

Characteristics of Support Vector Machines in Gene Expression Analysis

Daisuke Komura¹

komura@hal.rcast.u-tokyo.ac.jp

Hiroshi Nakamura¹

nakamura@hal.rcast.u-tokyo.ac.jp

Shuichi Tsutsumi¹

tsutsumi@genome.rcast.u-tokyo.ac.jp

Hiroyuki Aburatani²

haburata-tky@umin.ac.jp

Sigeo Ihara¹

ihara@genome.rcast.u-tokyo.ac.jp

¹ Research Center for Advanced Science and Technology, The University of Tokyo, 4-6-1 Komaba, Meguro, Tokyo 153-8904, Japan

² Genome Science Div., Center for Collaborative Research, The University of Tokyo, 4-6-1 Komaba, Meguro, Tokyo 153-8904, Japan

Keywords: support vector machine, Affymetrix GeneChip, feature selection

1 Introduction

Statistical analyses on DNA microarray expression data have made it possible to extract information from tissue and cell samples. Recently, Support Vector Machines (SVMs)[3], one of the supervised learning methods, have been employed to classify gene expression data and have shown greater performance than other learning methods, especially in the case where the number of features is larger than the number of samples[1]. However, the characteristics of SVMs are not clear. Thus, in this work, we will study how the number of features and regulatable parameters on SVMs affect performance of classification on DNA microarray. Gene expression profiles derive from various types of cancers using Affymetrix GeneChip, including our institute sample data.

2 Methods

In order to study the sensitivity of SVMs on various kernel parameters and the gene expression data, we have developed the Perl/Tk interface software based on svm-light[2](as shown in Fig.1(a)~ (c)). The DNA microarray data set we use consists of two classes; 16 cancerous livers and 8 noncancerous livers obtained from Affymetrix GeneChip. Before analysis, we normalize and filter the raw data. A quantile normalization procedure was used in the probe intensity distribution across different chips. We classify the data set by using SVMs with the polynomial kernel function: $K(x, y) = (x \cdot y + 1)^d$, where x and y are the vectors of the gene expression data. Here parameter d is an integer.

3 Results and Discussion

Parameter d is varied from 1 to 3, and the number of feature genes from 1 to 987. We perform two types of tests. In the first test, we select feature genes at random. In the second test, the feature genes are selected with calculating S/N ratio:

$$S(i) = \frac{\mu_+(i) - \mu_-(i)}{\sigma_+(i) + \sigma_-(i)},$$

where $\mu_{+(-)}$ is the mean of each class and $\sigma_{+(-)}$ is the standard deviation of each class. We sorted

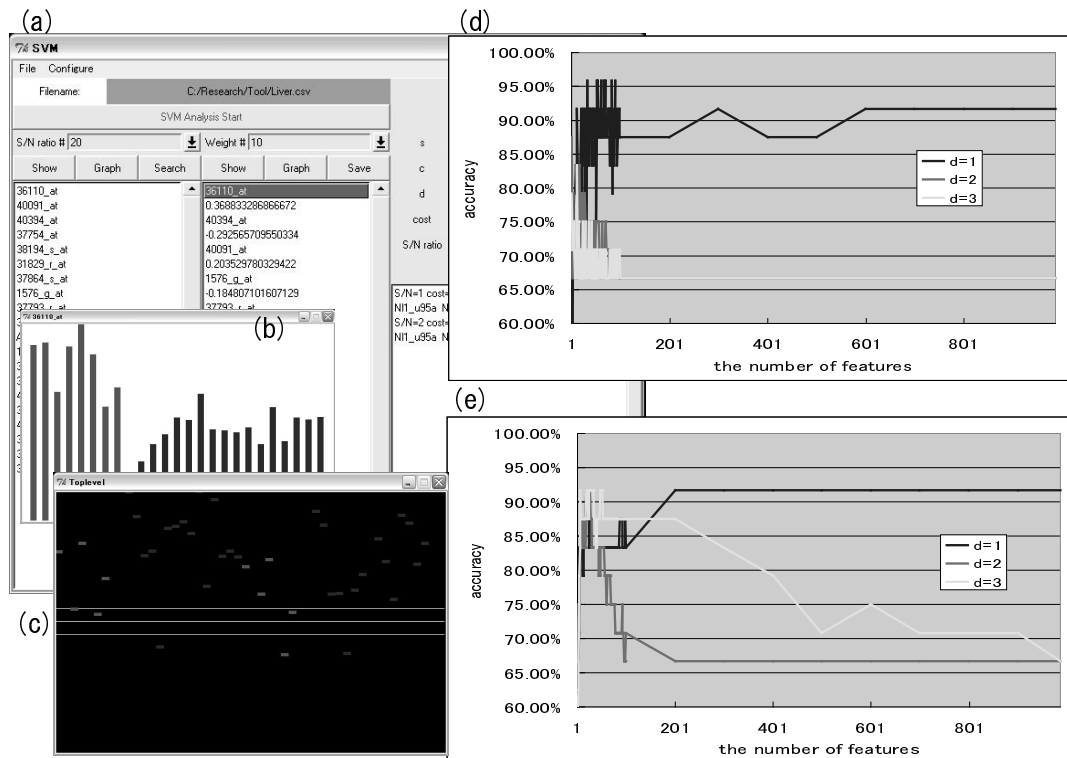


Figure 1: (a)(b)(c) The GUI interface, (d)(e) the accuracy of classification where the parameter d is varied from 1 to 3, and the number of feature genes (horizontal axis) from 1 to 987. In the chart (d), the feature genes are selected at random. In the chart (e), the feature genes are selected by calculating S/N ratio.

the genes by the score and select the feature genes from the top. The accuracy of class prediction is estimated by Leave-one-out method.

In Fig.1(d) and (e), the results where d equals to 1 perform the best. In the polynomial kernel function, parameter d decides a rough shape of a separator; in case d equals to 1, a linear classifier is generated, and in case d is equal to or more than 2, a nonlinear classifier is generated. In other words, these results indicate that the linear classification outperforms the nonlinear classifications on the data set. Overfitting causes the misclassification. If more samples are obtained (at least as many as the features) and they are not separable linearly, nonlinear classification may perform well. As far as the linear classification, increasing features do not adversely affect accuracy. This is because critical genes are likely to have larger weight in classifier and non-critical genes are not. This fact also suggests that SVMs can be used for feature selection in place of S/N ratio, with which the features do not always perform better than the features selected randomly.

Now we investigate the possibility of applying SVMs to feature selection. In addition, we will investigate about other kernel functions and other parameters on SVMs.

References

- [1] Furey, T.S., Cristianini, N.C., Duffy, N., Bednarski, D.W., Schummer, M., and Haussler, D., Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics*, 16(10):906–914, 2000.
- [2] Joachims, T., *Making large-scale SVM learning practical. Advances in kernel methods - support vector learning*, Scholkopf, B., Burges, C., and Smola, A. (ed.), MIT-Press, 1999.
- [3] Richard, O.D., Peter, E.H., and David, G.S., *Pattern Classification*, Wiley-Interscience, 2000.