Microarray Classification Using Intelligent Techniques

Smitarani Satpathy
Research Scholars
UU, India
Email-satpathy.smitabbsr@yahoo.com

Pratikshya Mahapatra
Student,
OUAT, India
Email-pratikshya123mahapatra@hotmail.com

ABSTRACT

Microarray data classification is one of the most emerging clinical applications in the medical community. The classification process takes the detection of relevant and irrelevant probes into account, which is fundamental for subsequent classification.

In this thesis, an efficient technique is proposed for the precise classification of microarray genes from the microarray gene expression dataset. The proposed classification technique performs the classification process with the aid of three phases namely, feature extraction, dimensionality reduction, and gene classification. Initially, Principal Component Analysis (PCA) is applied for dimension reduction and significant features are extracted from high dimension microarray data. The original data is projected to the lower dimension for selecting the eigenvector for the co-variance matrix shown cumulative variance up to a label of 100%. After the implementation of PCA, the reduced feature matrix so obtained is divided into two sets such as, training and test data set. Feed Forward Neural Network is a train using Back Propagation Algorithm and the performance accuracy of the classifier is tested using the test dataset (untrained samples). The total number of epochs required to reach the degree of accuracy is referred as the convergence rate. To check the efficiency of the algorithm the convergence rate is considered by plotting graphs between numbers of epochs vs. means square error (MSE). The two keys controlling parameters of BP algorithm are learning rate and momentum which affects the accuracy as well as the convergence rate. Varying the values of the parameters various graphs are plotted.

Keywords- Microarray, PCA, Feature Extraction, Feature Selection, FFBNN, SVM

1. Introduction

Bioinformatics can be defined as the application of computer technology to the management of biological information, encompassing a study of the inherent genetic information, molecular structure, resulting biochemical functions and the exhibited phenotypic properties [1]. The data mining techniques are effectively used to extract meaningful relationships from these data. The application of data mining techniques in Bioinformatics is effectively used to extract meaningful relationships from these data. Biological data mining is an emerging field of research and development posing challenges and providing possibilities in this direction. The broad areas of Bioinformatics are Genomic sequence, protein structure, gene expression micro arrays and gene regulatory networks. The effectiveness of the selected gene subset is measured by its prediction accuracy or error rate in classification. In microarray experiments, classification of data is a crucial step for the prediction of phenotype of cells [5].

Methods from bioinformatics and computational biology are increasingly used to augment or leverage traditional laboratory and observation-based biology. These methods have become critical in biology due to recent changes in our ability and determination to acquire massive biological data sets, and due to the ubiquitous, successful biological insights that have come from the exploitation of those data [2]. Bioinformatics involve the creation and advancement of algorithms using techniques including computational intelligence, applied mathematics and statistics, informatics, and biochemistry to solve biological problems usually on the molecular level. Major research efforts in the field include sequence analysis, gene finding, genome annotation, protein structure alignment analysis and prediction, prediction of gene expression, protein-protein docking/interactions, and the modeling of evolution. Microarray technology is used to categorize the tissue samples by using their gene expression profiles as one of the several types (or subtypes) of cancer. The gene expression profiles measured by microarray technology have given an exact, consistent and objective cancer classification than the standard histopathological tests. Classification analysis of microarray gene expression data has been performed extensively to find out the biological features and to differentiate intimately related cell types that usually appear in the diagnosis of cancer.

1.1 What is the purpose of microarray classification?

Classification of patient samples is an important aspect of cancer diagnosis and treatment. Microarrays offer hope that cancer classification can be objective and highly accurate, providing clinicians with the information to choose the most appropriate forms of treatment. Microarrays may be used to assay gene expression within a single sample or to compare gene expression in two different cell types or tissue samples, such as in healthy and diseased tissue. Because a microarray can be used to examine the expression of hundreds or thousands of genes at once, it promises to revolutionize the way scientists examine gene expression. This technology is still considered to be in its infancy; therefore, many initial studies using microarrays have represented simple surveys of gene expression profiles in a variety of cell types. Each microarray experiment generates thousands of data points and reports are written in a dense technical jargon. It is easy to feel lost when trying to make sense of it all. For this reason, it is important to clearly define certain technical terms as well as goals of microarray experiments. To understand how microarrays are used, the jargon “class” and, more specifically, “known class”

must be first defined. A class refers to any characteristic shared by one group of samples but not other samples: e.g., cancer versus normal tissue, metastatic versus primary tumor, responders to cancer treatment versus non-responders. A "known class" is any differentiating characteristic that the researcher will use to label the tumor samples under study *a priori* the data analysis. The two main goals of microarray studies are: 1) to identify molecular signatures associated with known classes, and 2) to discover new classes. To achieve those goals, two different approaches to data analysis are taken, the supervised method (first goal above) and the unsupervised method (second goal). To read and understand microarray-based studies, knowledge of these different methods, will greatly help to understand the authors' hypothesis and data interpretation.

2. Dimensionality reduction Using Principal component Analysis

There are two main reasons to keep the dimensionality of the pattern representation (i.e., the number of features) as small as possible: measurement cost and classification accuracy. A limited yet salient feature set simplifies both the pattern representation and the classifiers that are built on the selected representation. Consequently, the resulting classifier will be faster and will use less memory. Moreover, as stated earlier, a small number of features can alleviate the curse of dimensionality when the number of training samples is limited. On the other hand, a reduction in the number of features may lead to a loss in the discrimination power and thereby lower the accuracy of the resulting recognition system. The term feature selection refers to algorithms that select the (hopefully) best subset of the input feature set. Methods that create new features based on transformations or combinations of the original feature set are called feature extraction algorithms. Note that often feature extraction precedes feature selection; first, features are extracted from the sensed data (e.g., using principal component or discriminate analysis) and then some of the extracted features with low discrimination ability are discarded. The choice between feature selection and feature extraction depends on the application domain and the specific training data which is available. Feature extraction and dimension reduction can be combined in one step using Principal Component Analysis (PCA), Genetic Algorithm (GA), and Kernel-PCA.

The purpose of this study is to evaluate some of those recently proposed for tumor classification with gene expression data. In the feature selection phase, the features are selected from the dimensionality reduced microarray cancer gene dataset. After that, the selected features are given to the feed forward back propagation neural network (FFBNN) to perform the gene cancer classification process. The proposed tumor data classification process is explained in Let W_{ij}; $1 \leq i \leq S$, $1 \leq j \leq G$ be the microarray tumor data, where S represent the number of samples and G represents the number of genes.

The dimensionality reduction process is used on the microarray dataset for reducing the complexity in the gene classification. Because the dataset size is very high dimensional, which increases the procession time and also will not produce the accurate result for the classification. The high dimensional W_{ij} dataset is converted into low dimensional dataset by selecting the optimal number of genes. For optimal gene selection process, we are using PCA.

We now briefly discuss some of the commonly used methods for feature extraction and feature selection

Feature selection is a process, through which no new set of features will be generated, but only a subset of original features is selected and feature space is reduced. The problem of feature selection is defined as follows: given a set of d features, select a subset of size m that leads to the smallest classification error.

The Taxonomy of dimensionality reduction techniques can be divided into two categories, transformation or selection based reduction. Problem of feature selection is hence, an important issue in cancer classification. It has been shown that, in many applications feature selection process improves a classifier's prediction capability.

2.1. Principal Component Analysis (PCA)

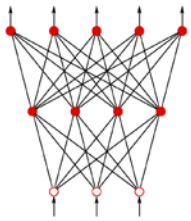
The Principal Component Analysis (PCA) was introduced by Karl Pearson in 1901. Principal component analysis (PCA) is the best, in the mean-square error sense, linear dimension reduction technique. Being based on the covariance matrix of the variables, it is a second-order method. PCA consists into an orthogonal transformation to convert samples belonging to correlated variables into samples of linearly uncorrelated features. The new features are called *principal components* and they are less or equal to the initial variables. If data are normally distributed, then the principal components are independent.

PCA mathematically transforms data by referring them to a different coordinate system in order to obtain on the first coordinate the first greatest variance and so on for the other coordinates. A number of correlated variables into a smaller number of uncorrelated variables called principal components. The algorithm solves for the eigenvalues and eigenvectors of a square symmetric matrix with sums of squares and cross products. The eigenvector associated with the largest eigenvalue has the same direction as the first principal component. The eigenvector associated with the second largest eigenvalue determines the direction of the second principal component. The sum of the eigenvalues equals the trace of the square matrix and the maximum number of eigenvectors equals the number of rows of the matrix.

3. Classifier design- using FFBNN

Classification operation performs the intelligent discrimination by means of features obtained from feature extraction phase. In this study FFBPNN is used. The feed forward neural network can be used for nonlinear transformation (mapping) of a multidimensional input variable into another multidimensional variable in the output. Presently, there is no satisfactory method to define how many neurons should be used in hidden layers.

A feed forward neural network is a biologically inspired classification algorithm. It consists of a (possibly large) number of simple neuron-like processing *units*, organized in *layers*. Every unit in a layer is connected with all the units in the previous layer. These connections are not all equal; each connection may have a different strength or *weight*. The weights on these connections encode the knowledge of a network. Often the units in a neural network are also called *nodes*.



Data enters at the inputs and passes through the network, layer by layer, until it arrives at the outputs. During normal operation, that is when it acts as a classifier, there is no feedback between layers. This is why they are called *feed forward.*

This fig. is an example of a 2-layered network with, from top to bottom: an output layer with 5 units, a *hidden* layer with 4units, respectively. The network has 3 input units. A multi-layer neural network can compute a continuous output instead of a step function. A common choice is the so-called logistic function:

$$y = \frac{1}{1 + e^{-x}}$$

In many classification problems, the classifier is expected to have some desired invariant properties. An example is the shift invariance of characters in character recognition; a change in a character's location should not affect its classification. If the preprocessing or the representation scheme does not normalize the input pattern for this invariance, then the same character maybe represented at multiple positions in the feature space. These positions define a one-dimensional subspace. As more invariants are considered, the dimensionality of this subspace correspondingly increases. The second main concept used for designing pattern classifiers is based on the probabilistic approach. The optimal Bayes decision rule (with the 0/1 loss function) assigns a pattern to the class with the maximum posterior probability. This rule can be modified to take into account costs associated with different types of misclassifications. For known class conditional densities, the Bayes decision rule gives the optimum classifier, in the sense that, for given prior probabilities, loss function and class-conditional densities, no other decision rule will have a lower risk (i.e., expected value of the loss function, for example, probability of error). If the prior class probabilities are equal and a 0/1 loss function is adopted, the Bayes decision rule and the maximum likelihood decision rule exactly coincide. In practice, the empirical Bayes decision rule, or "plug-in" rule, is used: the estimates of the densities are used in place of the true densities.

4. Experiment and Results

The datasets for this study is described in Table 1. They have been widely used for the benchmark problems. They consist of two binary cancer classification problems: Leukemia data set [7] and colon cancer dataset [12]. The initial leukemia data set consisted of 38 bone marrow samples obtained from adult acute leukemia patients at the time of diagnosis, of which 11 suffer

from acute myeloid leukemia (AML) and 27 suffer from acute lymphoblastic leukemia (ALL). An independent collection of 34 leukemia samples contained a broader range of samples: the specimens consisted of 24 bone marrow samples and 10 peripheral blood samples were derived from both adults and children. The number of input features was 7,129. The objective is to separate the AML samples from the ALL samples. The training set consisted of 38 patterns and 34 patterns were used for testing.

Table 1. Specification of the data sets

Data set	Training set	Testing set	Gene expression levels
Leukemia	38	34	7129
Colon	40	22	2000

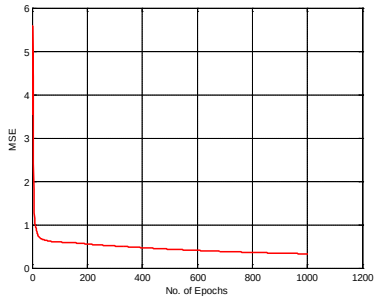
The colon cancer data set contains the expression of the 2000 genes with highest minimal intensity across the 62 tissues derived from 40 tumor and 22 normal colon tissue samples[12].The gene expression was analyzed with an Affymetrix (Sata Clara, CA U.S.A.) oligonucleotide array complementary to more than 6,500 human genes. The gene intensity has been derived from about 20 feature pairs that correspond to the gene on the DNA microarray chip by using a filtering process. Details for data collection methods and procedures are described in [12], and the data set is available from the website *http://microarray.princeton.edu/oncology/*.

We use the iris data as our bench mark input data which gives the 87.6% accuracy result. We use leukemia and colon tumor data for the experiment to calculate the learning parameter as well as the alpha, ita, means square error, cumulative variance 0.80 and 0.85, and evaluated classification models based on information gene pairs whose correlation coefficients are higher than *h*. Here we only report the experimental results *h* is 0.75

because the experimental results with different *h* values are similar.

Leukemia data has 72 samples and 7129 features. It is also a 2 class problem. Out of these samples we take 38 as training set and 34 as test set data.

Name	No. of samples	FFBNN
Colon Tumor	40, 22	93%, 91.33%
Leukemia	25, 47	95.85%, 95.5



Training Data

Figure 1

5. Conclusion

In this paper, we have selected the most widely used data sets in the literature for the evaluation of our algorithm. These data sets were obtained from Keng Ridge Bio-Medical (http://datam.i2r.a-star.edu.sg/datasets/krbd/) databases. The back propagation algorithm for training multilayer artificial neural network is studied and successfully implemented on FFBNN. This can help in achieving online training of neural networks on FFBNN than training in computer system and make a trainable neural network. In terms of hardware efficiency, the design can be optimized to reduce the number of multipliers. Though PCA takes the traditional features for classification, this work can further be extended by using GA for feature extraction and PNN, SVM for Classification. Because research said that there may be some more good features are left, but they can be given us good classification and accuracy result.

References

[1] Sushmita Mitra, Senior Member, IEEE, and Yoichi Hayashi, Senior Member, IEEE, 'Bioinformatics With Soft Computing', IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, VOL. 36, NO. 5, SEPTEMBER 2006.

[2] T.G. Smolinski et al. (Eds.): Comp. Intel. in Biomed. & Bioinform., SCI 151, pp. 3–47, 2008.springerlink.com.

[3] Juan F. De Paz1, Javier Bajo2, Sara Rodríguez1, and Juan M. Corchado1, 'Computational Intelligence Techniques for Classification in Microarray Analysis', 4, SCI 309, pp. 289–312.springerlink.com © 2010.

[4] Slonim DK (2002) From patterns to pathways: gene expression data analysis comes of age. Nat Genet 32: 502–508.

[5] Onur Dagliyan1, Fadime Uney-Yuksektepe2, I. Halil Kavakli1, Metin Turkay3, Optimization Based Tumor Classification from Microarray Gene Expression Data, February 2011 | Volume 6 | Issue 2 | e14579.

[6] J.H.Holland, Adaptation in Natural and Artificial Systems, MIT Press, Cambridge, Mass, USA, 1992.

[7] Veerabhadrappa, Lalitha Rangarajan,"Bi-level dimensionality reduction methods using feature selection and feature extraction**"** *in International Journal of Computer Applications (0975 – 8887) Volume 4 – No.2, July 2010.*

[8] S. Kumar, Neural Networks, McGraw-Hill, New York, 2005

[9] Ahmad M. Sarhan, "Cancer Classification Based on Micro array Gene Expression Data Using DCT and ANN", Journal of Theoretical and Applied Information Technology, Vol. 6, No. 2, pp. 208-216, 2009

[10] Bharathi and Natarajan, "Cancer Classification of Bioinformatics data using ANOVA", International Journal of Computer Theory and Engineering, Vol. 2, No. 3, pp. 369-373, June 2010

[11] Anil K. Jain, Fellow, IEEE, Robert P.W. Duin, and Jianchang Mao, Senior Member, IEEE "Statistical Pattern Recognition: A Review", vol.22, No.1, 2000

[12] Alon U, Barkai N, Notterman DA, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA 1999; 96: 6745 -50

[13]] Bo Li, Chun-Hou Zheng, De-Shuang Huang, Lei Zhang and Kyungsook Han, "Gene expression data classification using locally linear discriminate embedding", Computers in Biology and Medicine, Vol. 40, pp. 802–810, 2010

[14] Xiaosheng Wang and Osamu Gotoh, "A Robust Gene Selection Method for Micro array-based Cancer Classification", Journal of Cancer Informatics, Vol. 9, pp. 15-30, 2010

[15] Chhanda Ray, "Cancer Identification and Gene Classification using DNA Micro array Gene Expression Patterns", International Journal of Computer Science Issues, Vol. 8, Issue 2, pp. 155-160, March 2011.

[16] S. B. Cho and J. Ryu, "Classifying gene expression data of cancer using classifier ensemble with mutually exclusive features," *Proc. IEEE*, vol. 90, no. 11, pp. 1744–1753, Nov. 2002.

[17] S. Bicciato,M. Pandin, G. Didon`e, andC.DiBello, "Pattern identification and classification in gene expression data using an autoassociative neural network model," *Biotechnol. Bioeng.*, vol. 81, pp. 594–606, 2003.

[18] K. Deb and A. Raji Reddy, "Reliable classification of two-class cancer data using evolutionary algorithms," *BioSystems*, vol. 72, pp. 111–129, 2003.