A Data Mining Application to Leukemia Microarray Gene Expression Data Analysis

Yuh-Jye Lee[+] and Chia-Huang Chao

Abstract

Microarray technology makes biologists be capable of monitoring expression of thousands of genes in a single experiment on a small chip. How to extract knowledge and information from these microarray gene expression datasets has attracted great attention of many molecular biologist, statisticians and computer scientists. Recently, data mining has become a synonym for the process of extracting the hidden and useful information from datasets. In this paper, we developed a methodology to analyze the gene expression data by using the techniques in data mining such as feature selection and classification. We are given the leukemia microarry gene expression dataset that consists of 72 samples. Each sample has 7129 gene expression levels and comes with one of three class labels: AML, B-cell ALL and T-cell ALL. We selected 30 genes out of 7129 by using feature selection technique and prescribe a procedure to classify every sample into the correct category.

Keywords: microarray gene expression data, data mining, multicategory classification, clustering, support vector machine

Scopes: Data mining and Bioinformatics

[+] Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, No. 43, Sec. 4, Keelung Rd., Taipei, 106, Taiwan.

Email address: yuh-jye@mail.ntust.edu.tw.

A Data Mining Application to Leukemia Microarray Gene Expression Data Analysis

Yuh-Jye Lee and Chia-Huang Chao
yuh-jye@mail.ntust.edu.tw, M9115016@mail.ntust.edu.tw
Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology
No. 43, Sec. 4, Keelung Rd., Taipei, 106, Taiwan

Abstract

Microarray technology makes biologists be capable of monitoring expression of thousands of genes in a single experiment on a small chip. How to extract knowledge and information from these microarray gene expression datasets has attracted great attention of many molecular biologist, statisticians and computer scientists. Recently, data mining has become a synonym for the process of extracting the hidden and useful information from datasets. In this paper, we developed a methodology to analyze the gene expression data by using the techniques in data mining such as feature selection and classification. We are given the leukemia microarry gene expression dataset that consists of 72 samples. Each sample has 7129 gene expression levels and comes with one of three class labels: AML, B-cell ALL and T-cell ALL. 3571 genes were left after data preprocessing. We selected 30 genes out of 3571 genes by using feature selection technique and prescribe a procedure to classify every sample into the correct category.

Keywords: microarray gene expression data, data mining, multicategory classification, clustering, support vector machine

1. Introduction

Microarray technology makes biologists be capable of monitoring expression of thousands of genes in a single experiment on a small chip. Several microarray gene expression datasets are publicly available on the Internet [14, 15]. These datasets include a large number of gene expression values and need to have an accurate method to extract knowledge and useful information from these microarray gene expression datasets.

Support vector machines (SVMs) [12] have been shown have superior performance in the analysis of microarray gene expression data than other classification algorithms such as decision trees, Parzen windows and Fisher's linear discrimination [6, 8]. Hence we will use SVMs as the classification tool in this paper.

Here we use the acute leukemia dataset the same as Golub presented in [9] for our classification experiments and convert this multicategory classification problem into a series of binary classification problems. Besides, we propose a hierarchical model for selecting a small number of informative genes and classifying acute leukemia patients as acute myeloid leukemia (AML), B-cell acute lymphoblastic leukemia (B-cell ALL) or T-cell acute lymphoblastic leukemia (T-cell ALL) class based to those selected genes. Each level of this hierarchical model is composed of two phases, gene selection step and classifier construction step. Our model selects 30 genes out of 3571 genes and can classify 34 examples into three categories, which are defined above correctly based on the training result of 38 examples.

This paper is organized as follows. In section 2 we describe the leukemia gene expression dataset and provide the gene selection criterion in details. Section 3 gives a basic description of the essential of the smooth

support vector machine (SSVM) [13]. The classifiers used in this work are trained by the SSVM. The experimental results are shown in section 4. Section 5 includes the contributing of this method and proposing some possible extensions of this method in the future.

2. Gene selection

The acute leukemia dataset contains 38 training samples and 34 independent testing samples. All samples have 7129 features obtained from the microarray. Hence there is a total number of 72×7129 microarray gene expression data in this dataset. Each sample is labeled as AML, B-cell ALL or T-cell ALL. Therefore, the problem we faced is a multicategory classification problem. The class distribution of this dataset described as table 1.

Table 1: Class distribution of leukemia dataset

	Training	Testing	Total
AML	11	14	25
B-cell ALL	19	19	38
T-cell ALL	8	1	9
Total	38	34	72

Although there are more than 7000 features for each sample, many of them are uninformative or providing redundant information for class distinction. Before gene selection phase, we need to perform some preliminary data cleaning in order to discard those uninformative genes expression data and reduce the number of features, so as to decrease the computational time in following steps. In the preliminary processing stage, we eliminate those genes whose expression values do not have significant differences in different classes.

After this preliminary processing, we still retain 3571 features for each sample. For the sake of extracting the most informative genes for discriminating classes, we need a metric mechanism for gene selection with following property: The metric expressions of an informative gene in one class should be quite different from the other, but the variation in the same class is as little as possible [2]. We adopt a metric mechanism fit in with this property, which is presented in [9], called correlation metric $P(g_i)$.

$$P(g_i) = \frac{m_+ - m_-}{s_+ + s_-}$$

where g_i is the expression vector of i^{th} gene over all training samples, m_+ indicates the mean value of the i^{th} gene's expression in positive class, and s_+ is the standard deviation of the i^{th} gene's expression in positive class. m_- and s_- are defined for negative class in the same way.

Our goal is to select a small number of genes that can be used to construct robust, efficient and accurate classifiers for predicting future samples. In our experiments, we tried different gene selection methods such as choosing genes with the largest absolute value of P, choosing half from the top genes (highly correlative with positive class) and half from the bottom genes (highly correlative with negative class) or choosing only the top dozens of genes from all 3571 genes etc. We got the best result by choosing genes with the largest absolute value of P. Therefore, the genes that we used in our experiment are selected by using the method.

3. SVMs: Support vector machines

SVMs are widely used in many machine learning and data mining problems due to the superior performance in data analysis. The SVM algorithm finds the maximum margin hyperplane, which maximizes the minimum distance from the hyperplane to the closest training points [6]. The function corresponds to the optimal hyperplane used for classifying data is found by solving a convex quadratic optimization program [5, 12, 13] defined as:

$$\min_{(w,g,y)\in R^{n+1+m}} u e' y + \frac{1}{2} w'w$$
$$s.t. \quad D(Aw - eg) + y \geq e,$$
$$y \geq 0$$

where u is a positive weight value for controlling overfitting, $e \in R^m$ is a column vector with all ones , $A \in R^{m \times n}$ represents data points with m samples and n features, $y \in R^m$ is slack vector variable, and D is an $m \times m$ diagonal matrix with ones or minus ones along its diagonal to specify the class label of each sample.

By using smooth techniques introduced in [1, 13] and employing the KKT optimality conditions, the original problem is converted into the smooth support vector machine (SSVM) [13] defined as follows:

$$\min_{(w,g)\in R^{n+1}} \frac{u}{2} \| p(e - D(Aw - eg), a) \|_2^2 + \frac{1}{2}(w'w + g^2)$$

where $p(x,a) = x + \frac{1}{a}\log(1 + e^{-ax}), \quad a > 0$ is an accurate smooth approximation to plus function:

$$(x)_+ = \begin{cases} 0, & for \quad x < 0 \\ x, & for \quad x \geq 0 \end{cases}$$

This problem can be solved by the Newton-Armijo algorithm [13] which has be shown converges globally and quadratically [13]. The solution of this problem will give us the linear classification function:

$$f(x) = x'w + g$$

For a new unlabeled sample x, we plug it into the classification function. If the output value $f(x) > 0$, we classify this sample into positive category otherwise we classify it into negative category. In our experiments, all classifiers are established based on the SSVM algorithm.

4. Experiments and results

The leukemia dataset splits into AML, B-cell ALL and T-cell ALL categories. But the classifiers we build via SSVM algorithm are binary classifiers. Hence we have to convert this multicategory classification problem into a series of binary classification problems.

We calculate the distances between these three classes using their means of gene expression values. The distance between ALL (including B-cell ALL and T-cell ALL) and AML is larger than the distance between B-cell and T-cell ALL. So, it is nature to conclude that the difference between ALL and AML is more significant than B-cell ALL and T-cell ALL. This result consists with [2, 7, 9] and represents that the preliminary and normalization processing described in section 2 without changing the characteristics of expression data in different categories. Then the initial hierarchical classification model was built as Figure 1.

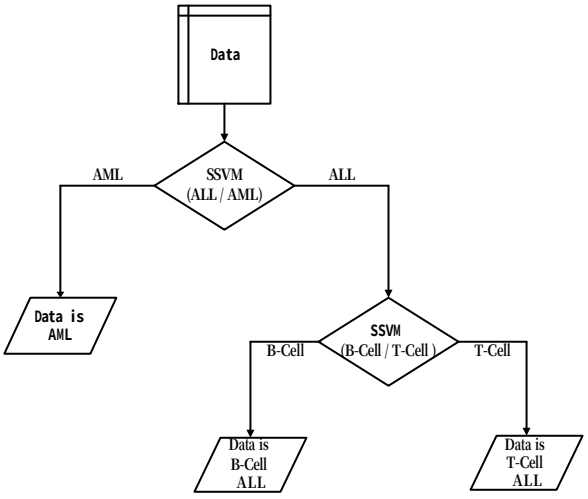

Figure 1. The initial hierarchical classification model

There are two diamonds we can see in figure 1. Each of them represents a classification function established by SSVM with linear kernel and deals with its own classification problem.

Combining with gene selection steps, we can get our hierarchical two-phase classification model, shown as figure 2. The dashed boxes are our two-phase classifiers. In the ellipses of gene selection I and gene selection II, 10 and 20 genes out of 3571 are selected for their own

classification tasks, respectively. The reasons that use 10 and 20 genes at different level are decided by accuracy, stability and robustness of SSVM classifiers built.

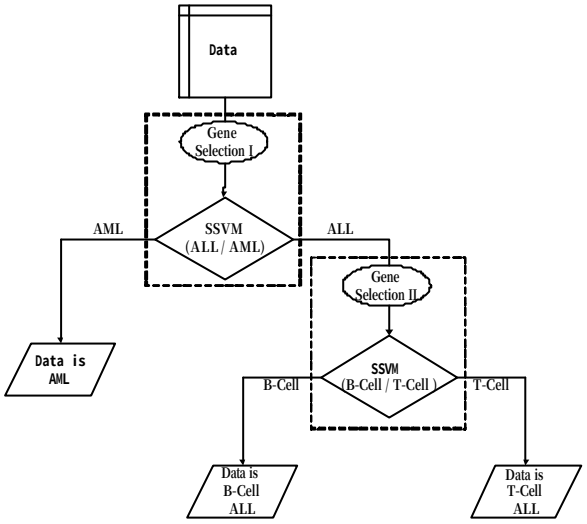

Figure 2. Hierarchical two-phase classification model

In order to fit these requirements, 10 genes for AML/ALL classifier and 20 genes for T-cell/B-cell classifier provid the best results in our experiments.

According to this model, 30 genes out of 3571 are selected and the two SSVM classifiers are also established when the training data was input. In testing stage, we can input data with genes sorted or unsorted. This hierarchical two-phase classification model will select genes correctly by itself at each level and classify testing data accurately.

The numerical results of our experiments and some other researchers' are summarized in table 2. The number of genes we used here is not the smallest subset that can archive the best result. Actually, we need only 18 genes (8 for AML/ALL classifier and 10 for B-cell/T-cell classifier) to have the same results in table 2 by advanced gene selection steps. But the built classifiers are too sensitive when noise data involved, and degrading the robustness and generic ability of classifiers. Hence we intend to use more genes but build robust classifiers.

For the purpose of showing the 30 genes selected for building two SSVM classifiers are informative, we depicted the means of these genes' expression data correspond to different classes in figure 3, 4, and 5. In figure 3, the means of 10 genes used for building ALL/AML SSVM classifier are graphed.

Table 2: Comparison of experimental results
(Tested by 34 independent testing samples)

Author	ALL / AML		B-ALL / T-ALL	
	# of genes	Errors	# of genes	Errors
T. Golub et al. [9]	50	2	N / A	N / A
S. Mukherjee et al. [7]	99	0	999	0
T. Furey et al. [10]	25	2	250	0
I. Guyon et al. [3]	8	0	N / A	N / A
D. Slonim et al. [2]	50	2	50	1
J Weston et al. [4]	20	0	5	0
Rui Xu et al. [11]	10	1	N / A	N / A
Our Results	*10*	*1*	*20*	*0*

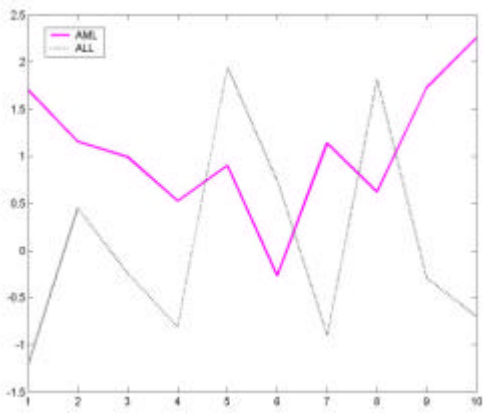

Figure 3. Means of 10 selected genes respect to ALL and AML categories. X-axis represents only a serial number of selected genes and without any specific meaning. Y-axis represents the mean value of selected gene expression data respects to different classes. The same specify is used in figure 4 and figure 5.

Similar to figure 3, figure 4 depicts the means of 20 genes used for building B-cell/T-cell SSVM classifier. It is apparent in these two figures that the means of identical gene respect to different classes are significantly

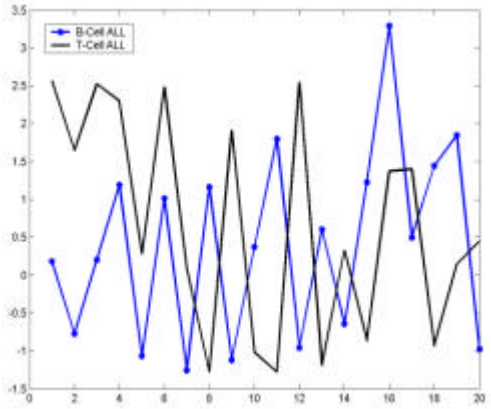

Figure 4. Means of 20 selected genes respect to B-cell ALL and T-cell ALL categories

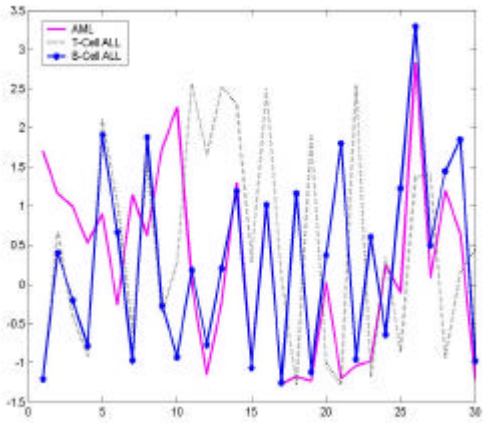

Figure 5. Means of 30 selected genes respect to AML, B-cell ALL and T-cell ALL categories. X-axis from 1 to 10 are genes selected for discriminating AML from ALL data, 11 to 30 are selected for classifying B-cell ALL and T-cell ALL.

dissimilar. Through these genes we can obtain sufficient information to build superior classifiers.

Finally, the means of all 30 genes selected by us are combined in figure 5. The genes correspond to x-axis from 1 to 10 are the same as the genes in figure 3, and genes numbered from 11 to 30 are the same as graphed in figure 4. We can obvious see that the means of gene expression data respect to AML are dramatically different from the means respect to two different ALL types in first 10 genes. And the means of last 20 genes respect to B-cell and T-cell samples are apparent different. By these figures we believed that the genes we used for establishing SSVM classifiers are informative and representative.

5. Conclusions and future works

To propose an efficient and powerful method for microarray gene expression data analysis is the object of this paper. Therefore, we design a hierarchical two-phase classification model to achieve this goal and get acceptable numerical results. We not only substantially reduce the number of genes needed from 7129 to 30, but also build classifiers with superior classification ability by using those limited informative genes.

In future works, we will apply this model to more different types of microarray gene expression datasets, more complicated kernels may be used and hope to get acceptable numeric results like in this leukemia dataset. In addition to, the function used to evaluate gene expression respect to different classes may be replaced by a more complex function instead of average function. The final goal we hope to fulfill is making all processing steps visualized as graphs rather than only numeric computation. For example, we may use only graphs like figure 3, 4 or 5 to accomplish gene selection steps and classify unknown samples (include multicategory classification) directly according to some information get from graphs. If this object can be achieved, the time needed for analyzing microarray gene expression data will be saved further, and the processes will be more understandable.

6. References

[1] Chun-Hui Chen and O. L. Mangasarian. "Smoothing methods for convex inequalities and linear complementarity problems", *Mathematical programming*, 71(1): 51-69, 1995.

[2] D. Slonim, P. Tamayo, J. Mesirov, T. Golub, E. Lander. "Class prediction and discovery using gene expression data", *Fourth Annual*

International Conference on Computational Molecular Biology, pp. 263-272, 2000.

[3] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. "Gene selection for cancer classification using support vector machines", *Machine Learning*, 2000.

[4] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik. "Feature selection for SVMs", *Advances in Neural Information Processing Systems 13*, T. K. Leen, T. G. Dietterich, and V. Tresp, Eds. Cambridge, MA: MIT Press, pp. 668–674, 2001.

[5] L. Kaufman. "Solving the quadratic programming problem arising in support vector classification", *in Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf, C. J. C. Burges, and A. J. Smola, eds., MIT Press, 47-167, 1999.

[6] M. Brown, W. Grundy, D. Lin, N. Cristianini, C. Sugnet, J. Manuel Ares, and D. Haussler. "Support vector machine classification of microarray gene expression data", Technical Report UCSC-CRL-99-09, Department of Computer Science, University of California, Santa Cruz, 1999.

[7] S. Mukherjee, P. Tamayo, D. Slonim, A. Verri, T. Golub, J. Mesirov, and T. Poggio. "Support vector machine classification of microarray data", AI Memo 1677, Massachusetts Institute of Technology, 1999.

[8] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, Chen-Hsiang Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. Mesirov, T. Poggio, W. Gerald, M. Loda, E. Lander, and T. Golub. "Multiclass cancer diagnosis using tumor gene expression signatures", *PNAS*, vol.98, no. 26, 15149-15154, December 18, 2001.

[9] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander. "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring", *Science*, 286 (5439): 531-537, October 1999.

[10] T. Furey, N. Cristianini, N. Duffy, D. Bednarski, M. Schummer, and D. Haussler. "Support vector machine classification and validation of cancer tissue samples using microarray expression data", *Bioinformatics*, vol. 16, pp. 906-914, 2000.

[11] Rui Xu, G. Anagnostopoulos, and D. Wunsch II. "Tissue classification through analysis of gene expression data using a new family of ART architectures", *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks* (IJCNN '02), vol. 1, pp. 300-304, Honolulu, Hawaii: IEEE, INNS, ENNS, 2002.

[12] V. Vapnik. *Statistical Learning Theory*, Wiley Interscience, 1998.

[13] Yuh-Jye Lee and O. Mangasarian. "SSVM: A smooth support vector machine for classification", *Computational Optimization and Applications*, 20, 5-22, 2001.

[14] http://bioinformatics.upmc.edu/Help/UPITTGED.html

[15] http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi