Features of Gene Extraction by Nonlinear Support Vector Machines in Gene Expression Analysis

Daisuke Komura[1]
komura@hal.rcast.u-tokyo.ac.jp

Hiroshi Nakamura[1]
nakamura@hal.rcast.u-tokyo.ac.jp

Shuichi Tsutsumi[1]
tsutsumi@genome.rcast.u-tokyo.ac.jp

Hiroyuki Aburatani[2]
haburata-tky@umin.ac.jp

Sigeo Ihara[1]
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 Machines, DNA microarray, feature selection, Recurisve Feature Elimination

1 Introduction

Statistical analysis on gene expression data from DNA microarray has enabled us to extract information from tissue and cell samples. Comparing two classes of gene expression datasets (e.g. datasets from normal tissues and cancerous tissues), we first choose discriminative genes, which have significantly different expression values between two classes and characterize each class. In the most case of statistical filter methods such as t-test, the chosen genes tend to be strongly expressed in one class and weakly in the other class. However, the lower ranked genes by the simple filter methods may also include discriminative and informative genes. The expression values of such genes have complicated distributions. Suppose there are two classes whose distribution is as depicted in Figure 1: there are two peaks of frequency of gene expression values in one class, while the peak in the other class is located between them. In this case, although such a gene is of importance, it is low ranked and thus discarded by conventional filter methods because of its low inner-class average and its high deviation.

In supervised classification problems, various wrapper methods such as Recursive Feature Elimination (RFE) [1] are proposed for feature selection. Since the main purpose of the wrapper method is to improve the performance of the classification algorithm, the genes chosen by the method have not been paid attention, especially in the nonlinear classification. In this paper, we use the wrapper methods based on a nonlinear classification algorithm in order to extract the discriminative genes that difficult to be extracted by conventional filter methods. RFE method based on nonlinear Support Vector Machines (SVMs) [2] is employed to this end because it is successfully applied to classification of gene expression data. We investigate the genes extracted by the RFE method based on SVMs with gaussian kernel function to indicate that it can extract discriminative genes which are not chosen by conventional filter methods.

2 Methods

In order to extract discriminative genes which are not chosen by conventional filter methods, we apply the RFE method based on nonlinear SVMs to acute lymphoblastic leukemia (ALL) dataset [3]. This dataset consists of six subtypes of ALL and is obtained by hybridization on the Affymetrix U95A Genechip containing probe sets for 12558 genes. We select 60 TEL-AML1, 32 T-ALL and 48 Hyperdiploid samples among them. We divide these samples into two classes: Class +1 consists

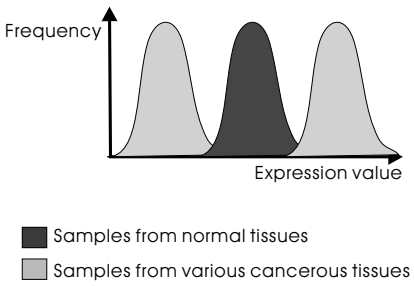

Figure 1: The case where the discriminative gene is low ranked by conventional filter methods.

Table 1: P-values for t-tests about the genes chosen by the RFE method based on SVMs with gaussian kernel. Class +1 consists of TEL-AML1 samples and Class -1 consists of T-ALL and Hyperdiploid samples.

rank	probe set	Class +1 vs. Class -1	TEL-AML1 vs. T-ALL	TEL-AML1 vs. Hyperdip
1	37519_at	0.20	0.29	0.20
2	776_at	0.21	0.43	0.070
3	41007_at	0.21	0.37	0.15
4	36927_at	0.21	0.014	0.027
5	36605_at	0.24	9.1E-09	0.024
6	38558_at	0.20	0.078	0.42
7	36115_at	0.21	0.33	0.20
8	40836_s_at	0.22	0.49	0.11
9	36591_at	0.23	3.0E-08	0.00034
10	847_at	0.22	0.34	0.069

of TEL-AML1 samples and Class -1 consists of T-ALL and Hyperdiploid samples. Before analysis, we discard genes with the values' overall average less than 2500 or its overall coefficient of variation less than 0.35 for each gene of the dataset. The genes whose p-value for t-test between two classes is more than 0.20 are also discarded to remove linearly discriminative genes. Finally 616 genes are remained. Then we classify the dataset by using SVMs with the gaussian kernel function $K(x, y) = \exp(-\gamma\|x - y\|^2)$ and extract genes by the RFE method. As a result of empirical parameter tunings, the value of γ is set to 0.001 in the numerical experiment.

3 Results and Discussion

Table 1 shows the p-values for three t-tests about each of top 10 genes ranked by the RFE method based on SVMs with gaussian kernel. There is a tendency that the higher ranked genes have lower p-values for t-tests of Class +1 vs. Class -1. However, the 5th and 9th genes have the relatively high p-values for t-tests of Class +1 vs. Class -1. This means that these genes would be discarded by t-test. Both two genes have very low p-values for the t-test of TEL-AML1 vs. T-ALL and those of TEL-AML1 vs. Hyperdiploid. Additionally, the average expression values of TEL-AML1 of these two genes are lower than those of T-ALL but higher than those of Hyperdiploid. These facts show that the method successfully chooses the genes whose distribution of the expression is shaped like the distribution depicted in Figure 1. We find that the 4th gene also has the kind of distribution.

In this paper, we indicate that RFE method based on nonlinear SVMs can extract discriminative genes which can not be extracted by filter methods. Such genes are also useful for clustering analysis. We expect that they deepen our understandings of various diseases.

References

[1] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V., Gene selection for cancer classification using support vector machines, *Machine Learning*, 46:389–422, 2002.

[2] Vapnik, V., *Statistical Learning Theory*, Wiley, New York, NY, 1998.

[3] Yeoh, E., Ross, M., Shurtleff, S., Williams, W., Patel, D., Mahfouz, R., Behm, F., Raimondi, S., Relling, M., Patel, A., Cheng, C., Campana, D., Wilkins, D., Zhou, X., Li, J., Liu, H., Pui, C., Evans, W., Naeve, C., Wong, L., and Downing, J., Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling, *Cancer Cell*, 1:133–143, 2002.