Prognosis Prediction by Microarray Gene Expression Using Support Vector Machine

Chihoko Tago tago@brs.kyushu-u.ac.jp
Taizo Hanai taizo@brs.kyushu-u.ac.jp
Masahiro Okamoto okahon@brs.kyushu-u.ac.jp
Laboratory for Bioinformatics, Graduate School of Systems Life Sciences, Kyushu University, 6-10-1 Hakozaki, Higashiku, Fukuoka 812-8581, Japan

Keywords: outcome prediction, support vector machine, microarray, k-NN

1 Introduction

Diffuse large B-cell lymphoma (DLBCL) is most common disease in non-Hodgkin's lymphomas. The research related to the prognostic prediction of patients is almost untouched despite of recent progress in clinical study of cancer. In recent years, the categorization of cancer was attempted based on expression profile of genes obtained by DNA microarray [3]. In this report, by using published microarray data [6], we have predicted survival of DLBCL patients with support vector machine (SVM) [5], which is one of the outstanding supervised machine learning method against classification problem. Furthermore, imputing missing value and gene selection, the models based on SVM were constructed. This model can achieve high accuracy for prognosis of patients.

2 Material and Method

2.1 Gene Selection and Prediction by SVM

Before classification by SVM, we performed gene selection by Fisher's criterion [4] and parameter increasing method (PIM). Gene selection should lead to the improvement in classification accuracy except for a surplus gene. As the first step, we selected 100 genes as the higher rank by Fisher's criterion, followed by, PIM to the 100 genes.

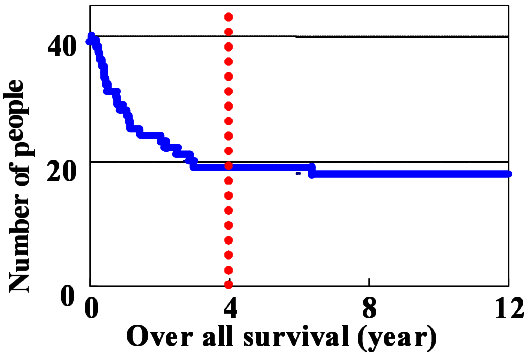

Figure 1: Kaplan-Meier plot of 40 patients.

The procedure for PIM can be surmarrized as follows; 1) set empty subset, 2) add one gene to subset followed by the caluculation of sum-squared error (f) by SVM, 3) continue procedure 2 until the value of f goes up. 4) determin the number of genes in subset.

2.2 Data for Analysis

In this study, we used published gene expression data about DLBCL [1, 6]. This data set consisted of 40 samples that have information about over all survival, and each sample has expression profile of 4026 genes. By setting threshold value of overall survival at 4 years (Fig. 1), data of prognosis can be categorized into two groups, survival or death.

2.3 Missing Values Estimation by K-Nearest Neighbor

Among 4026 genes, 1980 genes could not apply to classification because of missing values. However, since there might be involved important information for classification, we have imputed missing values by k-Nearest Neighbor (k-NN) method [2] before applying to SVM. K-NN method can select the gene with the minimum Euclid distance between the targeted gene with missing value followed by imputing the values of selected gene to the targeted gene.

3 Result and Discussion

The Leave One Out validation was carried out for the fairly test predictions for 40 patients. Results are shown in Table 1.

Table 1: Predictive accuracy by SVM.

Method	Initial number of genes	Selected number of genes	Predictive accuracy
Without k-NN	2046	2046	62.5%(25/40)
Without k-NN + gene selection	2046	11	90.0%(36/40)
With 1-NN + gene selection	4026	5	95.0%(38/40)

As shown in Table 1, the predictive accuracy was improved by applying gene selection and missing value estimation. The SVM model consisting of selected 5 genes shows the highest predictive accuracy (95%). These 5 genes might be biologically important genes for prognosis. The name of five selected genes for prognosis are JNK3, E2F-3, fvt-1, and remaining two genes are labeled "Unknown". It was reported that a normal cell change to cancer cell without E2F-3.

For further study we have to examine whether these selected 5 genes can be adopted as clinical marker genes against DLBCL.

References

[1] Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T., Hudson, J.Jr., Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Levy, R., Wilson, W., Grever, M.R., Byrd, J.C., Botstein, D., Brown, P.O., Staudt, L.M., Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature*, 403:503–511, 2000.

[2] Duda, R.O., Hart, P.E., and Stork, D.G., *Pattern Classification 2nd*, John Wiley & Sons, 2001.

[3] Golub, T.R. *et al.*, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, 286:531–537, 1999.

[4] Iizuka, N. *et al.*, Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection, *Lancet*, 361(9361):923–929, 2003.

[5] Vapnik, V.N., *The Nature of Statistical Learning Theory*, Splinger, 1995.

[6] The data set is available at `http://llmpp.nih.gov/lymphoma/`