http://dx.doi.org/10.5392/IJoC.2012.8.2.007

Improving the Error Back-Propagation Algorithm for Imbalanced Data Sets

Sang-Hoon Oh
Department of Information Communication Engineering
Mokwon University, Daejon, 302-729, Korea

ABSTRACT

Imbalanced data sets are difficult to be classified since most classifiers are developed based on the assumption that class distributions are well-balanced. In order to improve the error back-propagation algorithm for the classification of imbalanced data sets, a new error function is proposed. The error function controls weight-updating with regards to the classes in which the training samples are. This has the effect that samples in the minority class have a greater chance to be classified but samples in the majority class have a less chance to be classified. The proposed method is compared with the two-phase, threshold-moving, and target node methods through simulations in a mammography data set and the proposed method attains the best results.

Keywords*: Imbalanced Data, Error Back-Propagation, Error Function, Mammography.*

1. INTRODUCTION

The class imbalance problem has been emerged as a new challenge, since it is reported in a wide range of applications [1]-[4]. Particularly, for bi-class applications of imbalanced data sets, the usual class is represented by a large number of samples while the other unusual class with great interest is represented by only a few samples. This imbalanced class distribution of a data set has posed a serious difficulty for most classifier learning algorithms, which assume a relatively balanced class distribution and equal misclassification costs [5].

Multilayer perceptrons (MLPs) have been widely applied to pattern classification problems. A popular method of training MLPs is the error back-propagation (EBP) algorithm, which is a gradient descent with a fixed learning rate [6]. When applying the conventional EBP algorithm to the imbalanced data sets, the class boundary of the majority class is enlarged towards the minority class [7]. In order to prevent this boundary distortion, Bruzzone and Serpico proposed the two-phase method which strengthens the error function related to the minority class [8]. Also, the threshold-moving method adjusts the output thresholds of neural networks such that samples in the minority class become harder to be misclassified [9]. In spite of performance improvement, these methods show serious fluctuations in their learning curves because of the incorrect saturation of output nodes [10][11]. For preventing the above problems, an error function was proposed so that the weight-updating of neural networks was controlled with regards to the target node of each class [11]. This target node method showed better performances than the two-phase and threshold moving methods without serious fluctuations of learning curves. However, the target node method has a heuristic procedure to fix the imbalance of cases in which each output node is selected as the target node.

This paper proposes a new error function which can improve the EBP algorithm for imbalanced data sets without any heuristic procedure. In section 2, the conventional EBP algorithm is briefly introduced. A new error function for the imbalanced data set is proposed in section 3 and section 4 shows the effectiveness of the proposed method through simulations of the mammography data set. Finally, section 5 concludes this letter.

2. ERROR BACK-PROPAGATION ALGORIHTM

Consider an MLP consisting of *N* inputs, *H* hidden nodes, and *M* output nodes, which is denoted as an "*N-H-M* MLP". When a sample $\mathbf{x}^{(p)} = [x_1^{(p)}, x_2^{(p)}, \ldots, x_N^{(p)}]$ $(p = 1,2,\ldots,P)$ is presented to the MLP, the *j*-th hidden node is given by

$$h_j^{(p)} = h_j(\mathbf{x}^{(p)}) = \tanh((w_{j0} + \sum_{i=1}^{N} w_{ji} x_i^{(p)})/2), \quad j = 1,2,\ldots,H. \tag{1}$$

Here, w_{ji} denotes the weight connecting x_i to h_j and w_{j0} is a bias. The *k*-th output node is

$$y_k^{(p)} = y_k(\mathbf{x}^{(p)}) = \tanh(\hat{y}_k^{(p)}/2), \quad k = 1,2,\ldots,M, \tag{2}$$

where

$$\hat{y}_k^{(p)} = v_{k0} + \sum_{j=1}^{H} v_{kj} h_j^{(p)} . \tag{3}$$

* *Corresponding author, Email: shoh@mokwon.ac.kr*
Manuscript received Apr. 20, 2012; revised Jun 09, 2012; accepted Jun 11, 2012

www.kci.go.kr

Also, v_{k0} is a bias and v_{kj} denotes the weight connecting h_j to y_k.

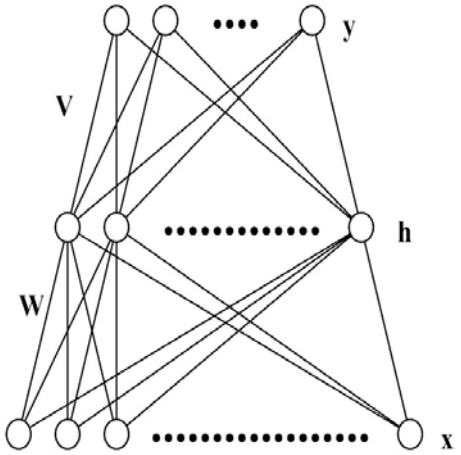


Fig. 1. The architecture of a multilayer perceptron.

Let the desired output vector corresponding to the training sample $\mathbf{x}^{(p)}$ be $\mathbf{t}^{(p)} = [t_1^{(p)}, t_2^{(p)}, \ldots, t_M^{(p)}]$, which is coded as follows:

$$t_k^{(p)} = \begin{cases} +1 & \text{if } \mathbf{x}^{(p)} \text{ originates from class } k \\ -1 & \text{otherwise.} \end{cases} \tag{4}$$

As a distance measure between the actual and desired outputs, we usually use the squared error function for P training samples defined by

$$E = \sum_{p=1}^{P} \sum_{k=1}^{M} \left(t_k^{(p)} - y_k^{(p)} \right)^2. \tag{5}$$

To minimize E, weights v_{kj}'s are iteratively updated by

$$\Delta v_{kj} = -\eta \frac{\partial E}{\partial v_{kj}} = \eta \delta_k^{(p)} h_j^{(p)}, \tag{6}$$

where

$$\delta_k^{(p)} = -\frac{\partial E}{\partial \hat{y}_k^{(p)}} = \left(t_k^{(p)} - y_k^{(p)} \right) \frac{\left(1 - y_k^{(p)}\right)\left(1 + y_k^{(p)}\right)}{2} \tag{7}$$

is the error signal and η is the learning rate. Also, weights w_{ji}'s are updated by

$$\Delta w_{ji} = -\eta \frac{\partial E}{\partial w_{ji}} = \eta x_i^{(p)} \sum_{k=1}^{M} v_{kj} \delta_k^{(p)} . \tag{8}$$

The above weight-updating procedure is the EBP algorithm [6].

3. PROPOSED ERROR FUNCTION

This paper considers the two-class problems with imbalanced data sets [1-4]. Assume that one is the minority class C_1 with P_1 training samples and the other is the majority class C_2 with P_2 training samples ($P_1 << P_2$). If we use the conventional EBP algorithm to train the MLP, weight-updating is overwhelmed by the majority class samples and this severely distorts the boundary between the two classes [7]. That is, the boundary of the majority class is enlarged to the boundary of the minority class. This gives a less chance to be classified for the minority class samples while samples in the majority class have a greater chance to be classified. Finally, we attain poor classification performance for the minority class even though samples in the minority class have a high misclassification cost.

In order to resolve the above problem, we propose a new error function which can intensify weight-updating for the minority class samples and weaken weight-updating for the majority class samples. Accordingly, the proposed error function is defined by

$$\begin{aligned} E_{prop} = &-\sum_{p \in C_1} \sum_{k=1}^{2} \int \frac{t_k^{(p)\,n+1} \left(t_k^{(p)} - y_k^{(p)} \right)^n}{2^{n-2} \left(1 - y_k^{(p)\,2} \right)} dy_k^{(p)} \\ &-\sum_{p \in C_2} \sum_{k=1}^{2} \int \frac{t_k^{(p)\,m+1} \left(t_k^{(p)} - y_k^{(p)} \right)^m}{2^{m-2} \left(1 - y_k^{(p)\,2} \right)} dy_k^{(p)} \end{aligned}, \tag{9}$$

where n and m (n<m) are positive integers and $t_k^{(p)}$ is coded as in (4). If n=m, the proposed error function is the same as the n-th order error function proposed in [10] which dramatically reduces the incorrect saturation of output nodes. Using the proposed error function, the error signal of the output layer is given by

$$\begin{aligned} \delta_k^{(p)} &= -\frac{\partial E_{prop}}{\partial \hat{y}_k^{(p)}} \\ &= \begin{cases} t_k^{(p)\,n+1} \left(t_k^{(p)} - y_k^{(p)} \right)^n / 2^{n-1} & \text{for } p \in C_1, \\ t_k^{(p)\,m+1} \left(t_k^{(p)} - y_k^{(p)} \right)^m / 2^{m-1} & \text{for } p \in C_2. \end{cases} \end{aligned} \tag{10}$$

The parameters n and m controls the updating amount of weights whether training samples are in the minority or majority classes. Since n<m and $-1 \le y_k^{(p)} \le 1$, the error signal for C_1 is greater than or equal to the error signal for C_2. The associated weights are updated in proportion to the error signals, which is the same procedure as in the EBP algorithm [6]. Thus, in order to prevent the boundary distortion, the proposed error function generates a stronger error signal for the minority class.

The proposed error function can be written by

$$E_{new} = \sum_{p=1}^{P} \sum_{k=1}^{2} l^{(r)}(t_k^{(p)}, y_k^{(p)}) \tag{11}$$

where

$$l^{(r)}(t, y) = -\int \frac{t^{r+1} (t - y)^r}{2^{r-2} \left(1 - y^2 \right)} dy \tag{12}$$

and

$$r = \begin{cases} n & \text{for } p \in C_1, \\ m & \text{for } p \in C_2. \end{cases} \tag{13}$$

In the limit $P \to \infty$, the minimizer of E_{new} converges (under certain regularity conditions, Theorem 1 in [12]) towards the

www.kci.go.kr

minimizer of the expectation function

$$E\left\{\sum_{k=1}^{2} l^{(r)}(T_k, y_k(\mathbf{X}))\right\} = \int \sum_{k=1}^{2} \left[Q_k(\mathbf{x}) l^{(r)}(1, y_k(\mathbf{x})) + (1 - Q_k(\mathbf{x})) l^{(r)}(-1, y_k(\mathbf{x}))\right] f(\mathbf{x}) d\mathbf{x}. \quad (14)$$

Here, $E\{\cdot\}$ is the expectation operator, $\mathbf{X}$ is the random vector denoting an input pattern, and T_k is the random variable denoting the target. Also,

$$Q_k(\mathbf{x}) = \Pr[\mathbf{X} \text{ originates from class } k \mid \mathbf{X} = \mathbf{x}] \quad (15)$$

and $f(\mathbf{x})$ is the probability density function of $\mathbf{x}$. Let us seek the function $\mathbf{b} = [b_1, b_2]^T$ minimizing the criterion (14) [in the space of all functions taking values in (-1,1)][13]. For a fixed $Q_k(\mathbf{x})$, $0 < Q_k(\mathbf{x}) < 1$, the optimal solution $\mathbf{b(X)} = [b_1\mathbf{(X)}, b_2\mathbf{(X)}]^T$ can be derived by

$$\frac{\partial}{\partial y_k}\left[Q_k(\mathbf{x}) l^{(r)}(1, y_k(\mathbf{x})) + (1 - Q_k(\mathbf{x})) l^{(r)}(-1, y_k(\mathbf{x}))\right] = 0, \quad k = 1,2. \quad (15)$$

As a result,

$$b_1(\mathbf{x}) = h_{m,n}^{-1}(Q_1(\mathbf{x})) \text{ and } b_2(\mathbf{x}) = h_{n,m}^{-1}(Q_2(\mathbf{x})), \quad (16)$$

where $h_{m,n} : (-1,1) \to (0,1)$ is given by

$$h_{m,n}(v) = \frac{\left(\frac{1+v}{2}\right)^m}{\left(\frac{1-v}{2}\right)^n + \left(\frac{1+v}{2}\right)^m}. \quad (17)$$

Fig. 2 shows the optimal solution with *n*=2 and *m*=4. Since $b_1(\mathbf{x}) > b_2(\mathbf{x})$ with the condition n<m, the proposed error function has an effect of threshold adjusting for classification of imbalanced data. Also, notice that the optimal solution is a strictly increasing function and the Bayes classifier can be defined by

$$\text{decide } k \text{ if } k = \arg_k[\max y_k(\mathbf{x})]. \quad (18)$$

This argument shows that the proposed error function is sensible in a bi-classification task of imbalanced data sets, provided that there are two output nodes whose targets are coded as in (4). If we use a different coding of targets, we should modify the proposed error function. The proposed error function works only for bi-class tasks. Multi-class tasks are more difficult than two-class tasks and a higher degree of class imbalance may increase the difficulty[9]. So, it is another big issue to handle the imbalance in multi-class tasks.

Since the class 1 is the minority class and the class 2 is the majority class, the cases that $t_1^{(p)} = 1$ and $t_2^{(p)} = -1$ is much less than the cases that $t_1^{(p)} = -1$ and $t_2^{(p)} = 1$. As shown in Fig. 2, $|b_1(\mathbf{x})| > |b_2(\mathbf{x})|$ for $0.5 < Q_k(\mathbf{x}) < 1$ and the cases that $t_1^{(p)} = 1$ is less than the cases that $t_2^{(p)} = 1$. For $0 < Q_k(\mathbf{x}) < 0.5$, $|b_1(\mathbf{x})| < |b_2(\mathbf{x})|$ and the cases that $t_2^{(p)} = -1$ is less than the cases that $t_1^{(p)} = -1$. Thus, the stronger optimal solution corresponds to the less cases of target selection.

In the target node method[11], on the contrary, $|b_1(\mathbf{x})| > |b_2(\mathbf{x})|$ for $0.5 < Q_k(\mathbf{x}) < 1$ and the cases that $t_1^{(p)} = 1$ is less than the cases that $t_2^{(p)} = 1$. For $0 < Q_k(\mathbf{x}) < 0.5$, $|b_1(\mathbf{x})| > |b_2(\mathbf{x})|$ and the cases that $t_2^{(p)} = -1$ is less than the cases that $t_1^{(p)} = -1$. That is, the weak optimal solution corresponds to the less cases of target selection for $0 < Q_k(\mathbf{x}) < 0.5$ and the stronger optimal solution corresponds to the less cases of target selection for $0.5 < Q_k(\mathbf{x}) < 1$. So, there must be a heuristic procedure to fix the imbalance of target selection in the target node method. However, in the proposed method, it is not necessary to fix the imbalance of target selection since the stronger optimal solution corresponds to the less cases and the weak optimal solution corresponds to the more cases for whole range of $Q_k(\mathbf{x})$.

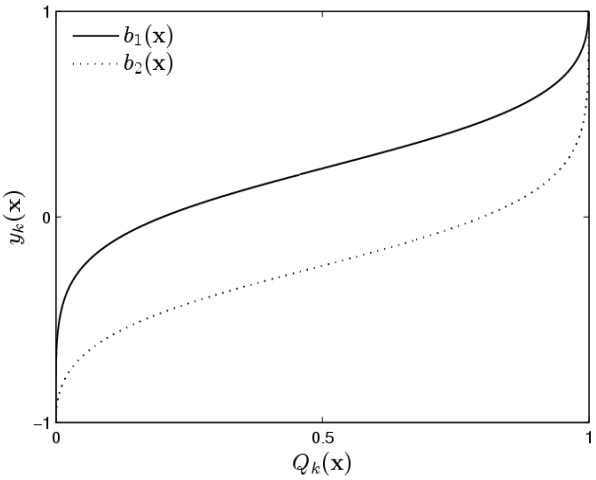


Fig. 2. The optimal solutions of $y_k(\mathbf{X})$ for minimizing $E\{\sum_{k=1}^{2} l^{(r)}(T_k, y_k(\mathbf{X}))\}$. $E\{\cdot\}$ is the expectation operator and $\sum_{k=1}^{2} l^{(r)}(T_k, y_k(\mathbf{X}))$ is the proposed error function when a random vector $\mathbf{X}$ is presented to an MLP. Also, $Q_k(\mathbf{x})$ is the posterior probability $\Pr[\mathbf{X} \text{ originates from class } k \mid \mathbf{X} = \mathbf{x}]$.

4. SIMULATIONS

The proposed method was compared with the conventional EBP [6], two-phase [8], threshold moving [9], and target node [11] methods through simulations of "Mammography" data set [4]. The "Mammography" data set has 260 minority class samples and 10293 majority class samples, that is, the minority ratio is 2.32%. The "5-fold cross-validation" technique was used for performance evaluations because its test data is not provided. The MLP consisted of six inputs, four hidden, and

www.kci.go.kr

two output nodes. Since no fair comparison was possible if the learning rates were kept the same for all methods[14], the learning rates were derived so that $E\{\eta|\delta_k^{(p)}|\}$ had the same value in each method. Here, we assumed that $y_k^{(p)}$ had uniform distribution on $[-1,1]$ [10]. As a result, we used $\eta = 0.001\times[(n+1)P_1+(m+1)P_2]/P$, $\eta = 0.001\times[(n+1)+(m+1)]/2$, and $\eta = 0.006$ for the proposed method, the target node method, and the other methods, respectively. Let A_1 denote the accuracy for C_1 and A_2 denote the accuracy for C_2. In this letter, the G-mean (geometric mean) of A_1 and A_2 is used as a performance measure since the total accuracy is not adequate for the imbalanced data problem [7]. Nine simulations were conducted using each method with the same initializations of uniform weights on $[-1\times10^{-4}, 1\times10^{-4}]$. Totally 45 cases of simulation results-that is, nine weight initialization cases times five validation set cases-were averaged to draw figures. For fair comparisons, we tried various parameter values (*T* for the two-phase method [8], *TH* for the threshold moving method [9], and (*n*, *m)* for the target node [11] and the proposed methods, respectively) and the best average case of the G-mean in each method is shown in Figures 3 and 4.

Fig. 3 shows the G-mean curve of each method. As expected, the conventional EBP method was the worst. The threshold moving (*TH*=15), two-phase (*T*=0.2), and target node (*n*=2, *m*=18) methods improved the performance. Among the improved methods, the best one was the proposed method (*n*=2, *m*=8). For more detailed comparisons, we drew curves of A_1, A_2, the G-mean and the total accuracy in Fig. 3. As shown in Fig. 4(a), the conventional EBP resulted in very low values of A_1 and the G-Mean, although the total accuracy was about 98%. The two-phase method (Fig. 4(b)) improved A_1 and the G-mean, however, there were fluctuations of A_1 and A_2. Even though the threshold moving method (Fig. 4(c)) was better than the conventional EBP, it was worse than the two-phase method. Also, its $|A_2 - A_1|$ is greater than that of the two-phase method. Fig. 4(d) shows that the target node method improved A_1 and the G-mean with less $|A_2 - A_1|$. Finally, as shown in Fig. 4(e), the proposed method attained the best of A_1 and the G-mean with the least $|A_2 - A_1|$.

5. CONCLUSIONS

In this paper, a new error function for the EBP algorithm was proposed especially to improve the classification of imbalanced data. The proposed error function generated a stronger error signal of output node for the minority class samples and this could prevent the invasion of class boundary from the majority class to the minority class. The effectiveness of the proposed method was verified through simulations of "Mammography" data. Comparing with the conventional EBP, two-phase, threshold moving and target node methods, the proposed method attained the best performance with the criteria of A_1, G-mean and $|A_2 - A_1|$.

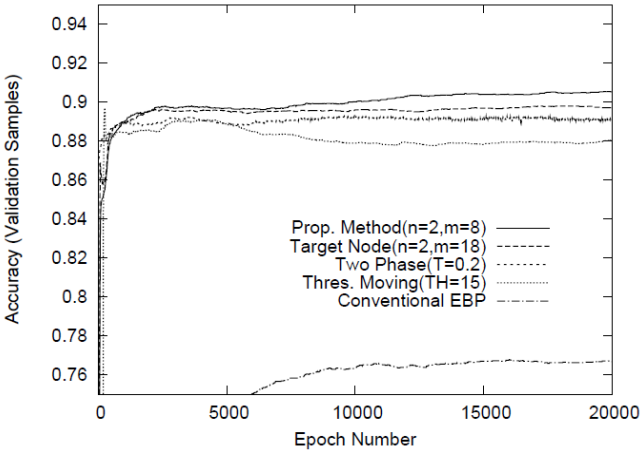


Fig. 3. The geometric mean of class accuracies for "Mammography" data. The order of curves in the legend coincides with the descending order of curves at the 20000th epoch.

REFERENCES

[1] H. Zhao, "Instance Weighting versus Threshold Adjusting for Cost-Sensitive Classification," Knowledge and Information Systems, vol.15, 2008, pp. 321-334.

[2] Y.-M. Huang, C.-M. Hung, and H. C. Jiau, "Evaluation of Neural Networks and Data Mining Methods on a Credit Assessment Task for Class Imbalance Problem," Nonlinear Analysis, vol.7, 2006, pp. 720-747.

[3] R. Bi, Y. Zhou, F. Lu, and W. Wang, "Predicting gene ontology functions based on support vector machines and statistical significance estimation," Neurocomputing, vol.70, 2007, pp.718-725.

[4] N. V. Chawla, K. W. Bowyer, L. O. all, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," J. Artificial Intelligence Research, vol.16, 2002, pp. 321-357.

[5] F. Provost and T. Fawcett, "Robust Classification for Imprecise Environments," Machine Learning, vol.42, 2001, pp. 203-231.

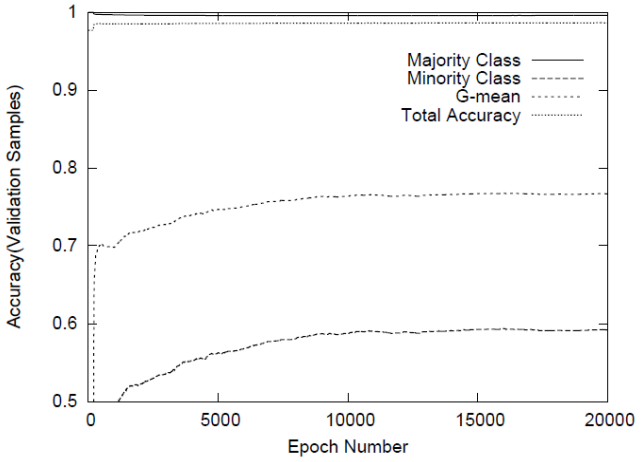
[6] D. E. Rumelhart and J. L. McClelland, *Parallel Distributed Processing*, Cambridge, MA, 1986.

[7] P. Kang and S. Cho, "EUS SVMs: ensemble of under-sampled SVMs for data imbalance problem, " *Proc. ICONIP'06*, 2006, p. 837-846.

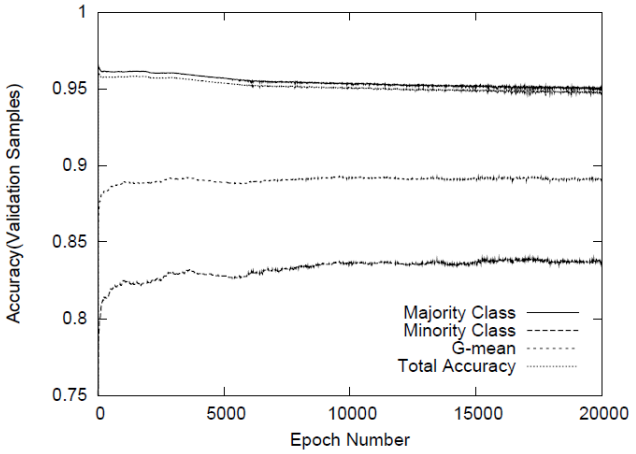
[8] L. Bruzzone and S. B. Serpico, "Classification of Remote-Sensing Data by Neural Networks," Pattern Recognition Letters, vol.18, 1997, pp. 1323-1328.

[9] Z.-H. Zhou and X.-Y. Liu, "Training Cost-Sensitive Neural Networks with Methods Addressing the Class Imbalance Problem," IEEE Trans. Know. and Data Eng., vol.18, no. 1, Jan. 2006, pp. 63-77.

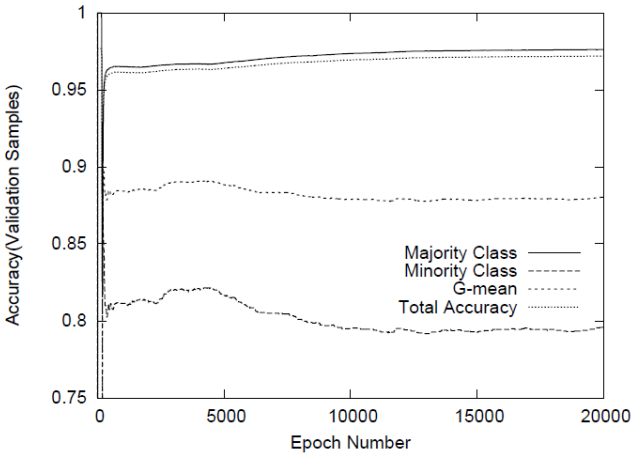
www.kci.go.kr



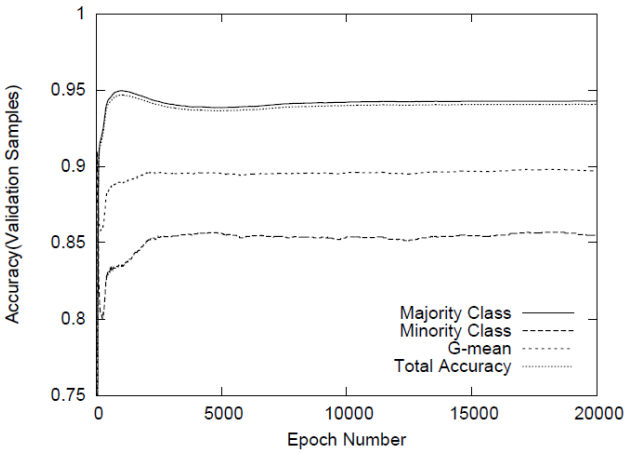
(a)



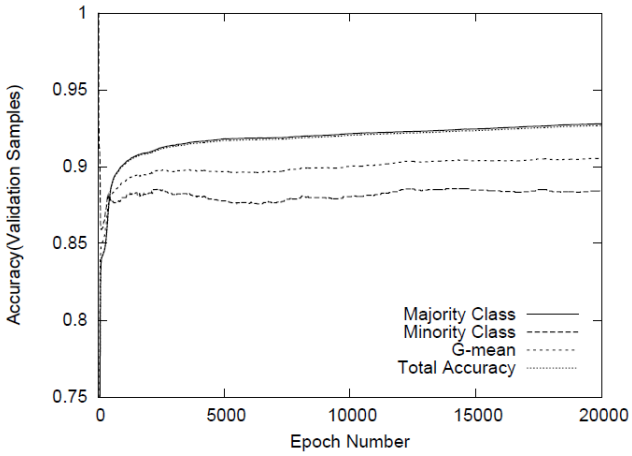
(b)



(c)



(d)



(e)

Fig. 4. Simulation results for "Mammography" data. (a) The conventional EBP method. (b) The two-phase method (T=0.2). (c) The threshold moving method (TH=15). (d) The target node method (n=2, m=18). (e) The proposed method (n=2, m=8). The curves in (a)-(e), sorted in descending order, are A_2 (the accuracy for majority class), total accuracy, G-Mean, and A_1 (the accuracy for minority class).

www.kci.go.kr

[10] S.-H. Oh, "Improving the Error Back-Propagation Algorithm with a Modified Error Function," IEEE Trans. Neural Networks, vol.8, 1997, pp. 799-803.
[11] S.-H. Oh, "Error Back-Propagation Algorithm for Classification of Imbalanced Data," Neurocomputing, vol.74, 2011, pp. 1058-1061.
[12] H. White, "Learning in Artificial Neural Networks: A Statistical Perspective," Neural Computation, vol.1, no.4, Winter 1989, pp. 425-464.
[13] S.-H. Oh, "A Statistical Perspective of Neural Networks for Imbalanced Data Problems," Int. Journal of Contents, vol.7,2011,pp.1-5.
[14] A. van Ooyen and B. Nienhuis, "Improving the convergence of the backpropagation algorithm," Neural Networks, vol.5, 1992, pp. 465-471.

Sang-Hoon Oh
received his B.S. and M.S degrees in Electronics Engineering from Pusan National University in 1986 and 1988, respectively. He received his Ph.D. degree in Electrical Engineering from Korea Advanced Institute of Science and Technology in 1999. From 1988 to 1989, he worked for LG semiconductor, Ltd., Korea. From 1990 to 1998, he was a senior research staff in Electronics and Telecommunication Research Institute (ETRI), Korea. From 1999 to 2000, he was with Brain Science Research Center, KAIST. In 2000, he was with Brain Science Institute, RIKEN, Japan, as a research scientist. In 2001, he was an R&D manager of Extell Technology Corporation, Korea. Since 2002, he has been with the Department of Information Communication Engineering, Mokwon University, Daejon, Korea, and is now an associate professor. Also, he was with the Division of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, USA, as a visiting scholar from August 2008 to August 2009. His research interests are machine learning, speech signal processing, pattern recognition, and bioinformatics.

www.kci.go.kr