ALL/AML Cancer Classification by Gene Expression Data Using SVM and CSVM Approach*

Xuegong Zhang **Haixin Ke**
zhangxg@mail.tsinghua.edu.cn hxke@simba.au.tsinghua.edu.cn
Institute of Bioinformactics/ Department of Automation, Tsinghua University, Beijing 100084, China

Keywords: cancer classification, gene expression data, support vector machines, pattern recognition

1 Introduction

Cancer classification plays an important role in cancer treatment. There has been no general approach for this problem now. The tasks for cancer classification are of two aspects: identifying new cancer classes and assigning tumors to known classes, which are called class discovery and class prediction by Golub *et al.* [1]. From mathematical point of view, class discovery is a cluster analysis problem, while class prediction is usually called classification problem (we'll use the later name to keep consist with pattern recognition literatures).

Until now, cancer classification has been based primarily on morphological appearance of tumor [1]. This has serious limitations because of ambiguity. Golub *et al.* presented a new approach to cancer classification based on gene expression monitoring by DNA microarrays in [1]. They chose acute leukemia as a test case, and the target is to distinguish between ALL (acute lymphoblastic leukemia) and AML (acute myeloid leukemia), which is a typical cancer classification problem not well solved despite many years of efforts. This paper is a report of our work on the classification (prediction) part of this problem following their original work.

Golub *et al.* adopted a feature selection (gene selection) procedure before classification. A metric was defined to evaluate the correlation of each gene to the classification. After some "good" genes were selected from all the 6817 genes, the classification is done by a weighted voting scheme. The classifier was trained on a 38-sample training set, and another 34-sample set was used for testing. With leave-one-out cross-validation on the training set with 50 selected genes, 36 out of 38 samples were correctly classified and 2 were rejected (no-call). The performance on the test set was that 29 samples out of 34 were correctly classified and the other 5 were rejected. If the classifier were compelled to give these 5 no-calls a prediction, the prediction would be wrong.

Since the feature selection procedure is of single selection type, and the classification method is also an intuitive one, we believe that there is still much space for the performance to be improved. In our approach to the problem, we took all the genes for the classification (the selection problem will be discussed in another paper), and applied the support vector machine(SVM) method and one of its improved version CSVM as the classifier. Thanks to the better generalization ability of SVM and CSVM, much better performance was obtained.

2 Brief Introduction to SVM and CSVM

The basic idea of SVM is that when designing (training) the linear classifier, besides the goal of getting minimum classification errors, the classification margin should be maximized at the same time. This makes SVM be superior to other classifiers by having better generalization performance [3, 4].

*This work is supported by Nature Science Foundation of China, under project number 69885004.

The name of SVM comes from the fact that the final classification function of SVM usually only depends on part of the training samples, which are called support vectors, because they "support" the boundary. These support vectors can be viewed as critical samples in the training set. This fact allows us to be able to identify those most informative samples from a data set.

However, this feature of SVM also has disadvantages, because it makes the final decision function sensitive to certain specific samples in the set. Especially for cases where very few training samples are available (such as this ALL/AML case), this side effect must be considered. To overcome this shortcoming of SVM, Zhang proposed a modified version named CSVM or central support vector machine [2]. It adds the class-mean information into the standard SVM, thus makes the decision function less sensitive to specific samples which may have the risk to be noise or outliers.

We choose SVM and CSVM because of their better generalization ability based on limited training samples. They are well suited for the kind of task where the feature dimension is very high (6817 here) while the number of training samples is very small (only 38 here). The details of SVM and CSVM can be found in [3, 4, 5] and [2].

3 Experiment Results

Our experiment was done on the same data set as used in [1] by Whitehead/MIT MPR group [6]. We firstly applied the standard linear SVM on the ALL/AML data. All the 38 training samples were classified without error. 22 of them were selected as support vectors, 15 of ALL and 7 of AML. Table 1 gives the function output for each of the training samples. The decision is ALL if $f(x) > 0$ and otherwise AML. Samples that were selected as support vectors are indicated by a "*" in the table. Further study is undergoing about why these samples were regarded as more critical. We hope this may reveal clinically helpful cues.

Applying this trained SVM on the test set, all but two the samples were classified correctly. The 2 misclassified samples are call #47 and call #55, which are ALL but were assigned to AML with SVM function values -0.089355 and -0.059675, values among the smallest in all the test outputs. It should be noted that in SVM, relatively the smaller the absolute value of the output is, the nearer the sample is to the classification boundary.

Two samples were misclassified by SVM. This is very possibly due to the small size of the training set, which causes the SVM result rely too much on specific samples that may bias from the underlying true distribution. Thus we applied CSVM. As expected, the CSVM classifier only classified all the training samples correctly, but also made no mistake on all the 34 test samples. From the CSVM result, we can also find critical samples (with meaning slightly different from support vectors), but the number is less than the SVM case. For this case, only 3 most critical samples were identified, 2 of ALL(call #12 and #25) and 1 of AML(call #35).

4 Discussions

This paper reports our experiment results on applying SVM and CSVM for the problem of ALL/AML cancer classification by gene expression data obtained by DNA microchips. With stand SVM approach, the training error rate is zero and testing error rate is 2 out of 34. With CSVM, both errors are zero. These results over-perform the original work by Golub *et al.* Besides the better performance in class prediction precision, our approach also identified from the training set several samples as most critical samples for the distinction.

It should be mentioned that although we did not take feature selection procedure in this experiment, SVM provides the possibility for doing so, which is currently under our study. Since SVM considers not only the contribution of single features (genes), but also considers the cooperation between different

Table 1: SVM output of the training samples.

call #	f(x)	call #	f(x)
1	1.326683	20	1.402707
2*	1.000234	21*	1.000001
3	1.311848	22*	1.000001
4*	1.000000	23	1.610979
5	1.065201	24	1.185595
6	1.554635	25*	1.000917
7*	0.999954	26	1.101593
8*	1.000382	27*	1.000477
9	1.592384	28*	-0.999880
10*	1.000001	29*	-0.999859
11*	1.000103	30	-1.132313
12*	1.000001	31*	-0.999999
13	1.212180	32*	-1.000000
14*	1.000148	33	-1.377774
15	1.566983	34*	-1.000000
16	1.166551	35*	-0.999999
17	0.999884	36*	-1.618616
18	1000818	37*	-1.399201
19	1000001	38*	-0.999984

features (genes), we believe that the genes thus selected will be closer to those genes intrinsic to the distinction of cancer classes.

References

[1] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gassenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, 286(15):531–537, 1999.

[2] Joachims, T., Making large-scale SVM learning practical, *Advances in Kernel Methods - Support Vector Learning*, B. Schokopf *et al.* (ed.), MIT Press, 1999.

[3] Vapnik, V.N., *The Nature of Statistical Learning Theory*, Springer-Verlag, 1995.

[4] Zhang, X., Introduction to statistical learning theory and support vector machines, *Acta Automatica Sinica (in Chinese)*, 26(1):32–42, 2000.

[5] Zhang, X., Using class-center vectors to build support vector machines, *Neural Networks for Signal Processing IX*, 3–11, 1999.

[6] http://www.genome.wi.mit.edu/MPR/data_set_ALL_AML.html