ELSEVIER

Biochemical Engineering Journal 33 (2007) 100–109

Biochemical Engineering Journal

www.elsevier.com/locate/bej

New cancer diagnosis modeling using boosting and projective adaptive resonance theory with improved reliable index

Hiro Takahashi [a,b,c], Yasuyuki Murase [a], Takeshi Kobayashi [d], Hiroyuki Honda [a,*]

[a] *Department of Biotechnology, School of Engineering, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan*
[b] *Research Fellow of the Japanese Society for the Promotion of Science (JSPS), Japan*
[c] *Genetics Division, National Cancer Center Research Institute, 5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan*
[d] *School of Bioscience and Biotechnology, Chubu University, Matsumoto-cho 1200, Kasugai, Aichi 487-8501, Japan*

Received 22 April 2006; received in revised form 20 June 2006; accepted 5 August 2006

Abstract

An optimal and individualized treatment protocol based on accurate diagnosis is urgently required for the adequate treatment of patients. For this purpose, it is important to develop a sophisticated algorithm that can manage large amount of data, such as gene expression data from DNA microarray, for optimal and individualized diagnosis. Especially, marker gene selection is essential in the analysis of gene expression data.

In the present study, we developed the combination method of projective adaptive resonance theory and boosted fuzzy classifier with SWEEP operator method for model construction and marker selection. And we applied this method to microarray data of acute leukemia and brain tumor. The method enabled the selection of 14 important genes related to the prognosis of the tumor. In addition, we proposed improved reliability index for cancer diagnostic prediction of blinded subjects. Based on the index, the discriminated group with over 90% prediction accuracy was separated from the others.

PART-BFCS with improved $\mathrm{RI}_{\mathrm{BFCS}}$ method does not only show high performance, but also has the feature of reliable prediction further. This result suggests that PART-BFCS with improved $\mathrm{RI}_{\mathrm{BFCS}}$ method has the potential to function as a new method of class prediction for diagnosis of patients.

Keywords: Cancer diagnosis; Fuzzy classifier; Projective adaptive resonance theory; Marker gene selection; Reliability index

1. Introduction

Cancer is a major cause of disease related to human deaths in many developed countries. Frequently, the prognosis of cancer patients with the same clinical diagnosis can be different. Therefore, it is important that the prognosis of cancer patients is accurately determined, and an adequate treatment is proposed. However, the sensitivity of cancer patients to radiotherapy and/or chemotherapy is determined by complex causality involving multiple factors, and not a single factor because the mechanisms of cancer development (or malignancy) are extremely complex. Gene expression data from DNA microarray are individualized and are useful in the diagnosis and prognosis of diseases [1]. However, to conduct analysis, it is necessary to select significantly differentially expressed genes that are strongly related to diagnosis or prognosis of disease because the performance of classification analysis can decline due to such large quantities of data.

Feature selection has been performed in order to screen candidate genes for modeling. There are two types of approaches—wrapper approach and filter approach. In the former approach, features (genes) are selected as a part of mining algorithms, such as support vector machines (SVM) [2], fuzzy neural network (FNN) combined with SWEEP operator (FNN-SWEEP) method [1], and boosted fuzzy classifier with SWEEP operator (BFCS) method [3]. On the other hand, in the filter approach, features are selected by filtering methods, such as *U*-test, *t*-test, signal-to-noise statistic (S2N) [4] and projective adaptive resonance theory (PART) [5], prior to the application of mining algorithms.

These methods were often used alone in previous studies. In the present study, we combined various wrapper and filtering approaches and then, we applied these methods to gene expression profile data of leukemia and central nervous system tumor. It is necessary that specific and essential marker genes are selected for cancer classification and diagnosis. Minimum gene sets with-

[*] Corresponding author. Tel.: +81 52 789 3215; fax: +81 52 789 3214.
E-mail address: honda@nubio.nagoya-u.ac.jp (H. Honda).

1369-703X/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.bej.2006.08.004

out false positive ones should be extracted. Therefore, various methods were compared under the condition of small inputs. The combination method of PART and BFCS was the best under this condition.

2. Materials and methods

2.1. Data processing

We used two kinds of gene expression profiles. The first one is the gene expression profiles, obtained from http://www.genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi, reported by Golub et al. [4]. The data set comprised 7129 human genes (probe sets) and 72 patients (47 acute lymphoblastic leukemia (ALL) and 25 acute myeloid leukemia (AML)), which were obtained from acute leukemia patients at the first time of diagnosis. In this experiment, the data set was partitioned into one data set comprised of two groups: 38 patients (27 ALL, 11 AML) as a modeling data set for constructing the class prediction model (predictor) and 34 patients (20 ALL, 14 AML) as a blinded data set for evaluating the constructed predictor. We excluded those genes for which all the 72 patients showed an intensity of less than 1000 signals [6] prior to applying the various filtering methods. Thus, 2476 genes were selected for the present study.

The second one is gene expression data set of medulloblastoma, which is a type of central nervous system (CNS) tumor, obtained from http://www.genome.wi.mit.edu/MPR/CNS, reported by Pomeroy et al. [7]. Patients with medulloblastoma are treated by combinations of surgery, radiotherapy, and chemotherapy. In the present data set, the following three drugs are mainly used for chemotherapy: vincristine, cisplatin, and cytoxan. Therefore, by using gene selection and prognosis modeling proposed in the present study, the gene related to the treatment response can be extracted. The data set comprised 7129 human genes (probe sets) and 60 patients from whom tumor specimens were obtained by surgery. Among these 60 patients, a few patients (16) had a short follow-up period. Therefore, we used the data of the remaining 44 patients for the construction of a 4-year survival prediction model. Of these 44 patients, 26 patients remained alive after 4 years and 18 patients had died. In this experiment, the data set was randomly partitioned into three data sets consisting of two groups: 30 or 29 patients (18 or 17 survivors, 12 dead) as a modeling data set for constructing the class prediction model (predictor) and 14 or 15 patients (8 or 9 survivors, 6 dead) as a blinded data set for evaluating the constructed predictor. We excluded those genes for which all the 44 patients showed the intensity of less than 1000 signals prior to applying the various filtering methods. Thus, 2713 genes were selected for the present study.

In order to validate performance of models, 10 independent predictors were constructed from these genes by the parameter increasing method (PIM). The prediction accuracy of the blinded data set was utilized for comparison of model performance, and the accuracy was calculated as the average of 10 independent combination predictors.

A total of 1000 genes were selected by various gene screening methods, e.g. Mann—Whitney's *U*-test, signal-to-noise statistic (S2N), and projective adaptive resonance theory (PART), prior to the model construction step. Subsequently, various modeling methods were applied as described in the following sections.

2.2. Determination of optimal input number

When a large number of inputs are provided in the model, the model is excess fitted to the training data and the robustness is lost. Therefore, in order to construct a model with relatively high robustness, we assumed that the number of IF-THEN rules should not exceed the sample number [1]. Then, we used a stopping condition in the present study such that the total input number became $N_{\text{attribute}}$ in all the selected weak learners; $N_{\text{attribute}}$ is defined according to the following condition:

$$N_{\text{attribute}} < \log_2 N \tag{1}$$

where $N_{\text{attribute}}$ indicates the optimum selected attribute number.

Using Eq. (1), $N_{\text{attribute}}$ is 4 since N is 30 (or 29) for the CNS data set and 5 since N is 38 for the leukemia data set.

2.3. Boosted fuzzy classifier with SWEEP operator (BFCS)

Boosting was proposed by Schapire [8], and thus far, several derivative boosting algorithms [9–11] have been developed. Boosting is useful for class prediction using high dimensional inputs and is very fast algorithms.

In the previous study, we developed a boosted fuzzy classifier with SWEEP operator (BFCS) method [3] on the basis of AdaBoost [9], which is the most basic boosting algorithm. This method enables the evaluation of reliability of the predictions for each patient. On the other hand, it is difficult to evaluate the reliability of the predicted results of the conventional boosting.

Fig. 1 shows the structure of BFCS. BFCS is composed of one-input type I fuzzy neural network (FNN) models [12]. In the present study, one-input FNN models were used as weak learners in the BFCS model, and they were combined by connection weights, which were determined by the AdaBoost algorithm. FNN has three types of weight parameters (w_c, w_g, and w_f) [12]. In the present study, parameter w_g is a constant value ($=2.0\ln((1.0+0.995)/(1.0-0.995))$) [12], and w_c is a threshold that has the best odds ratio in the case that only one input was used. w_c and w_g were determined; w_f was calculated by the SWEEP operator method [12].

2.3.1. Reliability index for BFCS (old RI_{BFCS})

Reliability index (RI) based on fuzzy inference has been proposed to evaluate the result of class prediction by Huang and Li [13]. We have developed a reliability index for BFCS (RI_{BFCS}) by modifying RI for boosting.

We modified RI equations as follows:

$$\text{RI}_{\text{BFCS}} = \begin{cases} \text{INT}(\text{diff}_{\text{BFCS}} \cdot 10) + 1, & \text{if } 0 \le \text{diff}_{\text{BFCS}} < 0.9 \\ 10, & \text{if } \text{diff}_{\text{BFCS}} \ge 0.9 \end{cases} \tag{2}$$

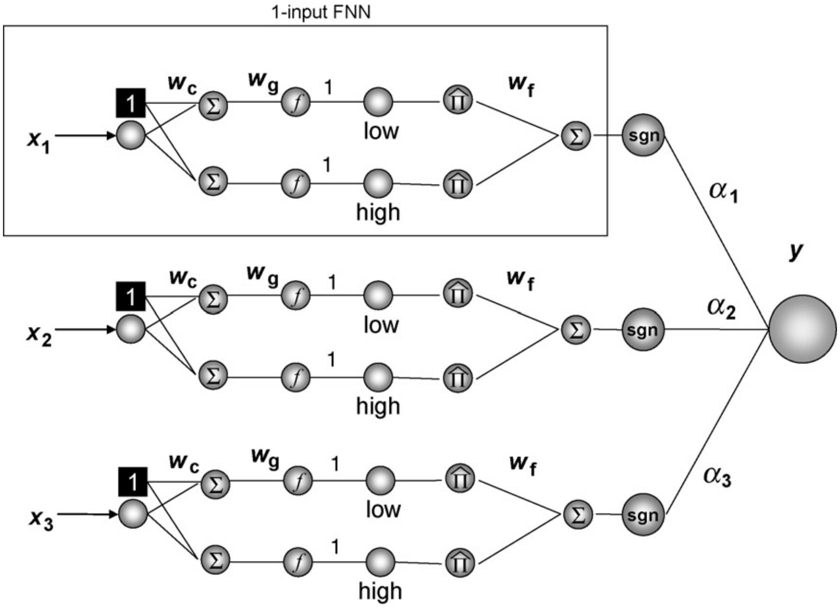


Fig. 1. Concept of the BFCS model.

where

$$\mathrm{diff}_{\mathrm{BFCS}} = \frac{\sum_t^T \left\{ \alpha_t \cdot \underset{v \in M_t}{\mathrm{argmin}}(\mathrm{diff}_v) \right\}}{\sum_t^T \alpha_t} \tag{3}$$

where T indicates the number of weak learners in the BFCS model, M_t indicates set of input variables in tth selected weak learner, α_t indicates the connection weight of the tth selected model in the construction of BFCS models, and diff_v is defined by the following equation:

$$\mathrm{diff}_v = u_{\mathrm{highest}}(X_v) - u_{\mathrm{next\ highest}}(X_v) \tag{4}$$

where v indicates the vth input in the BFCS model and $u(x_v)$ indicates the grade of the fuzzy membership function when the vth input x_v was inputted. It is defined by the following equation [12]:

$$u_v = \frac{1}{1 + \exp\{-w_g(x_v + w_c)\}} \tag{5}$$

$\mathrm{RI}_{\mathrm{BFCS}}$ is calculated for each example. Here, the greater $\mathrm{RI}_{\mathrm{BFCS}}$ the sample has, the more reliable its prediction.

2.3.2. *Improved reliability index for BFCS (new RI_{BFCS})*

In the present study, we propose improved reliability index by modifying equation of $\mathrm{RI}_{\mathrm{BFCS}}$ for more practical cancer diagnosis. For previous reliability index, argmin(diff_v)s in each weak learner, that mean distance from boundary line, are multiplied by α_t and summed. For improved reliability index, argmin(diff_v) in weak learner that output opposite to integrated model, is used as negative value. It is defined by the following equation:

$$\mathrm{diff}_{\mathrm{BFCS}} = \frac{\sum_t^T \left\{ \alpha_t \cdot g_t \cdot \underset{v \in M_t}{\mathrm{argmin}}(\mathrm{diff}_v) \right\}}{\sum_t^T \alpha_t} \tag{6}$$

where

$$g_t = \begin{cases} -1, & \text{if sign } (O_t) \neq \text{sign } (O_I) \\ +1, & \text{if sign } (O_t) = \text{sign } (O_I) \end{cases} \tag{7}$$

where O_t indicates output of tth model, and O_I indicates output of integrated model.

2.4. *k-Nearest neighbor (kNN)*

The k-nearest neighbor (kNN) methods are based on a distance function for pairs of tumor samples, such as the Euclidean distance. The kNN proceeds as follows to classify blind data set observations on the basis of the modeling data set. For each patient in the blind data set (a) finding the k-closest patients in the modeling data set and (b) predicting the class by majority vote; that is, choosing the class that is most common among those k-neighbors. The number of neighbors $k = 3$ was used because a similar cross-validation accuracy of model was obtained in the modeling data set for various k.

2.5. *Multiple regression analysis (MRA)*

The multiple regression analysis (MRA) is one of conventional methods. The MRA is a concerned with describing and evaluating the relationship between a patient's outcome and gene expression. MRA models are used to help us predict patient's outcome by using gene expression data.

2.6. *Weighted voting (WV)*

The weighted voting (WV) method was originally proposed by Golub et al. [4] to manage microarray data. The weights of each gene were calculated by the signal-to-noise. The linear models of one gene were assembled with gene weight.

2.7. Support vector machine (SVM)

The support vector machine (SVM) was originally proposed by Vapnik and Chervonenkis [14] and is used to avoid the "curse of dimensionality". SVM is superior to many other conventional methods and is frequently used in bioinformatics. In the present study, the SVM-LIGHT software package [15] was used. It was modified, and a PIM function was added to select a combination of inputs. In the present study, the regulatory parameter c was the default value of SVM_LIGHT ((avg(input vector)^2)$^{-1}$). A linear kernel was used because a similar cross-validation accuracy of model was obtained in the modeling data set for various kernels.

2.8. Fuzzy neural network (FNN) combined with SWEEP operator method (FNN-SWEEP)

The fuzzy neural network (FNN) combined with SWEEP operator method (FNN-SWEEP method) was also applied for model construction. The FNN-SWEEP method was originally proposed by Noguchi et al. [16] and was modified by Ando et al. [1] to manage microarray data. FNN has three types of weight parameters (w_c, w_g, and w_f) [12] as shown in Fig. 2. If w_c and w_g are fixed, FNN can be treated as multiple linear regression model in which w_f is variable parameter. Therefore, w_f was easily optimized without training. In the FNN-SWEEP method, only parameter w_f was optimized by the SWEEP operator method during the feature selection step. After the input combinations were determined, FNN models with the selected input combinations were optimized using a backpropagation algorithm on model construction step. In the backpropagation algorithm, the number of epochs was set to

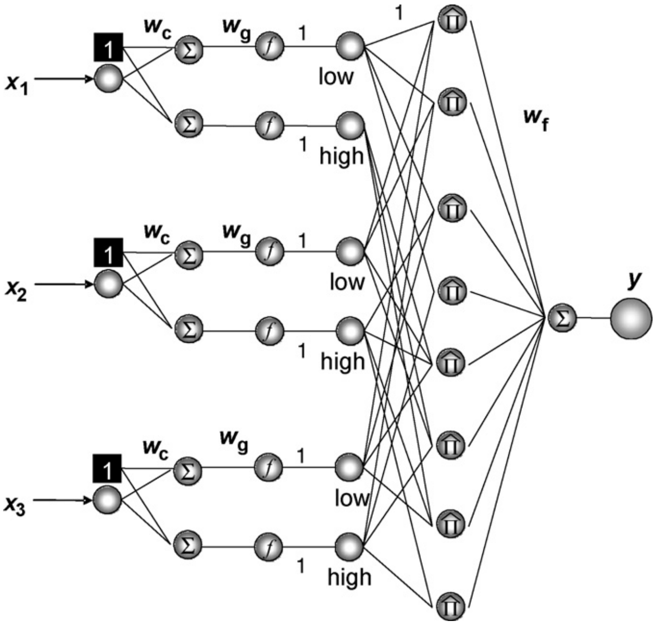


Fig. 2. Three-input type-I FNN model.

Table 1
Comparison of accuracies on various combination methods for leukemia data set (%)

	Inputs				
	1	2	3	4	5
BFCS with PART	77.9 ± 10.7	67.4 ± 7.6	84.7 ± 7.4	86.5 ± 4.4	89.1[a] ± 7.3
BFCS with S2N	78.8 ± 10.6	67.4 ± 7.6	84.4 ± 7.3	85.6 ± 5.7	83.2 ± 2.2
BFCS with *U*-test	78.8 ± 10.6	67.4 ± 7.6	84.4 ± 7.3	85.6 ± 5.7	83.2 ± 2.2
BFCS without screening	78.8 ± 10.6	67.4 ± 7.6	84.4 ± 7.3	85.6 ± 5.7	83.2 ± 2.2
SVM with PART	77.4 ± 10.0	79.4 ± 7.5	80.0 ± 8.2	80.9 ± 9.7	82.4 ± 8.4
SVM with S2N	76.2 ± 11.2	78.5 ± 7.0	81.8 ± 7.7	83.2 ± 9.0	82.4 ± 9.7
SVM with *U*-test	76.2 ± 11.2	78.5 ± 7.0	82.6 ± 6.2	84.1 ± 6.7	83.5 ± 8.0
SVM without screening	76.2 ± 11.2	78.5 ± 7.0	83.5 ± 6.2	84.7 ± 6.4	85.0 ± 7.7
FNN-SWEEP with PART	77.6 ± 12.2	77.1 ± 13.1	79.7 ± 9.1	80.3 ± 8.1	85.9 ± 7.7
FNN-SWEEP with S2N	77.9 ± 11.9	80.3 ± 7.8	81.8 ± 8.0	81.5 ± 8.2	81.5 ± 9.0
FNN-SWEEP with *U*-test	77.9 ± 11.9	80.3 ± 7.8	81.2 ± 7.5	82.6 ± 9.3	81.2 ± 8.5
FNN-SWEEP without screening	77.9 ± 11.9	80.3 ± 7.8	81.8 ± 8.0	84.4 ± 9.0	83.5 ± 8.7
*k*NN with PART	80.3 ± 11.8	75.3 ± 11.8	76.5 ± 11.8	80.0 ± 12.3	77.6 ± 12.5
*k*NN with S2N	79.1 ± 12.8	82.9 ± 12.8	82.6 ± 12.5	79.7 ± 9.8	79.4 ± 9.1
*k*NN with *U*-test	79.1 ± 12.8	84.1 ± 9.9	82.1 ± 9.0	81.5 ± 10.5	81.8 ± 10.8
*k*NN without screening	79.1 ± 12.8	79.4 ± 12.4	80.0 ± 11.3	78.8 ± 10.7	81.5 ± 9.3
MRA with PART	77.4 ± 11.2	79.4 ± 10.9	79.4 ± 10.3	75.3 ± 11.4	64.1 ± 8.2
MRA with S2N	77.9 ± 11.1	80.6 ± 8.8	83.2 ± 7.7	74.7 ± 9.6	64.7 ± 8.2
MRA with *U*-test	77.9 ± 11.1	80.6 ± 8.8	83.5 ± 8.0	76.2 ± 9.7	67.1 ± 7.0
MRA without screening	77.9 ± 11.1	80.6 ± 8.8	83.8 ± 8.2	76.2 ± 7.0	66.8 ± 6.8
WV with PART	79.7 ± 10.7	76.5 ± 12.5	82.4 ± 7.0	75.3 ± 8.6	72.4 ± 11.2
WV with S2N	78.2 ± 11.2	83.5 ± 7.5	70.9 ± 13.1	71.2 ± 12.6	70.6 ± 10.1
WV with *U*-test	78.2 ± 11.2	85.6 ± 5.8	76.2 ± 10.7	73.2 ± 14.2	76.2 ± 11.5
WV without screening	78.2 ± 11.2	78.8 ± 7.9	76.2 ± 13.6	77.1 ± 10.7	85.3 ± 9.4

The average blinded accuracies and their S.D.s were calculated from 10 combination models constructed by PIM.
[a] The highest accuracy.

5000, and the learning rate was set to 0.1, these values are the same as those reported by Ando et al. [1].

2.9. Model construction with parameter selection

The parameter increasing method (PIM) [17] was used to select input combinations for model construction of FNN-SWEEP, SVM, *k*NN, MRA, and WV. This was done as follows.

First, we predicted the subtype of each sample by using the prediction model with a single input. Prediction models for each probe were constructed in a series, and all the probes were ordered based on the accuracy of the constructed models. In the next step, the probe having the highest accuracy level was used for constructing a combination model.

Second, we selected a partner probe for the probe selected in the first step in order to increase the prediction accuracy. To accomplish this, we constructed a two-input model in which a ranked probe was designated as input 1, and input 2 (partner probe) was selected to provide the highest training accuracy while applying FNN-SWEEP (or SVM, *k*NN, MRA, and WV) and PIM to the modeling data. By repeating this step, a combination of $N_{\text{attribute}}$ candidate probes was identified for use as input probes in the model construction.

Finally, combinations of $N_{\text{attribute}}$ probes, i.e. from the first to the $N_{\text{attribute}}$th probe were evaluated. We constructed $N_{\text{attribute}}$ predictor models, beginning with one input using only the first-selected probe to $N_{\text{attribute}}$ inputs using all the $N_{\text{attribute}}$ probes. The predictor models were specifically constructed by using a backpropagation algorithm for FNN-SWEEP or quadratic programming for SVM. The performance of the prediction models was evaluated by applying them to the blinded data set.

For the two gene expression profile data, the genes with the first to the 10th highest accuracies were used as the first inputs for the construction of the 10 combination models by PIM. The S.D.s of blinded accuracies were calculated by using ones of these 10 combination models.

2.10. PART-BFCS method

Previously, we developed PART filtering method by modifying PART [18,19]. And, we developed and combined the PART filtering method as a gene filtering method and BFCS as a modeling method. In this PART-BFCS method, PART first preselects the genes that show small variances within a class. Then, BFCS rapidly selects these genes to build a highly accurate and reliable predictor.

PART has two important parameters, vigilance and distance parameters. The vigilance parameter was optimized so that modeling samples clustered well. The distance parameter was used to control the number of extracted genes. The genes extracted by PART showed low standard deviation (S.D.) in lower gene expression class. The predictor using genes with low S.D. in lower class showed high performance [5].

In BFCS model, one-input FNN models on the basis of neural network and fuzzy logic, were used as weak learners. FNN has three types of connection weights (w_c, w_g, and w_f). These parameters were optimized as mentioned in section of BFCS algorithm. The only one parameter that should be optimized is the number of input in boosting model. This parameter was optimized by using the number of samples.

Table 2
Frequency of construction of high performance model

	Methods	
	Leukemia[a]	CNS[b]
BFCS with PART	4/10	13/30
BFCS with S2N	0/10	3/30
BFCS with *U*-test	0/10	3/30
BFCS without screening	0/10	3/30
SVM with PART	2/10	2/30
SVM with S2N	1/10	2/30
SVM with *U*-test	2/10	0/30
SVM without screening	0/10	0/30
FNN-SWEEP with PART	0/10	3/30
FNN-SWEEP with S2N	0/10	0/30
FNN-SWEEP with *U*-test	0/10	0/30
FNN-SWEEP without screening	0/10	0/30
*k*NN with PART	0/10	0/30
*k*NN with S2N	0/10	0/30
*k*NN with *U*-test	0/10	0/30
*k*NN without screening	0/10	0/30
MRA with PART	0/10	0/30
MRA with S2N	0/10	0/30
MRA with *U*-test	0/10	0/30
MRA without screening	0/10	0/30
WV with PART	0/10	1/30
WV with S2N	0/10	2/30
WV with *U*-test	0/10	1/30
WV without screening	0/10	0/30

[a] Ten combination models from first to 10th models were constructed by PIM for each method in five-inputs. The accuracies of the models with first and second highest performance were 100% (=100 × 34/34) and 97.1% (=100 × 33/34), respectively. The number of the models with 100% or 97.1% accuracies were counted from 10 combination models.

[b] Ten combination models from first to 10th models were constructed by PIM for each method and each set (of three-fold cross-validation) in four-inputs. The accuracies of the models with first and second highest performance were 86.7% (=100 × 13/15) and 85.7% (=100 × 12/14), respectively. The number of the models with 86.7% or 85.7% accuracies, were counted from 30 combination models for three data sets.

3. Results and discussion

3.1. Comparison of the performance of PART-BFCS and the other methods

The performances of wrapper approaches with filter approaches as class predictors were investigated. For comparison, many combinations of various wrapper approaches, such as BFCS, SVM, FNN-SWEEP, *k*-nearest neighbor (*k*NN), multiple regression analysis (MRA), and weighted voting (WV), and various filtering approaches, such as *U*-test, S2N, PART, and no screening, were constructed. The performance of the predictors

Table 3
Comparison of cross-validation accuracies on various combination methods for CNS tumor data set (%)

	Inputs			
	1	2	3	4
BFCS with PART	65.6 ± 11.1	70.9 ± 14.3	74.5 ± 9.7	773[a] ± 8.8
BFCS with S2N	67.5 ± 11.5	67.4 ± 13.2	71.0 ± 9.7	71.1 ± 9.5
BFCS with *U*-test	67.5 ± 11.5	66.3 ± 12.8	71.2 ± 9.2	71.3 ± 9.1
BFCS without screening	67.5 ± 11.5	66.3 ± 12.8	71.2 ± 9.2	71.3 ± 9.1
SVM with PART	65.1 ± 14.9	65.6 ± 12.9	65.3 ± 12.9	65.8 ± 14.2
SVM with S2N	68.0 ± 11.6	66.8 ± 9.0	69.5 ± 8.1	68.3 ± 10.3
SVM with *U*-test	67.6 ± 11.9	66.3 ± 10.1	68.0 ± 8.8	65.7 ± 8.1
SVM without screening	67.6 ± 11.9	65.9 ± 9.6	68.2 ± 8.6	66.3 ± 10.0
FNN-SWEEP with PART	65.1 ± 11.3	66.5 ± 10.3	65.5 ± 12.6	62.2 ± 13.1
FNN-SWEEP with S2N	67.2 ± 12.6	62.9 ± 11.9	60.9 ± 10.4	59.1 ± 12.9
FNN-SWEEP with *U*-test	67.0 ± 12.6	62.5 ± 11.1	60.4 ± 10.1	59.1 ± 14.4
FNN-SWEEP without screening	67.0 ± 12.6	62.7 ± 10.6	60.3 ± 11.6	58.7 ± 11.9
*k*NN with PART	60.3 ± 11.6	59.3 ± 10.5	58.9 ± 12.2	59.8 ± 11.4
*k*NN with S2N	59.5 ± 11.9	57.2 ± 10.8	55.6 ± 10.9	55.0 ± 10.6
*k*NN with *U*-test	59.5 ± 10.9	58.6 ± 11.5	58.0 ± 9.9	57.1 ± 11.1
*k*NN without screening	58.0 ± 12.6	56.6 ± 11.7	57.5 ± 9.7	57.5 ± 9.0
MRA with PART	65.2 ± 11.2	64.2 ± 11.1	61.8 ± 14.6	55.2 ± 11.8
MRA with S2N	67.2 ± 11.9	63.3 ± 12.7	63.0 ± 10.9	56.9 ± 9.6
MRA with *U*-test	67.2 ± 11.9	61.8 ± 11.6	60.1 ± 10.4	55.1 ± 12.1
MRA without screening	67.2 ± 11.9	62.7 ± 11.3	59.1 ± 10.3	54.3 ± 13.6
WV with PART	61.7 ± 14.3	63.9 ± 12.9	60.9 ± 13.0	64.6 ± 12.0
WV with S2N	63.3 ± 13.8	63.3 ± 12.1	62.6 ± 12.1	63.1 ± 11.3
WV with *U*-test	66.1 ± 11.4	62.6 ± 9.3	62.3 ± 10.4	63.2 ± 10.1
WV without screening	66.1 ± 11.4	62.6 ± 11.4	63.0 ± 11.3	63.6 ± 9.6

The average blinded accuracies and their S.D.s were calculated from 10 combination models constructed by PIM.

[a] The highest accuracy.

was compared on the basis of the accuracy by using a blinded data set that was not used for modeling. By using 10 independent class predictor models, the average accuracy for blinded data set was calculated for the CNS and leukemia data sets.

The results of leukemia data are shown in Table 1. The result shows that average accuracy of the PART-BFCS models is the highest as shown in Table 1. In this experiment, top 10 independent class predictor models were constructed by PIM (parameter increasing method) [17] for each condition and data set. And the numbers of construction of high performance model (100% or 97.1% accuracy) were counted for each method as shown in Table 2. Four models among 10 models of five-input show 97.1% or more accuracy for PART-BFCS method. Next, the results for CNS data are shown in Table 3. The inputs used in the predictors were gradually increased from the one-input model to four-input model. As shown in Table 3, the PART-BFCS method clearly showed high performance when compared with the other methods in all input models with the exception of one-input model. The accuracy of the PART-BFCS method gradually increased and eventually, it reached 77.3% in the four-input models. On the other hand, SVM, FNN-SWEEP, *k*NN, MRA, and WV with various filtering showed an accuracy of 55.1–68.3%, which was lower than that of PART-BFCS. Average accuracy of three-input SVM models with S2N was the highest except BFCS models (69.5%). By using *U*-test, however, we found that the accuracy of BFCS with PART was significantly ($P = 5.94 \times 10^{-4}$) higher than one of SVM with S2N. In the four-input models, PART-BFCS method could constructed the most models that showed accuracies were 86.7% (first highest) or 85.7% (second highest), as shown in Table 2. These results could be explained by the facts that PART is the useful filtering method that could improve performances of simple models [5], BFCS is the modeling method in which the model is constructed by assembling simple models, such as one-input FNN. Otherwise, complex models are constructed by other modeling methods. Table 2 shows that the most high performance models were constructed by PART-BFCS method. Therefore, combination of PART and BFCS is the best one.

3.2. Evaluation of prediction results using improved RI_{BFCS}

PART-BFCS method can estimate assurance of results by calculating reliability index for BFCS ($\mathrm{RI}_{\mathrm{BFCS}}$). In the present study, we propose improved $\mathrm{RI}_{\mathrm{BFCS}}$ (new $\mathrm{RI}_{\mathrm{BFCS}}$) by modifying equation of $\mathrm{RI}_{\mathrm{BFCS}}$ (old $\mathrm{RI}_{\mathrm{BFCS}}$) for more practical cancer diagnosis. For acute leukemia and CNS data, both $\mathrm{RI}_{\mathrm{BFCS}}$ of each patient in blinded data were calculated (Fig. 3). Fig. 3 shows distributions of correct and incorrect sample for old and new $\mathrm{RI}_{\mathrm{BFCS}}$. It is necessary that there are many incorrect samples in low $\mathrm{RI}_{\mathrm{BFCS}}$ and many correct samples in high $\mathrm{RI}_{\mathrm{BFCS}}$. For old $\mathrm{RI}_{\mathrm{BFCS}}$, two distributions are not separated ($P = 0.169$, 0.311), as shown in Fig. 3A and B. On the other hand, they

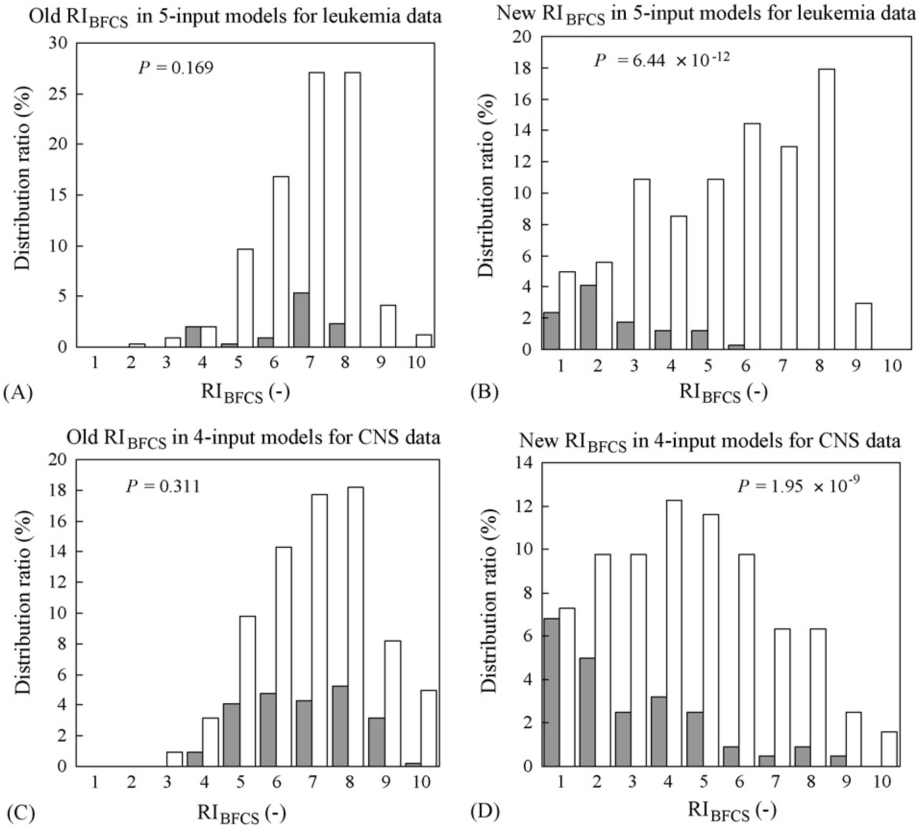


Fig. 3. Comparison of old and new $\mathrm{RI}_{\mathrm{BFCS}}$. White bars and gray bars indicate the distribution ratio of correct sample and incorrect samples, respectively. The *P*-values were calculated by Mann–Whitney test and indicate the difference in $\mathrm{RI}_{\mathrm{BFCS}}$ distribution between the correct and incorrect samples.

are clearly separated ($P = 6.44 \times 10^{-12}$, 1.95×10^{-9}) for new $\mathrm{RI}_{\mathrm{BFCS}}$, as shown in Fig. 3B and D. Based on this new index, the discriminated group with over 90% prediction accuracy was separated from the others. For example, the patients who had new $\mathrm{RI}_{\mathrm{BFCS}} > 5$ corresponded to 48.5% of all patients for leukemia data, and an accuracy of 99.4% was achieved. And, the patients who had new $\mathrm{RI}_{\mathrm{BFCS}} > 5$ corresponded to 29.3% of all patients for CNS data, and a accuracy of 90.7% was achieved. This result implies new $\mathrm{RI}_{\mathrm{BFCS}}$ more practical than old one. Old RI is mean distance from boundary line for each gene. BFCS is one of voting methods by assembling simple methods. Improved RI is modified by adding each signs of simple models in the BFCS model. Thus, improved RI is superior to old RI.

3.3. Comparison of selected genes with known prognostic marker genes

We investigated the presence of previously reported prognostic marker genes among the genes selected in the 10 constructed combinations of four-input PART-BFCS models. There were total of 40 genes in 10 models. Some genes were selected several times. In the case of PART-BFCS, 14 genes among 40 genes are independent, as shown in Table 4. Three genes among these 14 genes were reported to be prognostic markers for cancer: The *CCND1* gene was reported by Tan et al. [20] to be a high-risk marker gene. *CCND1* plays an important role in regulating the progress of the cell division during the G1 phase of the cell cycle. Overexpression of *CCND1* correlates with sensitivity to cisplatin [21]. The *LIF* gene was reported by Park et al. [22] to be a low-risk marker gene. *LIF* induces growth arrest and differentiation of cells. The *USP4* (*UNPH*) gene was reported by Frederick et al. [23] to be a low-risk marker gene. These observations accurately matched with low or high gene expression of the above-mentioned three marker genes, as shown in Table 4. These findings suggest that the PART-BFCS method may be used to identify new marker genes.

3.4. Comparison of genes used in PART-BFCS predictors and other predictors for CNS data

We firstly compared FNN-SWEEP and BFCS to investigate numerical character of the genes selected by PART-BFCS. Both FNN-SWEEP and BFCS are based on FNN. The one-gene predictors were constructed for each gene from second input to fourth input in the two methods. And then, average modeling accuracy of one-gene predictors for 10 combinations, was calculated (Table 5). The BFCS genes used as one-gene predictors showed clearly higher accuracy than FNN-SWEEP ones, as shown in Table 5. The average modeling accuracies of the genes from second to fourth were 83.3%, 77.3% and 79.0% for BFCS, and 72.0%, 68.7% and 65.7% for FNN-SWEEP, respectively. The PIM method was used in the FNN-SWEEP. This method is

Table 4
The genes used in PART-BFCS class predictor for one set of CNS tumor data set

Accession number	Gene name	Descriptions	Times of selection	Average intensity of surviving patients[a]	Average intensity of dead patients[b]	Threshold of model[c]	Prognostic markers
U20657	USP4	Ubiquitin specific protease 4 (proto-oncogene)	9	796	127	391	[d]
X59798	CCND1	Cyclin Dl (PRAD1: parathyroid adenomatosis 1)	6	0	2176	330	[e]
M73547	C5orfl8	Chromosome 5 open reading frame 18	6	1439	275	797	
AB000460	C4orf8	Chromosome 4 open reading frame 8	4	2429	1094	1605	
L33243	PKD1	Polycystic kidney disease 1 (autosomal dominant)	4	1498	228	815	
X13967	LIF	Leukemia inhibitory factor (cholinergic differentiation factor)	2	464	5	206	[f]
L10333	RTN1	Reticulon 1	2	4483	821	1747	
D30756	M17S2	Membrane component, chromosome 17, surface marker 2	1	591	59	236	
D83018	NELL2	NEL-like 2 (chicken)	1	2710	851	1416	
HG2238-HT2321	NUMA1	Nuclear mitotic apparatus protein l, alt. splice form 2	1	2833	1197	1721	
J04046	CALM3	Calmodulin 3 (phosphorylase kinase, delta)	1	3287	1022	1753	
S76475	NTRK3	Neurotrophic tyrosine kinase, receptor, type 3 (TrkC)	1	2002	96	687	
U25849	ACPI	Acid phosphatase 1, soluble	1	206	1082	602	
Y09616	CES2	Carboxylesterase 2 (intestine, liver)	1	2894	1065	1706	

[a] The average intensity of gene expression in the patients predicted as survivors.
[b] The average intensity of gene expression in the patients predicted as dead.
[c] The threshold of gene expression in the weak learner model.
[d] The marker gene reported by Frederick et al. [23] as a low-risk marker.
[e] The marker gene reported by Tan et al. [20] as a high-risk marker.
[f] The marker gene reported by Park et al. [22] as a low-risk marker.

very useful to select input combination that shows high accuracy by combining low accuracy inputs. But, the application of PIM to high dimensional data, such as microarray data, may cause overfitting. On the other hand, the boosting used in the BFCS is the method that can construct high-accuracy predictor by combining one-gene predictors. Thus, low-accuracy one-gene predictors are hardly selected. It may be for this reason that BFCS showed high performance.

Next, BFCS were compared with PART-BFCS. Average of gene expression for the 40 genes in 10 combinations of four-input models, was calculated for each class (survivors or dead). And average standard deviation (S.D.) of lower gene expression class for 40 genes is shown in Table 6. Table 6 shows that the S.D. of PART-BFCS was lower than one of BFCS. The values of S.D.s were 0.57 for BFCS and 0.39 for PART-BFCS. This tendency is corresponding to the fact previously reported by us [5]. The genes with low S.D. in lower class may show high generalization performance.

3.5. IF-THEN rules extracted from PART-BFCS model

After modeling, the IF-THEN rules for CNS prognosis were obtained from the model of highest blind accuracy among the 10 combinations. The model includes the *CCND1* gene and *USP4* (*UNPH*) gene as known prognostic markers. The IF-THEN rules have been obtained as a matrix that is classified based on expression level of such selected genes (Fig. 4). Using this matrix, simple and precise rules were obtained as follows. The simplest rule is that patients with high expression of *CCND1* gene are likely to exhibit poor prognosis. Six patients showed high expression of *CCND1* gene and five of them were actually dead patients, which corresponds to 28% (5/18) of all the dead

Table 5
Average modeling accuracy of one-input models between BFCS and FNN-SWEEP (%)

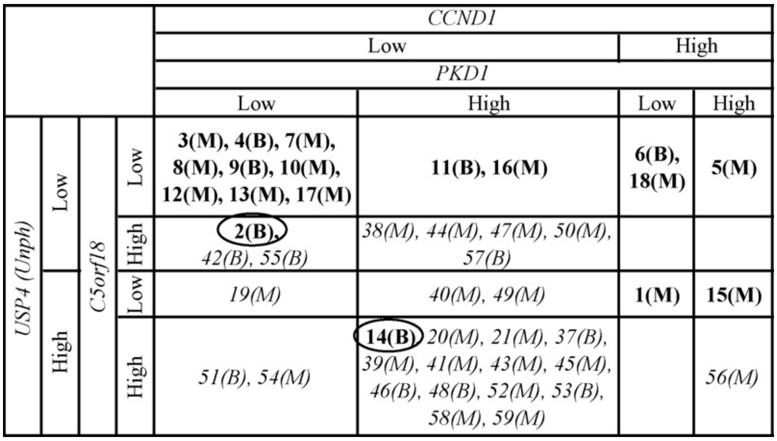
	Order of selection		
	Second	Third	Fourth
BFCS	83.3	77.3	79.0
FNN-SWEEP	72.0	68.7	65.7

In this experiment, 10 combinations of four-input models were constructed. Three genes from second to fourth in four-input were selected as combination of genes for each method. The modeling accuracies when these three genes were used alone as one-gene predictors, were calculated for 10 combinations.

Table 6
Average S.D. of gene expression in lower class between BFCS and PART-BFCS

Methods	The S.D.s of lower class
BFCS	0.56
PART-BFCS	0.39

Average of gene expression for the 40 genes in 10 combinations of four-input models were calculated for each class (survivors or dead). And then, average standard deviation (S.D.) of lower gene expression class for 40 genes was calculated.



				CCND1			
				Low		High	
				PKD1			
				Low	High	Low	High
USP4 (Unph)	Low	*C5orf18*	Low	**3(M), 4(B), 7(M), 8(M), 9(B), 10(M), 12(M), 13(M), 17(M)**	**11(B), 16(M)**	**6(B), 18(M)**	**5(M)**
			High	**2(B)**, *42(B), 55(B)*	*38(M), 44(M), 47(M), 50(M), 57(B)*		
	High		Low	*19(M)*	*40(M), 49(M)*	**1(M)**	**15(M)**
			High	*51(B), 54(M)*	**14(B)** *20(M), 21(M), 37(B), 39(M), 41(M), 43(M), 45(M), 46(B), 48(B), 52(M), 53(B), 58(M), 59(M)*		*56(M)*

Fig. 4. IF-THEN rules in the top two model of PART-BFCS. Since the expression level of each gene can be divided into either high or low groups using fuzzy logic, this model comprised 16 (=2^4) fuzzy rules. Numbers in each matrix are identical to the patient numbers previously described by Pomeroy et al. [7]. Numbers in bold type and italic type indicate the poor and good prognosis patients, respectively. Patient numbers are placed in the matrix according to the expression levels of each patient. Patient numbers in the circle represent incorrect classification by the PART-BFCS. (B) indicates sample in blinded data set. (M) indicates sample in modeling data set.

patients. Next simple rule is that patients with low expression of *CCND1* gene and low expression of *USP4* (*UNPH*) gene are likely to exhibit poor prognosis. Nineteen patients showed low expression of *USP4* (*UNPH*) gene and 12 of them were actually dead patients, which corresponds to 92% (12/13) of dead patients showing low expression of the *CCND1* gene. Nineteen patients showed high expression of *USP4* (*UNPH*) gene and low expression of *CCND1* gene, and 18 of them (95%) were actually surviving patients, which corresponds to 69% (18/26) of all the surviving patients. It was found that surviving or dead patients were clustered at specific parts of the matrix. The following rule was also found: patients were likely to exhibit a poor prognosis when the *USP4* (*UNPH*) expression was low and *C5orf18* expression was low. It was also found that on this matrix, two patients showing poor prognosis were incorrectly predicted as showing good prognosis. This may be due to the inability for complete removal of their tumors by CNS surgery.

4. Conclusions

In the present study, we investigated combinations of various filter and wrapper approaches, and found that combination method of PART and BFCS (a kind of boosting) is significantly superior to other methods with regard to high prediction accuracy for construction of class predictor from gene expression data. This method could select some marker genes related to cancer outcome. In addition, we proposed improved $\mathrm{RI}_{\mathrm{BFCS}}$ of PART-BFCS. Based on this new index, the discriminated group with over 90% prediction accuracy was separated from the others. It is necessary that there are about 90% or more prediction accuracy in the practical diagnosis application. These results suggest that the PART-BFCS method has a high potential to function as a new method of marker gene selection for the diagnosis of patients, using high dimensional data such as DNA microarray, mass spectrometry (MS), and two-dimensional polyacrylamide gel electrophoresis (2D-PAGE).

Acknowledgment

This work was supported by the Hori Information Science Promotion Foundation for financial support.

References

[1] T. Ando, M. Suguro, T. Hanai, T. Kobayashi, H. Honda, M. Seto, Fuzzy neural network applied to gene expression profiling for predicting the prognosis of diffuse large B-cell lymphoma, Jpn. J. Cancer Res. 93 (2002) 1207–1212.

[2] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (2002) 389–422.

[3] H. Takahashi, H. Honda, A new reliable cancer diagnosis method using boosted fuzzy classifier with SWEEP operator method, J. Chem. Eng. Jpn. 38 (2005) 763–773.

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

[5] H. Takahashi, T. Kobayashi, H. Honda, Construction of robust prognostic predictors by using projective adaptive resonance theory as a gene filtering method, Bioinformatics 21 (2005) 179–186.

[6] D.J. Park, P.T. Vuong, S. de Vos, D. Douter, H.P. Koeffler, Comparative analysis of genes regulated by PML/RAR alpha and PLZF/RAR alpha in response to retinoic acid using oligonucleotide arrays, Blood 102 (2003) 3727–3736.

[7] S.L. Pomeroy, P. Tamayo, M. Gaasenbeek, L.M. Sturla, M. Angelo, M.E. McLaughlin, J.Y. Kim, L.C. Goumnerova, P.M. Black, C. Lau, J.C. Allen, D. Zagzag, J.M. Olson, T. Curran, C. Wetmore, J.A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D.N. Louis, J.P. Mesirov, E.S. Lander, T.R. Golub, Prediction of central nervous system embryonal tumour outcome based on gene expression, Nature 415 (2002) 436–442.

[8] R.E. Schapire, The strength of weak learnability, Mach. Learn. 5 (1990) 197–227.

[9] Y. Freund, R. Schapire, A decision-theoretic generalization of online learning and an application to boosting, J. Comput. Syst. Sci. 55 (1997) 119–139.

[10] Y. Freund, Adaptive version of the boost by majority algorithm, Mach. Learn. 43 (2000) 293–318.

[11] J. Friedman, T. Hastie, R. Tibshirani, Additive logistic regression: a statistical view of boosting, Ann. Stat. 28 (2000) 337–407.

[12] S. Horikawa, T. Furuhashi, Y. Uchikawa, On fuzzy modeling using fuzzy neural networks with the back-propagation algorithm, IEEE Trans. Neural Networ. 3 (1992) 801–806.

[13] Y. Huang, Y. Li, Prediction of protein subcellular locations using fuzzy k-NN method, Bioinformatics 20 (2004) 21–28.

[14] V.N. Vapnik, A. Chervonenkis, A note on one class of perceptrons, Automat. Rem. Control 25 (1964) 821–837.

[15] T. Joachims, Making Large-scale SVM Learning Practical, MIT Press, Cambridge, 1999.

[16] H. Noguchi, T. Hanai, W. Takahashi, T. Ichii, M. Tanikawa, S. Masuoka, H. Honda, T. Kobayashi, Model construction for quality of beer and brewing process using FNN, Kagaku Kougaku Ronbunshu 25 (1999) 695–701.

[17] H. Noguchi, T. Hanai, H. Honda, L.C. Harrison, T. Kobayashi, Fuzzy neural network-based prediction of the motif for MHC class II binding peptides, J. Biosci. Bioeng. 92 (2001) 227–231.

[18] Y. Cao, J. Wu, Projective ART for clustering data sets in high dimensional spaces, Neural Networ. 15 (2002) 105–120.

[19] Y. Cao, J. Wu, Dynamics of projective adaptive resonance theory model: the foundation of PART algorithm, IEEE Trans. Neural Networ. 15 (2004) 245–260.

[20] P.G. Tan, Z. Xing, Z.Q. Li, Expression of cyclin D1 in brain gliomas and its significance, Ai Zheng 23 (2004) 63–65.

[21] J. Akervall, D.M. Kurnit, M. Adams, S. Zhu, S.G. Fisher, C.R. Bradford, T.E. Carey, Overexpression of cyclin D1 correlates with sensitivity to cisplatin in squamous cell carcinoma cell lines of the head and neck, Acta Otolaryngol. 124 (2004) 851–857.

[22] J.I. Park, C.J. Strock, D.W. Ball, B.D. Nelkin, The Ras/Raf/MEK/extracellular signal-regulated kinase pathway induces autocrine–paracrine growth inhibition via the leukemia inhibitory factor/JAK/STAT pathway, Mol. Cell. Biol. 23 (2003) 543–554.

[23] A. Frederick, M. Rolfe, M.I. Chiu, The human UNP locus at 3p21.31 encodes two tissue-selective, cytoplasmic isoforms with deubiquitinating activity that have reduced expression in small cell lung carcinoma cell lines, Oncogene 16 (1998) 153–165.