# A Learning Method of Directly Optimizing Classifier Performance at Local Operating Range

Lae-Jeong Park and Jung-Ho Moon

Department of Electrical Engineering, Kangnung National University
Kangnung, Gangwon-Do, 210-702, South Korea

{ljpark, itsmoon}@kangnung.ac.kr

## Abstract

This paper addresses an effective learning method that enables us to directly optimize neural network classifier's discrimination performance at a desired local operating range by maximizing a partial area under a receiver operating characteristic (ROC) or domain-specific curve, which is difficult to achieve with classification accuracy or mean squared error (MSE)-based learning methods. The effectiveness of the proposed approach is demonstrated in terms of fraud detection capability in the credit card fraud detection, compared with the MSE-based approach.
**Keyword**: Classification, Receiver Operating Characteristic, And Area Under Curve.

## I. Introduction

In general, design and learning of a classifier for financial real-world two-class classification problems are plagued by severely overlapping class distribution because samples in one class are often similar or even identical to those in the other class due to the nature of the problems. Examples are database marketing, churn prediction, and fraud detection.

Classification accuracy or minimization of misclassifications has been conventionally used to evaluate classifier's discrimination ability. For neural network (NN) classifiers, the mean squared error (MSE) between the actual output and the desired target is defined and minimized to reduce the number of misclassifications. In the severely overlapping class distribution problems, a trade-off between true positive rate (TPR) and false positive rate (FPR) is unavoidable and hence should be determined carefully. The MSE minimization is, however, unsuitable to evaluation of a NN classifier for the severely overlapping class problem because the MSE is not one-to-one correspondent to the performance in terms of TPR and FPR of the NN classifier.

Receiver operating characteristic (ROC) curve[1] has been recently used to evaluate classifier's discrimination performance in the skewed and overlapping data sets since its introduction by [1] in the data mining communities. The ROC curve makes it possible to visualize a trade-off of classifier's discrimination capability that is indistinguishable in the MSE measure. Moreover, so-called AUC

[1] The TPR is plotted on the *Y* axis and the FPR is plotted on the *X* axis. The pairs of TPR and FPR are obtained by varying a decision threshold on the single continuous output of a classifier.

(the area under ROC curve)[2] has been proposed as a single performance measure to deal with a problem with specifying classifier's performance in terms of a single operating point on its ROC curve [2]. The AUC reflects an average classifier's performance on the entire operating points. The larger AUC is, the better classifier's average discrimination performance is. The AUC may be an appropriate performance measure if the class ratio and misclassification costs are unknown and/or a single classifier with a fixed decision threshold must be chosen to handle every possible operating points. Recently, much attention have been paid to design and train classifiers by maximizing the AUC in financial and medical applications [3,4,5,6,7].

On the other hand, in some classification applications, it is often desirable or important to evaluate and optimize classifier's discrimination performance at a certain operating range, not in the entire operating range as in the AUC. For example, in medical diagnosis, TPRs of less than, say, 0.7-0.8 would be probably unacceptable, because patients with a disease should be detected even if it turns out that it is a false detection. In credit card fraud detection, a fraud detection system should not operate in a range of high FPRs because it cannot handle the overwhelming number of suspicious transactions. In order to produce a high-quality classifier in those application domains, a method for optimizing classifier's discrimination performance at a desired local operating range, for example, the TPR at a certain range of FPRs is required, but to my knowledge, how to optimize a partial area under a ROC curve or the like has been rarely addressed.

In this paper, an effective learning approach is proposed that makes it possible to optimize the discrimination performance of a NN classifier at a desired specific operating range by utilizing a partial area under a ROC or domain-specific curve, which is difficult to achieve with common MSE-based learning methods. The performance of the proposed approach is examined and compared with the MSE-based approach in credit card fraud detection, which is a representative one of the severely overlapping real-world classification problems.

## II. Optimization of Classifier Performance at Local Operating Range

### A. *ROC, AUC, and Partial AUC*

The ROC curve of a NN classifier is obtained by varying a threshold $\theta$ on the continuous output of the classifier, ranging from [0,1]. An example of histograms of classifier outputs of positive and negative samples is illustrated in Fig. 1. Samples whose their outputs are greater than $\theta$ are classified as positives, otherwise they are classified as negatives. The more overlapping the two classes are, the more likely it is that the two histograms overlap each other. Given a value of $\theta$, a point of (FPR,TPR) on the ROC space is determined by

$$FPR = \frac{negatives\ incorrectly\ classified}{total\ negatives},$$

$$TPR = \frac{positives\ correctly\ classified}{total\ \ positives}. \tag{1}$$

[2] AUC stands for the Area Under the Curve. For a ROC curve $r(x)$, the AUC is represented by $\int_0^1 r(x)\cdot dx$, where $x$ is FPR.

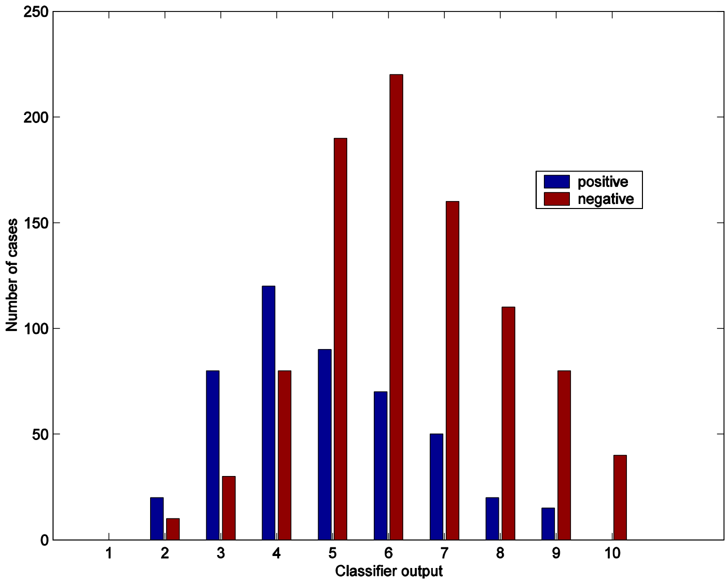


Figure 1 Two histograms of classifier’s outputs of samples of the two classes

Fig. 2 shows ROC curves of three distinct NN classifiers. Classifier *B* is completely outperformed by classifier *A*. The less likely it is that the two histograms of classifier outputs overlap each other, the more bowed the ROC curve is toward the left corner of (0,1), which represents the perfect discrimination. The AUC is a single measure to quantify the difference of classifier performances across all decision thresholds. The AUC of classifier *A* is definitely larger than that of classifier *B*. The AUC has been recently adopted as an informative classifier performance measure compared with the error rate and therefore several methods to optimize the AUC directly have been proposed. Verrelst et. al. adopted simulated annealing to maximize the AUC to produce a MLP classifier in ovarian tumor malignancy prediction problem [3]. Yan et. al. proposed a gradient-based training algorithm for directly maximizing the AUC by using a differentiable objective function that is approximation to the Wilcoxon-Mann-Whitney statistic, which is equivalent to the AUC [4]. In some applications, it is meaningful to focus on an area under only a portion of the ROC curve, for example, within a specific range of FPRs or TPRs because no interest lies in the entire range of FPRs or TPRs. For example, in a diagnostic test, much attention is generally paid to the portion of the ROC curve where TPR is greater than a predetermined threshold. For those applications, since a classifier with a larger AUC value may have lower discrimination performance than classifiers with smaller AUC values in a specific range of FPRs or TPRs (as shown in Fig. 2, classifier *C* is slightly better than classifier *A* at a low FPR range, but poor in terms of AUC values), it is required to train a classifier by optimizing discrimination performance at a desired local operating range that may be indistinguishable by the AUC performance measure. To my knowledge, there have been few attentions on importance of classifier's discrimination performance at a local operating range in machine learning and data mining communities, and therefore, there have been few researches on methods designed to train a NN classifier by directly optimizing a partial area under the ROC curve or an application-specific curve, so-called PAUC (Partial AUC).

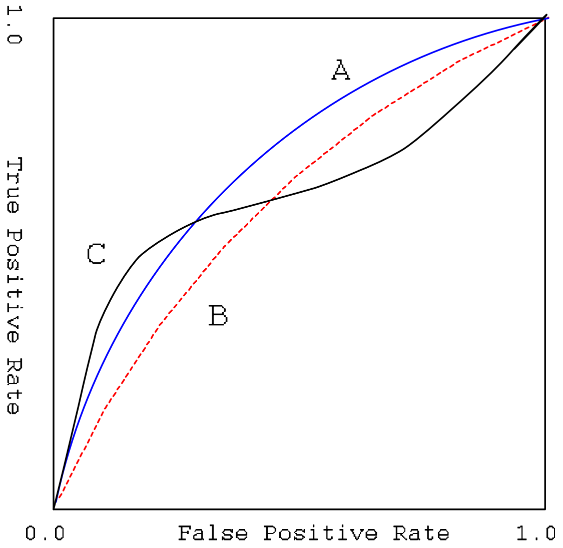


Figure 2 ROC curves of three distinct classifiers

## B. *PAUC Maximization*

Unfortunately, the backpropagation (BP) algorithm that minimizes the MSE of a NN classifier is not suited for optimizing classifier's performance in terms of the PAUC. It is because the MSE minimization at the outputs does not correspond directly to the PAUC-based optimization. A nonstandard training method is required because it is impossible to compute the partial derivatives of the PAUC with respect to the weights. A NN classifier is, therefore, trained by an effective search algorithm or Evolutionary Programming (EP) rather than gradient-descent algorithms so that the weights are chosen in terms of the PAUC optimization. An objective function for the PAUC optimization is given in the form of

$$\frac{\int_{\alpha}^{\beta} C(x)\cdot dx}{\int_{0}^{1} C(x)\cdot dx}, \tag{2}$$

where [α,β] specifies a desired operating range, $C(x)$ is a ROC curve or a domain-specific curve, and $x$ is a FPR or problem-dependent variable. Table 1 shows a procedure for numerically calculating the PAUC value of a NN classifier. In the STEP 3, NN classifiers that the number of (TPR, FPR) points in [α,β] is smaller than κ, is excluded or penalized during the evolutionary search in order not to produce NN classifiers that have a risk of generating an abrupt transition of classifier output histograms, which may be associated with poor discrimination performance.

Table 1. A procedure for numerical PAUC calculation of a NN classifier

| | |
|---|---|
| Input: | [α,β] : a range of FPRs<br>$N_\theta$ : The number of thresholds between [0,1] |

STEP 1: $N_\theta$ pairs of (TPR, FPR) are obtained by varying the threshold on the NN classifier’s output

STEP 2: A ROC curve is constructed by using $N_\theta$ pairs of (TPR, FPR)

STEP 3: A partial area under the curve in the range of $[\alpha,\beta]$ is calculated with the trapezoidal integration rule

## III. Experiments on Fraud Detection

### A. *Credit Card Fraud Detection*

Credit card fraud detection [8,9] is one of real-world two-class classification problems where the class distribution is severely overlapping and the class ratio is highly skewed. Fraudulent transactions resemble legitimate ones so that it is impossible to detect fraudulent transactions without a lot of false detection. Hence, a classifier for credit card fraud detection should be operated on a very low range of FPRs in order to reduce the number of legitimate transactions incorrectly detected. Therefore, it is crucial to optimize classifier's discrimination performance at a specific local range of FPRs, which is determined based on the several operational constraints. It should be noted that learning of classifiers is performed not at a single operating point but in a local operating range, because an operating point changes with time, from a month to another, depending on the changing class distribution.

Two data sets of credit card transactions labeled as legitimate or fraudulent were provided by a credit card company in Korea. A set consisting of about 51,260 transactions collected selectively during one year is used as a training data set. Another data set of about 7 millions transactions during three months is used to evaluate the NN classifier. A multi-layered perceptron (MLP) with one hidden layer and sigmoidal function is used as the NN classifier. The six features are extracted from each transaction by using machine learning techniques and are then used for the inputs of the NN classifier.

### B. *Application-specific PAUC Maximization*

For credit card fraud detection, it is practically desirable that an operating range of the NN classifier is represented by a rejection rate (true positives plus false positives) since the rate is a major monitoring variable that is carefully controlled due to the constraint imposed on the capacity of suspicious transaction investigation. Instead of a ROC curve, a domain-specific curve $C(x)$ is introduced that represents the number of correctly detected fraudulent transactions with respect to a rejection rate, $x$.

The EP with Gaussian mutation is used to train a NN classifier in order to maximize Eq. (2). With the domain-specific curve, the objective is to maximize an average of the correctly detected frauds in the chosen range of the rejection rates by maximizing the partial area under the curve. A penalty term is added to the cost function of EP so as to penalize classifiers that have a small number of (TPR, FPR) points falling on $[\alpha,\beta]$ since their decision boundaries locate the regions on the feature space where samples are distributed densely. The value of $\kappa$ is set to 20 by trial-and-error. The EP parameter values for all experiments are as follows. The number of generations and the population size are chosen as 1000 and 20, respectively, for a reasonable convergence speed. The tournament size is 10 and σ in the Gaussian mutation is 1.

### C. *Experimental Results and Discussions*

Two sets of experiments were performed to evaluate and compare the PAUC-based NN classifier with the MSE-based one. For fair comparison, the MSE minimization was performed by the EP, too. For each criterion, ten NNs were trained with 10, 15, 20 hidden nodes. The desired operating range was chosen as [0.001,0.004] to reflect reality. Table 2 shows the averaged MSE and PAUC values of ten classifiers trained with the PAUC criterion and ten ones with the MSE criterion. As expected, the averaged MSE value of classifiers trained with the MSE criterion is significantly smaller than that of classifiers trained with the PAUC criterion, and averaged PAUC value of classifiers trained with the PAUC criterion is significantly larger than that of classifiers trained with the MSE criterion. An explanation could be that, by maximization of Eq. (2) the decision boundary of a NN classifier is formed in such a way that $C(x)$ within the operating range of $[\alpha,\beta]$ increases regardless of large MSE contribution of the samples associated with the rejection rates of $[\beta,1]$.

Table 2. Averages and standard deviations of MSE and PAUC values of twenty NN classifiers, a half trained with the PAUC criterion, and the others with the MSE criterion

| Learning criterion | (MSE, PAUC) |
|---|---|
| MSE minimization | (3.07±0.03, 2.91±0.10) |
| PAUC maximization | (4.05±0.85, 3.05±0.02) |

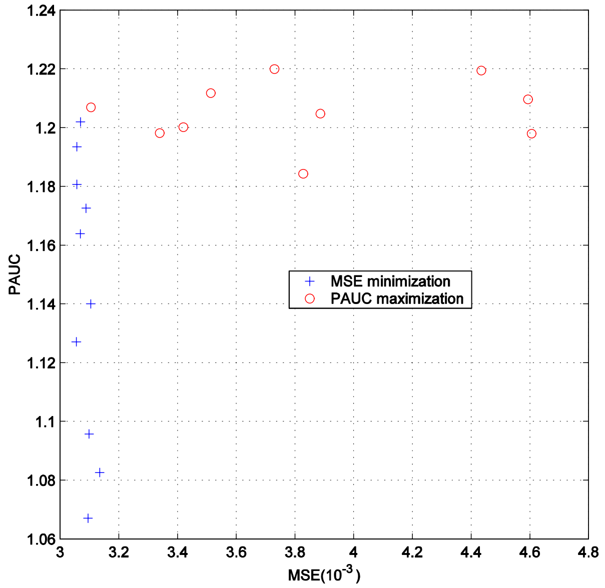


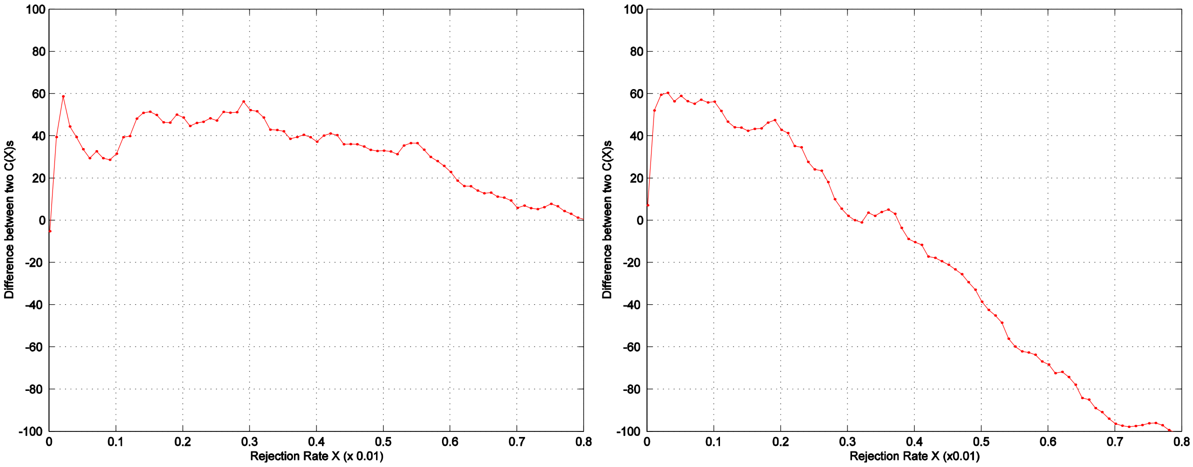
Figure 3 X-Y plot of twenty NN classifiers' performance on the (MSE, PAUC) space, a half trained with the PAUC, the others with the MSE

Fig. 3 shows a X-Y plot of the twenty classifiers' performance on the (MSE, PAUC) space. As seen in Fig. 3, classifiers trained with the PAUC criterion (or MSE criterion) have a large σ in the MSE values (or PAUC values) compared with σ in the PAUC values (or MSE values), which implies that the MSE minimization does not correspond to the PAUC maximization one-to-one. The similar results between AUC and MSE performance indices have been observed in [6,7]. It has been revealed analytically in [6] that algorithms designed to minimize the error rate may not lead to the best AUC possible values. In [7], by empirical comparison of error surfaces in the weight space of MSE, AUC, and partial AUC performance measures, they showed that MSE minimization tends to maximize AUC, but not partial AUC defined at a range of high true positive rates.

Fig. 4 shows a difference between the two averaged curves of the top five classifiers trained with the PAUC maximization and the top five ones with the MSE minimization on the test set. Note that a coordinate value of 0 in the Y axis represents that the PAUC maximization gives no performance improvement over the MSE minimization. Even though training of a NN classifier with the PAUC criterion generates classifiers with large MSE values, compared with the MSE-based classifiers, it enables a NN classifier to detect more fraudulent transactions at the specified rejection rate of [0.001,0.004] by increasing the partial area under the curve within the operating range.

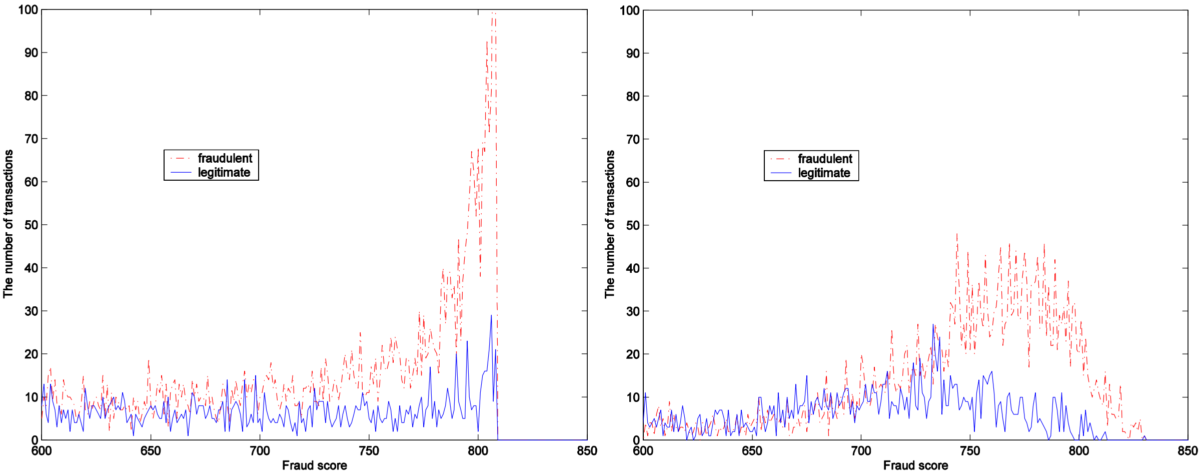
The performance improvement at the local operating range results from the difference of shape of the two histograms of classifier's outputs of legitimate and fraudulent transaction samples. Fig. 5 shows the histograms of classifier's outputs of legitimate (solid line) and fraudulent (dash-dotted line) transaction samples for MSE-based and PAUC-based classifiers with 10 hidden nodes. The fraud score is calculated by 500(1+*o*), where *o* is the classifier output ranging between -1 and 1. For the MSE-based classifier, the histogram of fraudulent samples has the spike-like shape near at the fraud score of 800, which is caused by the force to minimize the MSE during training (The target outputs of the fraudulent samples are set to +1). On the other hand, for the PAUC maximization, instead of adjusting the classifier outputs to the associated target values, a classifier is trained in such a way that the output histograms of the two classes are overlapped as less as possible at a high range of fraud scores that corresponds to the specified rejection rate [0.001,0.004]. Hence, it can be seen in Fig. 5(b) that the spike-like shape of the histograms of Fig. 5(a) disappears and that the two histograms flatten and get less overlapped (more separable) around at the fraud score of 800.

On the other hand, it is observed that the curve $C(x)$ of the PAUC-based classifier increases more slowly at a medium range of the rejection rates than that of MSE-based classifiers to compensate the increase in the range of [0.001,0.004]. The improvement is consistent with the number of hidden nodes between 10 and 20.



(a) (b)

Figure 4 Difference between the two averaged curves of the top five NN classifiers in terms of the PAUC and MSE. (a) 10 hidden nodes. (b) 20 hidden nodes



(a) (b)

Figure 5 Histograms of classifier outputs of legitimate and fraudulent transaction samples. (a) histograms of MSE-based classifier. (b) histograms of PAUC-based classifier

## IV. Conclusion

A learning method has been proposed that enables us to directly optimize neural network classifier's discrimination performance at a desired specific operating range by maximizing a partial area under a ROC or domain-specific curve, which is difficult to achieve with MSE-based learning methods. The effectiveness of the proposed approach has been demonstrated and compared with the MSE-based approach in terms of discrimination performance in credit card fraud detection in which interests lie in only a range of very low false positive rates. The experimental results with real credit card transactions data have demonstrated that the proposed approach makes it possible to detect more fraudulent transactions in a desired operating range.

## References

[1] F. Provost and T. Fawcett, “Analysis and visualization of classifier performance comparison under imprecise class and cost distributions,” in *Proc. of Conf. Knowledge Discovery and Data Mining*, 1997, pp. 43–48.

[2] A. P. Bradley, “The use of the area under the ROC curve in the evaluation of machine learning algorithms,” in *Pattern Recognition*, vol. 30, 1997, pp. 1145–1159.

[3] H. Verrelst, Y. Moreau, and D. Timmerman, “Use of a multi-layer perceptron to predict malignancy in ovarian tumors,” in *Proc. of Conf. on Advances in Neural Information Processing Systems*, 1998, pp. 978–984.

[4] L. Yan, R. Dodier, M. C. Mozer, and R. Wolniewicz, “ Optimizing classifier performance via the Wilcoxon-Mann-Whitney statistics,” in *Proc. of Int. Conf. on Machine Learning*, 2003, pp. 848–855.

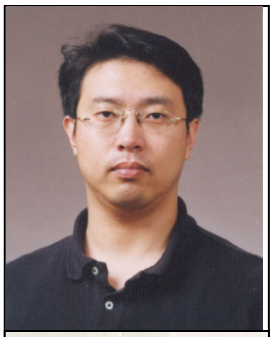
[5] B. Sahnier, H.-P. Chan, N. Petrick, S. S. Gopal, and M. M. Goodsitt, “Neural network design for optimization of the partial area under the receiver operating characteristic curve,” in *Proc. IEEE Conf. on Neural Networks*, 1997, pp. 2468–2471

[6] C. Cortes and M. Mohri, “AUC optimization vs. error rate minimization,” in *Proc. of Conf. on Advances in Neural Information Processing Systems*, 2003.

[7] M. K. Markey, J. Y. Lo, R. Vargas-woracek, G. D. Tourassi, and Jr. C. E. Floyd, “Perceptron error surface analysis: A case study breast cancer diagnosis,” in *Computers in Biology and Medicine*, vol. 32, 2002, pp. 99–109.

[8] R. J. Bolton and D. J. Hand, “Statistical fraud detection: A review,” in *Statistical Science*, vol. 17, 2002, pp. 235–255.

[9] P. K. Chan, W. Fan, A. L. Prodromidis, and S. J. Stolfo, “Distributed data mining in credit card fraud detection,” in *IEEE Intelligent Systems*, vol. 14, 1999, pp. 67–74.



Lae-Jeong Park received the B.S. degree in Electrical Engineering from Seoul National University in 1991, and the M.S. and Ph.D. degrees in Electrical Engineering from the Korea Advanced Institute of Science and Technology in 1993 and 1997, respectively. He was with the Information Technology Lab. at LG Corporate Institute of Technology, Seoul, Korea. He is currently an assistant professor in the Department of Electrical Engineering at Kangnung National University, Korea. His current research interests are machine learning, pattern recognition, and evolutionary and neural computation.

Jung-Ho Moon received the B.S. degree in control and instrumentation engineering from Seoul National University in 1991, the M.S. and Ph.D. degrees in electrical engineering from Korea Advanced Institute of Science and Technology in 1993 and 1998, respectively. In 2002, he joined the faculty of the Department of Electrical Engineering, Kangnung National University, Korea, where he is an Assistant Professor. From 1998 to 2000, he worked for Samsung Electronics Co., Ltd as a Senior Research Engineer, and from 2001 to 2002, he worked for Humax Co., Ltd as a Senior H/W Engineer. His research interests include digital control, disk drive servo systems, intelligent control, and embedded systems.