ENHANCE ACCURACY OF TUMOR CLASSIFICATION FROM GENE EXPRESSION OF MICROARRAY

NÂNG CAO ĐỘ CHÍNH XÁC PHÂN LOẠI UNG THƯ THÔNG QUA BIỂU HIỆN GEN TỪ CÁC THÍ NGHIỆM MICROARRAY

Do Van Dinh[1,*],
Tran Hoai Linh[2], **Dang Thuy Hang**[3]

DOI: https://doi.org/10.57001/huih5804.64

ABSTRACT

Gene expression microarray data is one of the most popular for dianosis of cancer. However, the microarray data have thousands of genes and very few samples, it is crucial to develop techniques to effectively gene selection for analysis. So, dimension reduction is an important issue for analysis, of which principle component analysis (PCA) is one of the frequently used methods, and in the previous works, the top several principle components are selected for modeling according to the descending order of eigenvalues. While in this paper, we argue that not all the first features are useful, but features should be selected form all the components by feature selection methods. We demonstrate a framework for selecting good feature subsets from all the principle components, leading to enhance classifier accuracy rates on the gene expression microarray data. As a case study, we have considered PCA for dimension reduction, decesion tree algorithms (DT) for feature selection, and then Multi Layer Perceptron network (MLP) for classification. Experimental results illustrate that our proposed framework is effective to enhance classification accuracy rates.

Keywords: *PCA; DT; MLP; feature selection; microarray; classification.*

TÓM TẮT

Dữ liệu biểu hiện gen từ các thí nghiệm microarray là một dữ liệu phổ biến cho chẩn đoán ung thư. Tuy nhiên, điểm đặc biệt của loại dữ liệu này là có rất ít mẫu trong khi số biểu hiện gen lại lên tới hàng nghìn mẫu nên rất khó để lựa chọn được các gen có hiệu quả cho việc phân tích. Do đó, giảm chiều dữ liệu là phương pháp cần thiết trước khi dữ liệu đưa vào phân tích và phân tích thành phần cơ bản (PCA) là phương pháp được sử dụng để giảm chiều dữ liệu đầu vào. Trong bài báo này, có thể nhận thấy không phải phải thành phần dữ liệu đầu tiên là các thành phần dữ liệu tốt nhất do đo cần phải sử dụng thêm phương pháp lựa chọn đặc tính sau khi giảm chiều dữ liệu để chọn ra các đặc tính tốt nhất cho việc phân loại. Vì vậy, chúng tôi đề xuất sử dụng PCA để giảm chiều dữ liệu sau đó dùng thuật toán cây quyết định (DT) để lựa chọn ra các đặc tính phù hợp nhất và mạng MLP để phân loại dữ liệu. Các kết quả đạt được cho thấy đề xuất của chúng tôi cho hiệu quả tốt.

Từ khóa: *PCA; DT; MLP; lựa chọn đặc tính; microarray; phân loại.*

[1]Sao Do University
[2]Hanoi University of Science and Technology
[3]Le Quy Don Technical University
[*]Email: dodinh75@gmail.com/ dinh.dv@saodo.edu.vn

Received: 25/08/2022
Revised: 15/10/2022
Accepted: 22/11/2022

1. INTRODUCTION

DNA microarray experiments are used to collect information from tissue and cell samples regarding gene expression differences for tumor diagnosis [1-3]. The microarray data have thousands of genes and very few samples, it is crucial to develop techniques to effectively gene selection for analysis. To overcome this problem, we can either select a small subset of interesting genes (gene selection) or construct K new components summarizing the original data as well as possible, with K components << Sample. There are some methods of input space dimension reduction such as correlation [4] and variable ranking [5]. Each method puts more emphasis on one aspect than another. In this paper, we propose using PCA for dimension reduction and then using decesion tree algorithms to search the space of eigenvectors with the goal of selecting a subset of eigenvectors encoding important information. This approach has the advantage of simple, general, and powerful.

2. METHOD

Our tumor classification system using supervised learning has three main step. The main difference from the traditional approach is that performs feature selection among the principle components extracted by feature extraction. Dimension reduction step consists of two parts, feature extraction and feature selection, here feature extraction is performed by principle components analysis, and feature selection is performed by decesion tree algorithms. They are explained in the following subsection.

2.1. Feature Extraction Using PCA

Principal components analysis (PCA) is a statistical technique for determining the key variables in a multidimensional data set that explain the differences in the observations, and can be used to simplify the analysis and visualization of multidimensional data sets [6].

Consider a data matrix:

$$\mathbf{X}=\left\{x_{ij}\right\} \in R^{nxp} \qquad (1)$$

with n - number of rows (i.e. the number of vectors), p - number of colums (i.e. the data dimensions). PCA is mathematically defined as an orthogonal linear transformation that transform the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on.

The original data change in different intervals, so we need to standardized values in the column of matrix X:

$$\hat{\mathbf{X}}=\left\{\hat{x}_{ij}\right\} \tag{2}$$

With

$$\hat{x}_{ij}=\frac{x_{ij}-g_i}{\sqrt{n}} \tag{3}$$

$$g_i=\frac{1}{n}\sum_{i=1}^{n}x_{ij} \tag{4}$$

where, is the average value of the jth column of X, given by (4).

Then, $\hat{\mathbf{X}}$ will be used to calculate the covariance matrix of the data set given as:

$$\mathbf{V}=\hat{\mathbf{X}}^{T}\cdot\hat{\mathbf{X}} \tag{5}$$

$\mathbf{V}\in R^{pxp}$ is a matrix of size pxp.

Next, PCA finds the eigenvalues and the eigenvectors and arranged them in descending order. Suppose p eigenvalues of V are $\lambda_1\geq\lambda_2\geq....\geq\lambda_p$ and p eigenvectors are $\mathbf{u}_1,\mathbf{u}_2,...,\mathbf{u}_p$. Then the axis of the new space is eigenvector u_i. If we use only some k first dimensions (where $k\in p$) then the new matrix created from the eigenvectors is:

$$\mathbf{U}=\left[\mathbf{u}_1\middle|\mathbf{u}_2\middle|...\middle|\mathbf{u}_k\right]\in\mathbf{R}^{pxk} \tag{6}$$

and the new co-ordinates are:

$$\mathbf{F}=\dot{\mathbf{X}}\cdot\mathbf{U} \tag{7}$$

2.2. Feature Selection Using Decesion Tree Algorithm

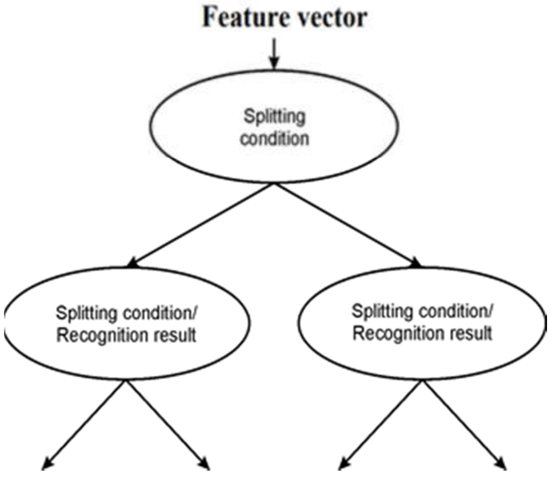
Feature vector
Splitting condition
Splitting condition/ Recognition result
Splitting condition/ Recognition result

Figure 1. An Example of Binary Decision Tree

The decision tree is a classical model for data recognition and classification [7, 8]. Among different model of decision trees, we will apply in this paper the linear model of binary tree. It means the tree will use only simple single conditions such as "if xi op A" at its nodes, where the op includes comparing operators such as =, >, <, >=, <=.

The general structure of the decision tree is given in Figure 1. There are a number of algorithms to train a tree for a given set of data samples. However, the algorithms to train a tree for a given set of data samples in this paper is ID3 [7, 8], which use the node entropy gain function to optimize the structure of the tree and the splitting conditions for each node of the tree. According to that, if at a node V we have N samples x_1, x_2, ..., x_N belonging to M classes C_1, C_2,..., C_M then the entropy of the node is given as:

$$E(V)=\sum_{i=1}^{M}-p_i\log_2(p_i) \tag{8}$$

where, $p_i=\frac{\overline{\left\{x_j : x_j\in C_i\right\}}}{N}$ is the probability that a sample x_j of the node belongs to the class C_i. Now with a splitting condition S, the samples from node V are classified to subnodes SV_i (for binary tree i = 1 or 2) with the appropriate numbers of samples are N_i ($\sum_i N_i=N$). At that time, the entropy gain for node V with splitting condition S is given as:

$$\mathrm{Gain}\left(V,S\right)=E\left(V\right)-\sum_i\frac{N_i}{N}E\left(SV_i\right) \tag{9}$$

A good splitting condition is the one with maximum value of entropy gain for a given node.

2.3. Multi Layer Perceptron network

MLP is a network of simple neurons called perceptrons [9, 10]. The basic concept of a single perceptron was introduced by Rosenblatt in 1958. The perceptron computes a single output from multiple real-valued inputs by forming a linear combination according to its input weights p and then possibly putting the output with i = 1,..., p, $x_i\in R^N; d_i\in R^K$ through some nonlinear activation function. Mathematically this can be written as:

$$Y=\psi\left(\sum_{i=1}^{n}W_ix_i+b\right)=\psi\left(w^Tx+b\right) \tag{10}$$

3. EXPERIMENTS

3.1. Data sets

In the paper, three real data sets obtained from different data sources. Due to the fact, the genetic data from microarray experiments have fewer sample numbers, while some variables (genes) more so to increase the reliability of solutions and research our research use the sample data of different diseases. Briefly described as below:

- Leukemia data sets were divided into two types of samples: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) with 47 patients with acute lymphoblastic leukemia (ALL) and 25 patients with acute myeloid leukemia (AML). Each of the 72 patients had bone marrow samples obtained at the time of diagnosis. Furthermore, the observations have been assayed with Affymetrix Hgu6800 chips, resulting in 5327 gene expressions (Affymetrix probes).

- Prostate data sets were divided into two types of samples: 52 patients with tumor and 50 patients with normal, resulting in 10509 gene expressions.

- Diffuse large B-cell lymphoma (DLBCL) data sets were divided into two types of samples: 58 DLBCL patients and 19 FL (Follicular lymphoma) patients were analyzed according to an Institutional Review Board approved protocol, resulting in 5469 gene expressions.

3.2. Experimental testing and result

To evaluate the performance of the proposed approach, we use the hold out validation procedure. Each data set is merged as a whole set, then we split the whole set into the training set and test set (2/3 for training data and the rest for test). The training data set is split by keeping 2/3 samples for training, the rest for validation. Classification error of MLPs is obtained on test data sets.

In order to demonstrate the importance of feature select ion of dimension reduction, we have performed four series experiments here:

(i) MLP has achieved satisfactory results, and here it is used without any feature reduction on the data sets;

(ii) PCA+MLP, PCA is a feature extraction method, it is used for dimension reduction without feature selection and the classification of SVM is used. The size of top eigenvectors of PCA is obtained by validating the classifier on the validation data set, as is a traditional way;

(iii) PCA+DT+MLP, beyond the baseline method, we proposed to use DT to select an optimum subset of eigenvectors, since we consider not all the top eigenvectors are useful for discrimination but the tail eigenvectors also contain useful information for discrimination.

Table 1. Show the number of features selected by each on three data sets

Data sets	MLP	PCA + MLP	PCA + DT + MLP
Leukemia	1 ÷ 300	6 (PCA1 ÷ PCA6)	4 (PCA1, PCA2, PCA4, PCA6)
Prostate	1 ÷ 500	6 (PCA1 ÷ PCA6)	4 (PCA8765, PCA6417, PCA3571, PCA897)
DLBCL	1 ÷ 500	8 (PCA1 ÷ PCA8)	4 (PCA1, PCA2, PCA6, PCA10)

From Table 1, we can find feature extraction using PCA + MLP and PCA + DT + MLP have less number inputs of three data sets than using only MLP.

The average error rates are shown in Table 2.

Table 2. Resul of classification

Data sets	MLP		PCA + MLP		PCA + DT + MLP	
	Learn error	Test error	Learn error	Test error	Learn error	Test error
Leukemia	3	1	1	1	1	0
Prostate	1	1	2	1	0	0
DLBCL	4	1	1	0	0	0

From Table 2, we can find feature selection by PCA and DT do great help in reducing features and get better result on classification.

3.3. Discussions

The difficulties of building a classifier for gene expression microarray data are dimension reduction. Here we use the PCA + DT + MLP framework to get a simpler, gender and efficiency classifier. Observing the tables shown in Section 3.2, several interesting comments can be made as below:

(i) The feature subsets selected by the DT approach improve classification performance, all for the different data sets.

(ii) The DT solutions are quite compact: The final feature subsets found by DT are very compact; the significant reduction in the number of eigenvectors speeds up classification substantially.

(iii) We can find DT both reduces the average error rate and the number of features selected.

4. CONCLUTION

We have investigated a systematic feature reduction framework by combing feature extraction with feature selection. To evaluate the proposed framework, we used three typical data sets. In each case, we used PCA for feature extraction, DT as feature selection, and MLP for classification. Our experimental results illustrate that the proposed method improves the performance on the gene expression microarray data in the accuracy. Further study of our experiment indicates that not all the top eigenvectors of PCA are useful for classification, the tail eigenvector also contain discriminative information. Therefore, it is necessary to combine feature selection with feature extraction for dimension reduction for analyzing high dimensional problems.

REFERENCES

[1]. T. Golub, D. Slonim, P. Tamayo, et al, 1999. *Molecular classification of cancer: Class discovery and class prediction by gene expression.* Bioinformatics & Computational Biology, 286 (1999, 531-537. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2]. U. Alon, et.al, 1999. *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays*. In Proceedings of the National Academy of Sciences of the United States of America, 1999, 6745–6750.

[3]. S. Dudoit, J. Fridlyand, T. Speed, 2002. *Comparison of discrimination methods for the classification of tumors using gene expression data*. Journal of the American Statistical Association, 97 (457) pp. 77-87.

[4]. Vinay Soni, Ritesh Joshi, 2012. *A Novel Dimension Reduction Technique based on Correlation Coefficient*. International Journal of Scientific & Technology Research Volume 1, Issue 4, May 2012.

[5]. Isabelle Guyon, Andre Elisseeff, 2003. *An introduction to variable and feature selection*. Journal of Machine Learning Research 3, pp. 1157-1182.

[6]. Jang B. Rampal, 2001. *DNA Array Method and Protocol*. Methods in Molecular Biology (MIMB), Vol. 170, pp. 229-230.

[7]. L. Monson, 1997. *Algorithm Alley Column: C4.5*. Dr. Dobbs Journal, Jan 1997.

[8]. J. Ross Quinlan, 1993. *C4.5 Programs for Machine Learning*. Morgan Kaufmann, 1993.

[9]. Tran Hoai Linh, 2014. *Neural network and its application in signal processing.* Publisher of Bach khoa.

[10]. Nguyen Quan Nhu, 2010. *Research and application of neural network and fuzzy logic to the problem of electricity load forecasting short-term.* Doctoral thesis, Hanoi University of Science and Technology

THÔNG TIN TÁC GIẢ

Đỗ Văn Đỉnh[1], Trần Hoài Linh[2], Đặng Thúy Hằng[3]

[1]Trường Đại học Sao Đỏ

[2]Trường Đại học Bách khoa Hà Nội

[3]Trường Đại học Kỹ thuật Lê Quý Đôn