1) \\ 0 & ;\ \text{otherwise} \end{cases} \qquad (2)$$

The concatenated matrix Y_{conc} of dimension $n_g \times n'_s$; $n'_s = n_s . n_c , n'_s << cn_g$, which is highly dimensional and so the dimensionality of the matrix is reduced using LPP. The LPP is a linear dimensionality reduction algorithm that shares most of the properties of data representation of nonlinear techniques, namely, locally linear Embedding or Laplacian Eigenmaps(Chen, 2009). The LPP procedure for dimensionality reduction constitutes of three steps, namely, (1) generation of Distance matrix (2) determining adjacency matrix and (3) Calculating dimensionality reduced matrix.

2.1.1 Generation of distance matrix

For the concatenated matrix Y_{conc}, the distance matrix of size $n_g \times n_g$ is determined as follows:

$$D_{xy} = \sqrt{\sum_{l=0}^{n'_s} \left(Y_{conc\ xl} - Y_{conc\ yl}\right)^2}$$

$$0 \le x, y \le n_g \qquad (3)$$

The determined distance matrix is based on the Euclidean distance calculated by considering each row of the Y_{conc} as a network node. The resultant D_{xy} is subjected to calculate adjacency matrix, which can be determined based on the relationship of an element with every neighbor elements.

2.1.2 Determination of adjacency matrix

In the virtual network consisting of n_g nodes, the adjacency matrix is a $n_g \times n_g$ with binary entries representing if there is an edge between two nodes. Here, the adjacency matrix W is determined with the aid of the D_{xy} as follows:

$$W_{xy} = \begin{cases} 1 & ;\ if\ \ D_{xy} > 0 \\ 0 & ;\ \text{otherwise} \end{cases} \qquad (4)$$

From the Eq. (4), it can be seen that the adjacency matrix W_{xy} is constituted of binary values depending upon the distance calculated in D_{xy}.

2.1.3 Calculation of dimensionality reduced matrix

From the adjacency matrix W , a diagonal matrix A is determined as follows

$$A_{xy} = \begin{cases} S_x; \text{ if } x = y \\ 0 \text{ ; otherwise} \end{cases} \tag{5}$$

where,

$$S_x = \sum_{y=0}^{n_g - 1} W_{xy} \tag{6}$$

Based on the A, which is obtained from the Eq. (5), Z_1 and Z_2 are calculated as follows:

$$Z_1 = \frac{1}{2}\left(A_p + A_p^T\right) \tag{7}$$

$$Z_2 = \frac{1}{2}\left(L_p + L_p^T\right) \tag{8}$$

In Eq. (7) and (8), A_p and L_p can be determined by $A_p = Y_{conc}.A.Y_{conc}'$ and $L_p = A - W$, respectively. The obtained Z_1 and Z_2 are subjected to a generalized eigenvector problem (He, 2003) as follows:

$$Z_2 E = \lambda Z_1 E \tag{9}$$

$$\hat{Y} = E^T Y_{conc} \tag{10}$$

The $\hat{Y}$ obtained from the above equation is the dimensionality reduced gene data with size $n_s' \times n_s'$.

2.2 Dimensionality diminution through PCA

Let M_{xy} ; $0 \le x \le n_g$, $0 \le y \le n_s$ here n_g indicates the number of genes in the sample in which it has been taken from and n_s indicates the number of samples in which has been taken for the process. These microarray gene data is of higher dimension and hence it must be reduced in order to do that the PCA mechanism is utilized. In this dimensionality diminution, data which is in high dimension is converted to low dimension.

The dimensionality diminution is the process of reducing the large dimensional data, in order to make comfort to the classification process. PCA is one of the dimensionality reduction techniques that are utilized. PCA is a great tool for the data analysis process and also it can be utilized for the dimensionality reduction without any loss of information. The following steps are the dimensionality reduction steps of our microarray gene data.

Step 1: Compute the mean of the microarray gene data

$$\mu = \frac{1}{N_s * N_g} \sum_{i=1}^{Ns} \sum_{j=1}^{Ng} M_{ij} \tag{11}$$

Step 2: Subtract the mean from each gene data to find the mean deviation

$$\delta = M_{xy} - \mu \tag{12}$$

Step 3: Compute the covariance matrix for the data δ

$$Cov = \frac{\delta * \delta^T}{y - 1} \tag{13}$$

Step 4: Compute the Eigen values and Eigen vector and determine

$$\lambda = \mu * \mathrm{E}^T \tag{14}$$

Step 5: After computing the Eigen values and vector then the embedding process is as follows in

$$\hat{M} = \frac{\lambda}{\mathrm{E}^T} * \mu \tag{15}$$

3. Implementation of LPCA

In microarray data analysis, the process of information retrieval system includes diagnosis of disease, categorizing disease and getting information which is useful to give possible treatments (Zhang, 2007; Nevine, 2005). This makes the gene classification as one of the main tasks in microarray gene expression analysis (Leung, 2009) because it is a basis for prediction of the functions of unknown genes (Hori, 2001). The step in the analysis of gene expression data is the detection of samples or genes with similar expression patterns. The accurate classification of tumors is essential for a successful diagnosis and treatment of cancer (Mohammad, 2011). One of the problems associated with cancer tumor classification is the identification of unknown classes using gene expression profiles. Several classification algorithms have been developed for gene expression data (Jose, 2012).

Also techniques to systematically evaluate the quality of the clusters have been presented. The process of the SVM is to identify the class of the gene (Jacinth, 2011). From the training gene data, the SVM learns well about the class under which the given gene dataset is present. Once the SVM is trained well, it attains the ability to classify any gene dataset in the similar fashion. In the classification firstly, the gene dataset to be classified is subjected to dimensionality reduction using LPP or PCA. The dimension-reduced matrix is given to the trained SVM and so the class of the given microarray gene data is obtained in an effective manner.

4. Results and Discussion

In the issue of bridging the existing gap between biomedical researchers and clinicians who work in the domain of cancer diagnosis, prognosis and treatment, we have developed and made accessible an efficient retrieval technique. The proposed technique LPCA for microarray gene classification has been implemented in the working platform of MATLAB (version 7.11). For evaluating the proposed technique, we have utilized the microarray gene samples of human acute leukemia. The data has been taken by two different classes of microarray gene data, namely, acute myeloid leukemia (AML) and acute lymphoblast leukemia (ALL) (Sushmita, 2002). Thus obtained microarray gene expression data is of dimension, $n_c = 2$ $n_g = 7192$ and $n_s = 38$.

The samples were passed through diminution of dimension. First samples were passed through the technique LPP stage by stage from equation (1) to (10). $\hat{Y}$ was obtained in the dimensionality reduced gene data with size $n_s' \times n_s'$.

Secondly PCA technique was used for dimensionality reduction of the same set of microarray gene data. The step from 1 to 5 was carried out and $\hat{M}$ was obtained. Table I gives the dimensionality reduced data with the aid of LPP and PCA.

Finally the dimensionality reduced data set was used for classification. A sample of microarray gene dataset of two classes that has been used for processing is given in the Table II. Some six samples for each cancer class from broad institute database were acquired and they are given in the Table II and III for training and testing, respectively. In the testing, only the samples have been given and the proposed technique decides its belonging class.

Table 1. Microarray gene data dimension utilized for the evaluation process for no. of samples = 38 and no. of Genes = 7192

Type of Gene Data	Dimensionality Reduced Data with the aid of LPP	Dimensionality Reduced Data with the aid of PCA
ALL	27 X 38	29X40
AML	11 X 38	14X42

Table 2. A sample of microarray gene data corresponds to the cancer classes

Class	ALL		
Gene at (endogenous control)	19769_ B-cell	23953_ B-cell	28373 _B-cell
AFFX-BioB-5	-214A	-135A	-106A
AFFX-BioB-M	-153A	-114A	-125A
AFFX-BioB-3	-58A	265A	-76A
AFFX-BioC-5	88A	12A	168A
AFFX-BioC-3	-295A	-419A	-230A
AFFX-BioDn-5	-558A	-585A	-284A
AFFX-BioDn-3	199A	158A	4A
AFFX-CreX-5	-176A	-253A	-122A
AFFX-CreX-3	252A	49A	70A
AFFX-BioB-5	206A	31A	252A
Class	AML		
Gene at (endogenous control)	AML_ 12	AML_ 13	AML_ 14
AFFX-BioB-5	-20A	7A	-213A
AFFX-BioB-M	-207A	-100A	-253A
AFFX-BioB-3	-50A	-57A	136A
AFFX-BioC-5	101A	132A	319A
AFFX-BioC-3	-370A	-377A	-209A
AFFX-BioDn-5	-529A	-478A	-557A
AFFX-BioDn-3	14A	-351A	40A
AFFX-CreX-5	-365A	-290A	-243A
AFFX-CreX-3	153A	283A	119A
AFFX-BioB-5	29A	247A	-131A

In the aspect of efficiency, most algorithms aim to produce the best result based on the input parameters (Alireza, 2009). Performance comparisons are carried out with the pair LPP, SVM and PCA, SVM methods (Changjing, 2005). The efficacy of the techniques has been determined by comparing it with classification technique using Support Vector Machine (SVM).

Table 3. A sample of the microarray gene data corresponds used to test the proposed technique

Class	ALL		
Gene at(endogenous control)	19769 TA+ Norel	406 TA+ (ML) Norel	4466 Norel
AFFX-BioB-5	-214A	-342A	-87A
AFFX-BioB-M	-153A	-200A	-248A
AFFX-BioB-3	-58A	41A	262A
AFFX-BioC-5	88A	328A	295A
AFFX-BioC-3	-295A	-224A	-226A
AFFX-BioDn-5	-558A	-427A	-493A
AFFX-BioDn-3	199A	-656A	367A
AFFX-CreX-5	-176A	-292A	-452A
AFFX-CreX-3	252A	137A	194A
AFFX-BioB-5	206A	-144A	162A
Class	AML		
Gene at(endogenous control)	15 (PK) Norel	19 (PK) Norel	10 (PK) Relap
AFFX-BioB-5	-21A	-202A	-112A
AFFX-BioB-M	-13A	-274A	-185A
AFFX-BioB-3	8A	59A	24A
AFFX-BioC-5	38A	309A	170A
AFFX-BioC-3	-128A	-456A	-197A
AFFX-BioDn-5	-245A	-581A	-400A
AFFX-BioDn-3	409A	-159A	-215A
AFFX-CreX-5	-102A	-343A	-227A
AFFX-CreX-3	153A	283A	119A
AFFX-BioB-5	29A	247A	-131A

Table 4. Performance comparison for the classified data with the aid of dimensionality reduced data (LPP)

Type of Gene Data	Specificity (In %)	Sensitivity (In %)	Accuracy (In %)	Error Rate (In %)
ALL	92.62	86.26	97.30	2.703
AML	72.50	97.63	92.22	2.778

Table 5. Performance comparison for the classified data with the aid of dimensionality reduced data (PCA)

Type of Gene Data	Specificity (In %)	Sensitivity (In %)	Accuracy (In %)	Error Rate (In %)
ALL	92.59	54.35	82.12	13.596
AML	54.55	93.10	84.59	14.741

The comparison of the proposed technique with the LPP and PCA-based gene classification techniques with respect to the performance metrics, specificity, sensitivity, accuracy and error rate are given in the Table IV and Table V. The results show that this estimation approach may represent an effective tool to support biomedical knowledge discovery and healthcare applications. Thus the results state that in the proposed work when technique like LPP is chosen for the dimensionality reduction of microarray gene data and also for its effective retrieval. Our LPCA system implements tool that allows the use of combined techniques that can be applied to gene selection, clustering, knowledge extraction and prediction for aiding diagnosis in cancer research. For biomedical researches, LPCA offers a core workbench for designing and testing new techniques and experiments. For pathologists or oncologists, LPCA implements an effective and reliable system that can diagnose cancer subtypes based on the analysis of microarray data.

5. Conclusion

In this paper, we have proposed an effective dimensionality reduction technique between LPP and PCA. Here we utilize SVMs for classifying ALL and AML genes and for this process, initially these diseases equivalent gene data was trained with the SVMs separately. In the testing process any of these genes given and they have identified in which the class they belongs to. The proposed work has been obtained better results as they are compared with dimensionality reduction algorithm i.e. PCA. As the LPP have good positive features in their task of dimensionality reduction respectively. The comparative results have shown that the proposed technique LPP possesses better accuracy and lesser error rate than PCA techniques. LPP is the technique that can effectively support the integrative work of programmers, biomedical researches and clinicians working together in a common framework. Hence, this means of gene classification have paved the way for effective information retrieval in the microarray gene expression data. The results discussed here have shown the efficient performance of the proposed work.

Acknowledgements:

I express my sincere thanks to Supervisor Dr. R.M Suresh for his continuous motivation, innovative ideas and assistance throughout this paper. My heart filled gratitude to doctoral committee members Dr. Raja and Dr. Palanivel for their timely support.

Corresponding Author:
J. Jacinth Salome
Department of Computer Science
L.N Government College
Ponneri 601204, Tamil Nadu, India
E-mail: jsalomej@gmail.com

References

1. He. X and Niyogi. P (2003), Locality preserving projections in Adv in Neural Information Processing Systems, Cambridge, MA: MIT Press.
2. Jose Kaldas, Graphical Models for Biclustering and Information Retrieval in Gene Expression Data, Doctoral Dissertation, Aalto University Publication series, ISBN 978-952-60-4558-0, 2012.
3. Kitsana Waiyamai, Chidchanok Songsiri and Thanawin Rakthanmanon, Object-Oriented Database Mining: Use of Object Oriented Concepts for Improving Data Classification Technique, Lecture Notes in Computer Science, Vol: 3036, pp. 303-309, 2004.
4. Nevine M. Labib and Michael N. Malek, Data Mining for Cancer Management in Egypt Case Study: Childhood Acute Lymphoblastic Leukemia, World Academy of Science, Engineering and Technology, Vol. 8, 2005.
5. Alireza Osareh and Bita Shadgar, Classification and Diagnostic Prediction of Cancers using Gene Microarray Data Analysis, Journal of Applied Sciences, Vol. 9, No. 3, 459-468, 2009.
6. Dietmar Wolfram, Applications of Informetrics to Information Retrieval Research, Informing Science, Vol. 3, No. 2, 2000.
7. Gunanidhi Pradhan, Vishal Korimilli,Suresh Chandra Satapathy, Dr. Sabyasachi Pattnaik and Dr Bhabatosh Mitra, Design of Simple ANN (SANN) model for Data Classification and its performance Comparison with FLANN (Functional Link ANN), International Journal of Computer Science and Network Security, Vol.9 No.10, 105-115, 2009.
8. Mohammad Othman Nassar, Feras Al Mashagba and Eman Al Mashagba, Improving User Query for the Boolean model using Genetic Algorithm, International Journal of Computer Science Issues, Vol. 8, Issue 5, No. 1, 66-70, 2011
9. Jacinth Salome and Suresh, An Effective Classification Technique for Microarray Gene Expression by Blending of LPP and SVM, European Journal of Scientific Research, Vol.64, No.1, 34-43, 2011
10. Jian J. Dai, Linh Lieu, and David Rocke, Dimension reduction for classification with gene expression microarray data, Statistical Applications in Genetics and Molecular Biology, Vol. 5, No. 1, 1–21, 2006.
11. Ramamohanarao. K, Bailey. J, Fan (2005), Efficient Mining of Contrast Patterns and Their Applications to Classification, Journal of Intelligent Sensing and Information Processing, ISBN: 0-7803-9588-3, 39-47.
12. Roberto Ruiz, Jose C. Riquelme and Jesus S. Aguilar-Ruiz, Incremental wrapper-based gene selection from microarray data for cancer classification, Pattern Recognition, Vol. 39 , No. 12, 2383-2392, 2006.
13. Wai-Ho Au, Keith C. C. Chan, Andrew K. C. Wong and Yang Wang, Attribute clustering for grouping, selection, and classification of gene expression data, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 2, No. 2, 83-101, 2005.
14. Chen (2008), Homeland Security Data Mining Using Social Network Analysis, European Conference on Intelligence and Security Informatics, (pp. 4-4), Esbjerg, Denmark.
15. Hori, Inoue, Nishimura and Nakahara (2001), Blind Gene Classification – An Appln of a Signal Separation Method, International Conference on Genome Informatics volume12, (pp. 255 – 256), Tokyo.
16. Lee (2007), A Model for Information Retrieval Agent System Based on Keywords Distribution, IEEE International Conference on MUE, (pp. 413-418), 26-28 April, Seoul.
17. Mark A. Iwen, Willis Lang and Jignesh M. Patel, Scalable Rule-Based Gene Expression Data Classification, IEEE 24th International Conference on Data Engineering, (pp.1062-1071), 2008.
18. Satchidananda Dehuri and Sung-Bae Cho, Multi-objective Classification Rule mining Using Gene Expression Programming, Third International Conference on convergence and Hybrid Information Technology, Vol. 2, (pp. 754-760), 11-13 November, Busan, 2008.
19. Seungchan Kim, Younghee Tak and Luis Tari, Mining Gene Expression Profiles with Biological Prior Knowledge, IEEE Life Science Systems and Applications Workshop, July, (pp. 1-2), Bethesda, MD, 2006.
20. Sushmita Mitra, Sankar K. Pal and Pabitra Mitra, Data Mining in Soft Computing Framework: A Survey, IEEE Transactions On Neural Networks, Vol. 13, No. 1, 2002.
21. Zhang, Li and Hu (2007), A Hybrid Gene Selection Method for Cancer Classification, Workshop on Data Mining in Bioinformatics, Vienna, Austria.
22. Slavkov. I, Dzeroski. S, Struyf. J, Loskovska. S (2005). Constrained Clustering of Gene Expression Profiles, International Conference on data mining and data warehousing & Information Security, pp. 212-215, October 10-17, Slovenia
23. ALL/AML datasets from http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi.

4/16/2013