PERGAMON

Computers in Biology and Medicine 32 (2002) 237–246

Computers in Biology and Medicine

www.elsevier.com/locate/compbiomed

Combining neural network predictions for medical diagnosis

Yoichi Hayashi[a], Rudy Setiono[b,*]

[a] *Department of Computer Science, Meiji University, 1-1-1 Higashimita, Tama-ku, Kawasaki 214-8571, Japan*
[b] *School of Computing, National University of Singapore, 3 Science Drive 2, Singapore 117543, Singapore*

Received 23 May 2001; accepted 21 November 2001

Abstract

We present our results from combining the predictions of an ensemble of neural networks for the diagnosis of hepatobiliary disorders. To improve the accuracy of the diagnosis, we train the second level networks using the outputs of the first level networks as input data. The second level networks achieve an accuracy that is higher than that of the individual networks in the first level. Compared to the simple method which averages the outputs of the first level networks, the second level networks are also more accurate. We discuss how the overall predictive accuracy can be improved by introducing bias during the training of the level one networks.

Keywords: Feedforward neural network; Neural network ensemble; Simple averaging; Medical diagnosis; Hepatobiliary disorder

1. Introduction

Many authors have shown that combining the predictions of several models often results in a prediction accuracy that is higher than that of the individual models. The general framework for predicting using an ensemble of models consists of two levels and is often referred to as *stacked generalization* [1]. In the first level, various learning methods are used to learn different models from the original data set. The predictions of the models from the first level along with the corresponding target class of the original input data are then used as inputs to learn a second level model.

As neural networks are among the most popular models for pattern classification, numerous papers that report on theoretical and experimental results on combining the neural network predictions can be found in the literature. Among the second level models proposed for combining the

* Corresponding author. Tel.: +65-874-6297; fax: +65-779-4580.
E-mail addresses: hayashiy@cs.meiji.ac.jp (Y. Hayashi), rudys@comp.nus.edu.sg (R. Setiono).

0010-4825/02/$ - see front matter
PII: S0010-4825(02)00006-9

network predictions are the simple averaging method and the generalized ensemble method [2] and the weighted least-squares method [3]. In these methods, the second level model is a simple weighted predictions of the component networks in the first level. The three methods differ in their computation of the ensemble weights given to the component networks in the ensemble. While the ensemble weights for the simple averaging method are equal for all component networks, the generalized ensemble weights depend on the correlation matrix of the errors of the component networks. In the weighted least-squares method, the weights are computed as the product of the component networks' outputs and the target vector of the training samples.

The accuracy of the different methods for combining regularized neural networks have been compared on a breast cancer database [4]. The regularized neural networks investigated are networks that have been trained to minimize a cost function involving the sum of squared error function and a quadratic penalty term of the network weights. The first level models are neural networks that have been initialized with different random initial weights and neural networks that have been trained using different subsets of the data. The different data subsets are obtained by randomly drawing samples from the original data set with replacement. The second level models used include simple averaging method and a variance-based weighting method of the first level neural networks.

Another application of the network ensemble approach for the diagnosis of breast cancer has also been reported recently [5]. The network ensemble is adapted so that it is less likely to make false positive diagnosis (malignant diagnosis for benign data). The adaptation is achieved through training neural networks using different proportions of malignant to benign data. The first level models are two groups of neural networks. The networks in the first group have been trained with greater proportion of benign samples, while those in the second group with greater proportion of malignant samples. The second level model is a threshold decision mechanism which, based on a certain empirically determined threshold, decides whether the output of the first group or the second group is to be taken as the final output.

In this paper, we present our experimental results on combining neural network predictions for the diagnosis of hepatobiliary disorders. The data have been collected from a total of 536 patients who were admitted to a university-affiliated hospital in Japan. Nine real-valued measurements from biomedical tests were obtained from these patients. The hepatobiliary disorders alcoholic liver damage (ALD), primary hepatoma (PH), liver cirrhosis (LC), and cholelithiasis (C) constitute the four output classes. Because there are four possible outcomes of a diagnosis, for the first level models we have used four sets of neural networks. Networks in each set have been trained so that they are likely to be more accurate for one type of disorder than the other three disorders. The predictions of the networks in the first level are combined by a second level neural network. We have been able to achieve significant improvement in accuracy by applying neural networks as the second level model compared to the simple averaging method.

The outline of this paper is as follows. In Section 2, we describe the data that have been collected in more detail. We also describe the neural network topology used in this section. In Section 3, we present the results of our experiments using neural network for combining the predictions of the first level networks. Finally, in Section 4 we conclude the paper.

2. Diagnosis of hepatobiliary disorders using neural networks

2.1. The data set

The hepatobiliary disorder data set contains 536 samples with nine input attributes. The attributes correspond to measurements from biomedical tests. They are glutamic oxaloacetic transaminase (GOT [1]), glutamic pyruvic transaminase (GPT [2]), lactate dehydrase (LDH), gamma glutamyl transpeptidase (GGT), blood urea nitrogen (BUN), mean corpuscular volume of red blood cells (MCV), mean corpuscular hemoglobin (MCH), total billirubin (TBil) and creatinine (CRTNN). Table 1 lists the nine input attributes along with their unit measurements, minimum values, mean values, and maximum values.

The patients had been clinically and pathologically diagnosed by physicians at a university-affiliated hospital in Japan and each was diagnosed as suffering from one of the four hepatobiliary disorders: ALD, PH, LC and C. In our previous experiments using this data set [6,7], the samples had been randomly split into a training data set containing 373 samples and a test data set containing the remaining 163 samples. The class distribution of the samples in the training and test data sets is summarized in Table 2.

2.2. The neural networks

We train 30 neural networks to find out what kind of accuracy level can be achieved by these networks. The network topology is the standard feedforward network with a single hidden layer. Each network has 10 input units, nine units for the nine attributes of the data and one unit for the hidden unit bias. The input value of the 10th unit is fixed at one for all samples. The number of hidden units is 12, and the number of output units is 4. Samples with target outputs ALD, PH, LC, and C are given the binary target values of $(1,0,0,0)$, $(0,1,0,0)$, $(0,0,1,0)$, and $(0,0,0,1)$, respectively.

Table 1
The nine measurements of the hepatobiliary data set

Measurement	Unit	Min. value	Mean value	Max. value
GOT	Karmen unit	8.0	113.0	4356.0
GPT	Karmen unit	3.0	54.5	1124.0C
LDH	iu/l	179.0	476.3	6327.0
GGT	μ/ml	4.0	144.1	3075.0
BUN	mg/dl	3.3	17.2	91.0
MCV	fl	66.7	96.1	160.5
MCH	pg	20.3	32.1	52.5
TBil	mg/dl	0.1	3.2	37.0
CRTNN	mg/dl	0.4	1.1	4.3

[1] Also known as aspartate aminotransferase (AST).
[2] Also known as alanine aminotransferase (ALT).

Table 2
Class distribution of the samples in the training and the test data sets

Class	Training set	Test set
ALD	83	33
PH	127	51
LC	89	35
C	74	44
Total	373	163

Before the network training starts, all the input attributes are normalized so that they range in the interval $[0, 1]$. Two criteria for measuring the accuracy rates have been used in previous studies on this data set [6,8]. Using the *best choice* criterion, a sample is correctly classified if the network unit with the largest output corresponds to the position of one in the actual target value. Using the *second best choice*, we also consider the network output unit having the second largest output as more than one of the disorders can occur together in the same patient.

The networks are trained to minimize the sum of squared errors:

$$E(\mathbf{w}, \mathbf{v}) = \sum_{i=i}^{P} \sum_{j=1}^{4} (S_i^j - t_i^j)^2, \tag{1}$$

where P is the number of training samples, $\mathbf{w}$ and $\mathbf{v}$ are the network weights from the input units to the hidden units, and from the hidden units to the output units, respectively. The target value for sample $\mathbf{x_i}$ at output unit j is t_i^j and the network output S_i^j is computed as

$$S_i^j = \sum_{k=1}^{12} A_i^k v_k^j, \tag{2}$$

where v_k^j is the weight of the network connection from hidden unit k to output unit j. The hidden unit k's activation for input sample $\mathbf{x_i}$ is

$$A_i^k = \sigma\left(\sum_{\ell=1}^{10} x_i^\ell w_\ell^k \right), \tag{3}$$

where x_i^ℓ is the ℓth component of $\mathbf{x_i}$ and w_ℓ^k is the weight of the network connection from input unit ℓ to hidden unit k. The hidden unit activation function σ is the hyperbolic tangent function:

$$\sigma(\xi) = \frac{1 - \mathrm{e}^{-\xi}}{1 + \mathrm{e}^{-\xi}}. \tag{4}$$

The average accuracy rates of the 30 networks on the training and test data sets are summarized in Table 3. Using the best choice and second best choice criteria the average predictive accuracy rates of the 30 neural networks are 72.35% and 90.27%, respectively. These rates are higher than the rates obtained by neural networks that have been trained to minimize an augmented error function and have few hidden units and connections after being pruned [7]. The average accuracy rates of

Table 3
The average training and predictive accuracy of neural 30 networks

Class	Best choice	Second best choice
Training set		
ALD	64.13/83 (77.27%)	77.47/83 (93.34%)
PH	115.57/127 (91.00%)	126.00/127 (99.21%)
LC	57.30/89 (64.38%)	76.37/89 (85.81%)
C	65.40/74 (88.38%)	70.73/74 (95.58%)
Overall	302.40/373 (81.07%)	350.57/373 (93.99%)
Test set		
ALD	22.53/33 (68.27%)	28.87/33 (87.48%)
PH	40.00/51 (78.43%)	47.40/51 (92.94%)
LC	19.43/35 (55.51%)	29.17/35 (83.34%)
C	35.97/44 (81.75%)	41.70/44 (94.77%)
Overall	117.93/163 (72.35%)	147.14/163 (90.27%)

the pruned networks are 54.44% and 84.64%, respectively. In Section 3, we show how the accuracy of the networks can be improved by combining the network predictions.

3. Combining network predictions

3.1. Simple averaging

Simple averaging of the predictions have been known to improve the performance of the individual predictions.

Table 4 shows the accuracy obtained by averaging the predictions from N networks, where N is 5, 10, or 15. The accuracy rates are averaged over five groups of randomly selected N networks from the 30 networks that we have trained. From the figures in this table, we see that there is a 1% increase in predictive accuracy over the average accuracy of the individual networks. When we also consider the second best choice, the improvement in the predictive accuracy is around 1.5%. The figures also indicate that the highest test set accuracy rates using the best choice criterion are obtained by averaging the predictions of $N = 5$ networks, while using the second best choice the best rates are obtained by the groups of $N = 15$ networks.

3.2. Averaging biased neural networks

The accuracy of the ensemble can be expected to be better than the accuracy of the individual networks if the component networks which make up the ensemble differ in their predictions. One way of creating networks with different error patterns is by introducing bias during learning. By adjusting the proportions of the different classes of training samples, the networks can be trained

Table 4
The accuracy from averaging the outputs of N neural networks

N	Training set		Test set	
	Best choice (%)	Second best choice (%)	Best choice (%)	Second best choice (%)
5	82.09	95.23	73.62	90.92
10	82.58	95.04	73.49	91.78
15	82.04	95.08	73.25	91.92

Table 5
The average accuracy of 10 networks trained with target $(0, 0, 0, 2.5)$ for samples of class C

Class	Training set		Test set	
	Best choice (%)	Second best choice (%)	Best choice (%)	Second best choice (%)
ALD	63.25	89.88	50.61	80.00
PH	84.09	96.93	69.41	90.20
LC	58.31	79.44	49.14	77.14
C	97.43	98.92	95.00	97.05
Overall	75.95	91.58	68.16	87.18

so that they are biased towards a certain class. In a two-class problem, this approach was used to reduce the number of false positive diagnosis of breast cancer patients [5].

Instead of changing the composition of the training samples, in our experiments we modify the target of a certain class of samples. For example, instead of having target values of $(0, 0, 0, 1)$ for samples from C patients, we change this target to $(0, 0, 0, \alpha)$ for some $\alpha > 1$. By modifying the target for samples of class C in this manner, the prediction of the networks will be biased towards C. More samples of this class will be correctly classified. The overall predictive accuracy of the networks, however, may be lower as samples from other classes are now also more likely to be classified as C. The figures in Table 5 show the average accuracy from 10 neural networks that have been trained with target values of $(0, 0, 0, 2.5)$ for all samples of class C. The target values for the other three classes are unchanged. As expected, these networks predict samples in class C both in the training set and in the test set with greater accuracy compared to the individual original networks (Table 3).

In order to investigate the effect of having biased neural networks on the overall predictive accuracy, we train neural networks in groups of 10. All 10 networks in a group are trained to be biased towards one class by multiplying the target value for samples in this class by α. These networks have the same topology, 10 input units, 12 hidden units and four output units, but they are initialized with different random initial weights. From each of the four groups of networks, M are selected randomly. The average predictions from $4 \times M$ networks are computed and their accuracy recorded for $M = 3$, 5 and 10. The accuracy rates are summarized in Table 6 for different values

Table 6
Predictive accuracy obtained by averaging the predictions of biased networks

α	$M = 3$ (%)	$M = 5$ (%)	$M = 10$ (%)
1.5	76.12, 92.64	77.30, 92.02	74.85, 92.02
2.0	76.07, 92.64	76.69, 92.64	77.30, 93.25
2.5	78.53, 92.64	79.75, 93.25	78.53, 94.48

The accuracy shown are for the best choice and the second best choice criteria.

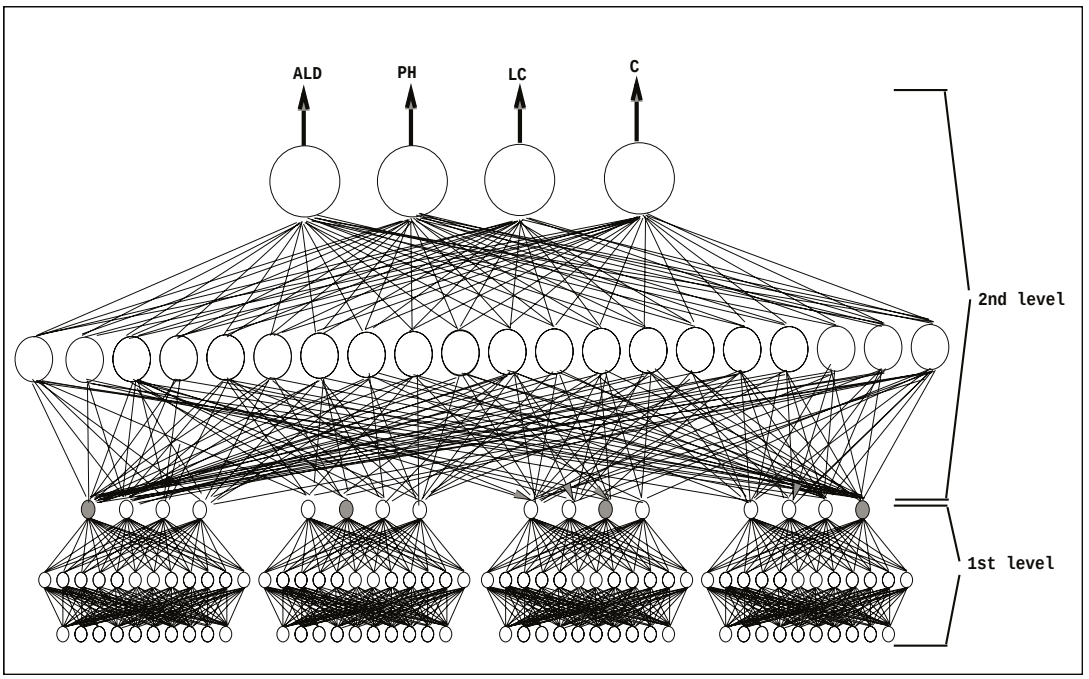

Fig. 1. A second level network is used to combine the predictions of the first level neural networks.

of α. Comparing the figures in this table to those in Table 4, we can see some clear improvement in the predictive accuracy. When there is no bias introduced during the training of the component networks, the predictive accuracy rates are no more than 74% and 92% using the best choice and second best choice criteria, respectively (Table 4). With bias introduced in the networks, the accuracy is as high as 79.75% using the best choice criterion and 94.48% using the second best criterion.

3.3. Using neural networks as second level model

Instead of simply taking the average of the predictions of the individual networks, a more sophisticated model that makes use of these predictions can be expected to give an even better improvement in accuracy. We have trained second level neural networks to combine the predictions of the first level networks (Fig. 1). A second level network has 16 input units which correspond to the outputs

Table 7
Predictive accuracy obtained by applying neural networks as second level model

α	$M = 3$ (%)	$M = 5$ (%)	$M = 10$ (%)
1.5	77.30, 87.73	77.91, 84.66	78.53, 88.34
2.0	80.98, 90.18	78.53, 87.12	80.37, 89.57
2.5	78.53, 90.18	83.47, 91.41	80.37, 89.57

The accuracy shown are for the best choice and the second best choice criteria.

of the four groups of the first level biased networks as described in Section 3.2. The number of output unit is four, and the number of hidden units is chosen to be 20. The training samples for the second level networks are generated as follows. For each original nine-dimensional training sample, the average output of M networks from each of the four groups are computed. This average output is a new 16-dimensional training sample for the second level networks. The target for this new sample is the same as the class target of the original sample. The accuracy achieved by the second level networks for the different values of α and M are given in Table 7. The accuracy on the test set using the best choice criterion achieved by the two-level model is as high as 83.47%, when the inputs of the second level networks are averages of $M = 5$ biased neural networks at the first level that have been trained with $\alpha = 2.5$.

3.4. Comparison with other models

Methods that have been used to analyze the hepatobiliary disorder data set include the fuzzy neural network model [9] and fuzzy multilayer perceptron (FMLP) [8]. The fuzzy neural network model of Hayashi et al. is similar to the traditional backpropagation neural networks except for its use of fuzzy numbers and fuzzy arithmetic as the input data to the model and the means to train the network, respectively. The fuzzy MLP, originally proposed by Pal and Mitra [10] also has the standard MLP as its backbone. Fuzziness is incorporated at the input and the output of the MLP. The method is capable of handling inexact or linguistic data. The inputs to the MLP are combinations of membership values in the set low, medium, and high. The output units of the MLP represent class membership values of the samples. Both Hayashi's fuzzy neural network model and the fuzzy MLP of Pal and Mitra update the weights of the network by backpropagating the errors.

The predictive accuracy rates of the fuzzy neural network (FNN) model are 56.4% and 77.3%, using the best choice and second best choice criteria, respectively. The corresponding figures for the FMLP are 76.0% and 88.9%. In comparison, the two-level neural network (2-LNN) model that we propose here achieves accuracy rates that is as high as 83.47% using the best choice criterion and 91.41% using the second best choice criterion. The highest accuracy rate using the second best choice criterion obtained by averaging the network predictions is 94.48%. It is obtained by combining the outputs of four groups of 10 biased neural networks (BNN). Using the best choice criterion, the highest accuracy rate obtained by averaging biased neural networks is 79.75%. The comparison of the performance of the various methods is summarized in Fig. 2.

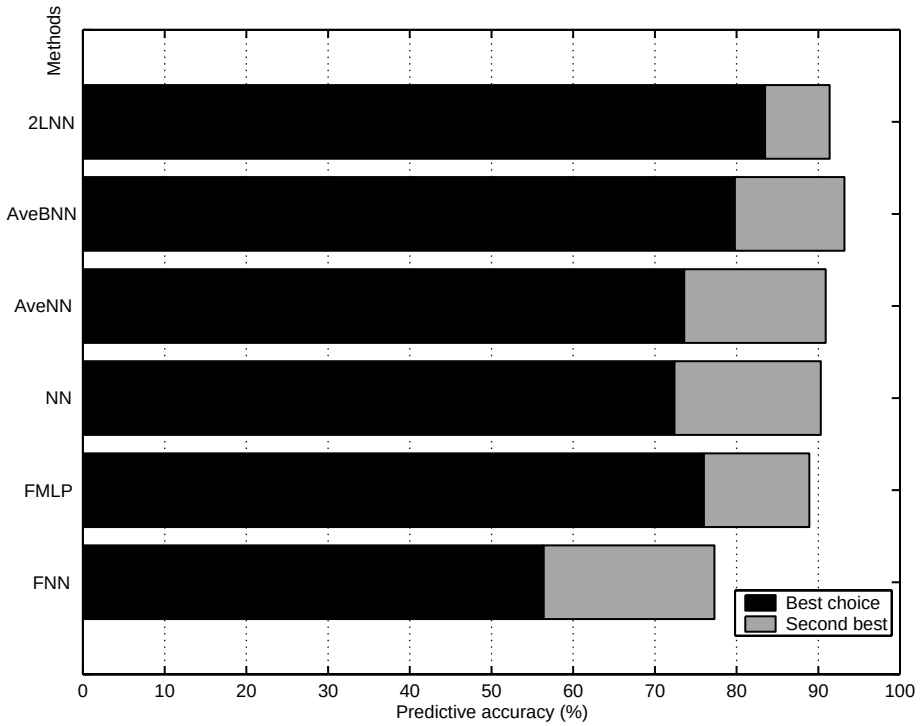

Fig. 2. Comparison of the accuracy rates of various methods on the hepatobiliary disorders data set. FNN is the accuracy of fuzzy neural network, FMLP is the best accuracy of fuzzy multilayer perceptron, NN is the average accuracy of neural networks, AveNN is the best accuracy from averaging NN predictions, AveBNN is the best accuracy from averaging biased NN predictions, 2LNN is the best accuracy from the two-level neural network method.

4. Summary

We proposed the use of neural networks to combine the predictions of a neural network ensemble that has been trained for diagnosing hepatobiliary disorders. In order to generate networks with differing error patterns we generated biased networks by training the networks in four separate groups. Networks in each group were trained with different targets. The learning targets were modified so that the trained networks would predict one particular disorder with higher accuracy than the other three types of disorders. Averaging the predictions of these biased networks resulted in an improvement in accuracy over the predictions of the individual networks and the predictions obtained by averaging neural networks with no bias introduced during their training. Further improvement in accuracy was obtained by training new neural networks to combine the predictions of the original networks. The accuracy rates achieved by the two-level neural network model are higher than the other methods that have been applied to the same hepatobiliary data set such as fuzzy multilayer perceptrons with many hidden layers and nodes.

References

[1] D.H. Wolpert, Stacked generalization, Neural Networks 5 (1992) 241–259.
[2] M.P. Perrone, L.N. Cooper, When networks disagree: ensemble methods for hybrid neural networks, in: R.J. Mammone (Ed.), Artificial Neural Networks for Speech and Vision, Chapman & Hall, London, 1993, pp. 126–142.
[3] L.-W. Chan, Weighted least square ensemble networks, Proceedings of the IJCNN'99, 1999, pp. 1393–1396.
[4] M. Taniguchi, V. Tresp, Combining regularized neural networks, in: W. Gerstner, A. Germond, M. Hasler, J. Nicoud (Eds.), Proceedings of the ICANN'97, Lecture Notes in Computer Science 1327, Springer, Berlin, 1997, pp. 349–354.
[5] A.J.C. Sharkey, N.E. Sharkey, S.S. Cross, Adapting an ensemble approach for the diagnosis of breast cancer, in: L. Niklasson, M. Boden, T. Ziemke (Eds.), Proceedings of the ICANN'98, Springer, London, 1998, pp. 281–286.
[6] Y. Hayashi, Neural expert system using fuzzy teaching input and its application to medical diagnosis, Inf. Sci. Appl. 1 (1994) 47–58.
[7] Y. Hayashi, R. Setiono, K. Yoshida, A comparison between two neural network rule extraction techniques for the diagnosis of hepatobiliary disorders, Artif. Intell. Med. 20 (2000) 205–216.
[8] S. Mitra, Fuzzy MLP based expert systems for medical diagnosis, Fuzzy Sets Syst. 65 (1994) 285–296.
[9] Y. Hayashi, J.J. Buckley, E. Czogala, Fuzzy neural network with fuzzy signals and weights, Int. J. Intell. Syst. 8 (4) (1993) 527–537.
[10] S.K. Pal, S. Mitra, Multi-layer perceptron, fuzzy sets and classification, IEEE Trans. Neural Networks 3 (1992) 683–697.

Yoichi Hayashi received the B.E. degree in Management Science, and the M.E. and Dr. Eng. degrees in Systems Engineering, all from the Science University of Tokyo, Japan, in 1979, 1981, and 1984, respectively. He joined Ibaraki University, Japan, as an Assistant Professor in 1986 and was a Visiting Professor at the University of Alabama at Birmingham and University of Canterbury for 10 months. Currently, he is a Professor of Computer Science at Meiji University, Japan. He has published 140 papers in academic journals and international conference proceedings in the fields of computer and information sciences. His current research interest includes artificial neural networks, fuzzy logic, soft computing, expert systems, computational intelligence, data mining and medical informatics. Dr. Hayashi is an Associate Editor of *IEEE Transactions on Fuzzy Systems*. He is a member of the IEEE, ACM, AAAI, IFSA, INNS, NAFIPS, IPSJ and EICE.

Rudy Setiono received the B.S. degree in Computer Science from Eastern Michigan University in 1984, the M.S. and Ph.D. degrees in Computer Science from the University of Wisconsin-Madison in 1986 and 1990, respectively. Since August 1990, he has been with the National University of Singapore where he is currently an Associate Professor at the Information Systems Department, School of Computing. He is IEEE Senior Member and an Associate Editor of *IEEE Transactions on Neural Networks.*