http://dx.doi.org/10.5392/IJoC.2012.8.4.030

Contour Plots of Objective Functions for Feed-Forward Neural Networks

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

ABSTRACT

Error surfaces provide us with very important information for training of feed-forward neural networks (FNNs). In this paper, we draw the contour plots of various error or objective functions for training of FNNs. Firstly, when applying FNNs to classifications, the weakness of mean-squared error is explained with the viewpoint of error contour plot. And the classification figure of merit, mean log-square error, cross-entropy error, and n-th order extension of cross-entropy error objective functions are considered for the contour plots. Also, the recently proposed target node method is explained with the viewpoint of contour plot. Based on the contour plots, we can explain characteristics of various error or objective functions when training of FNNs proceeds.

Keywords: *Feed-forward Neural Network, Error Function, Objective Function, Contour Plot.*

1. INTRODUCTION

Feed-forward neural networks (FNNs) can approximate any function with enough number of hidden nodes and this characteristic supports applications of FNNs to many fields[1]. Although there have been theoretical proofs for the capability of FNNs[1]-[4], how to train FNNs is still a challenging problem and various objective functions have been proposed to improve the training of FNNs[5]-[11]. These functions have peculiar properties and they can be compared various ways. The easiest way is to train FNNs using the objective functions for some application problems and to compare their performances. However, this performance comparison with simulation results is heuristic and problem-dependent.

Statistical comparisons are more concrete and there have been many researches of statistically analyzing error functions [12]-[15]. Firstly, there were analytic results that FNNs are Bayesian optimal classifiers if FNNs are trained with infinite number of training samples[12], [13]. Also, the optimal outputs of FNNs have been derived as a function of a posteriori probability that an input sample belongs to a certain class[12]-[14]. Furthermore, the relationship between two optimal outputs has been derived for two-class imbalanced data problems[15]. All these analytic results are functions of one argument and can be plotted as curves [13]-[15].

On the contrary, training of FNNs is done based on the objective functions. If FNNs are stuck in a local minima of the error objective functions, FNNs can scarcely improve performances[16]. If FNNs are in a flat region of the error objective functions with some high values, error decreases very slowly during training and it will take a very long time to escape that region. This period dominates learning time[17]. Thus, the contour of objective functions is very important for training of FNNs. In this paper, we draw contour plots of objective functions which are proposed to improve performances of FNNs. In section 2, we briefly introduce the various objective functions. For two dimensional plots of the contour, we assume that FNNs are with two output nodes. The drawings of contours for the objective functions are in section 3 with discussions. And section 4 concludes this paper.

* *Corresponding author, Email: shoh@mokwon.ac.kr*
Manuscript received Sep. 07, 2012; revised Oct 16, 2012; accepted Oct 26, 2012

2. OBJECTIVE FUNCTIONS

Consider a FNN with M output nodes whose activation functions are the bi-modal sigmoid. The traditional mean-squared error (MSE) objective function is defined by[5]

$$E_{MSE} = \sum_{k=1}^{M} (t_k - y_k)^2. \quad (1)$$

Here, y_k denotes the k-th output node of FNN and t_k is its desired value. In classification problems, $t_k = \pm 1$ where 1 is the desired value for a correct class node and -1 is for the other nodes. That is, if input samples are in the class c, $t_c = 1$ and $t_k = -1 (k \neq c)$. Parameters of FNNs are updated to decrease the error function during training[5]. For performance improvement of MSE, additive noise in the desired value is adopted as[6]

$$E_{AN} = \sum_{k=1}^{M} (t_k + n_k - y_k)^2, \quad (2)$$

where n_k is the zero-mean white noise with variance σ^2.

The classification figure of merit (CFM) objective function has three essential features that distinguish it from the traditional MSE[7]. Firstly, it has not the desired value but the index of output node representing the correct classification outcome. Secondly, CFM yields decreasing marginal "rewards" for increasingly ideal output pattern. Thirdly, CFM yields decreasing marginal "penalties" for increasingly bad misclassifications. This is to discourage the FNN from attempting to learn outliers heavily.

The resulting CFM objective function compares the output node value that should be at high state with values of all output nodes that should be at low state. It then goes through a sigmoidal function to each of these differences. Using CFM, learning focuses most heavily on the reduction of misclassifications rather than reducing the difference between the output node value and its target output. Thus, CFM is defined by[7]

$$O_{CFM} = \sum_{k \neq c} \frac{1}{1+\exp\left(-\beta\left(y_c - y_k\right)\right)}, \quad (3)$$

where y_c denotes the correct class node and $y_k (k \neq c)$ denotes the other output node. And training of FNN is done to increase O_{CFM}.

Let's consider a training of FNN based on MSE. When outliers are presented to the FNN, the difference between the output node value and its desired value is very large and this causes a heavy updating of parameters of FNN. Thus, outliers disturb learning for whole training samples and degrade the performance of FNNs. In order to suppress the huge updating of parameters by outliers, the mean log-square (MLS) error objective function is defined by[8]

$$E_{MLS} = \sum_{k=1}^{M} \ln\left(1 + \frac{1}{2}\left(t_k - y_k\right)^2\right). \quad (4)$$

In pattern recognition applications, the desired value of FNN is one of the two extreme values of sigmoidal activation function. If any output node value is near the wrong extreme value, we say the node is "incorrectly saturated"[10][17]. When an output node is incorrectly saturated, the gradient of MSE is small and this causes slow learning convergence[14]. In order to resolve this problem, the cross-entropy (CE) error objective function is defined by[9]

$$E_{CE} = -\sum_{k=1}^{M}\left[\left(1+t_k\right)\ln\left(1+y_k\right) + \left(1-t_k\right)\ln\left(1-y_k\right)\right]. \quad (5)$$

However, CE suffers from overspecialization for training samples since the gradient of CE for correctly saturated node is too strong. For improvement of CE, the *n*-th order extension of CE (*n*CE) error objective function is defined as[10]

$$E_{nCE} = -\sum_{k=1}^{M} \int \frac{t_k^{n+1}\left(t_k - y_k\right)^n}{2^{n-2}\left(1-y_k^2\right)} dy_k, \quad (6)$$

where *n* is a natural number.

Also, *n*CE is modified to attack imbalanced data problems, which is one of recent challenging problems[11]. Assume that the FNN has two output nodes in order to handle two-class imbalanced data problems. Class 1 denotes the minority class whose desired values are $t_1 = 1$ and $t_2 = -1$. Also, class 2 is the majority class whose desired values are $t_1 = -1$ and $t_2 = 1$. Here, y_k is called the target node of class *k*. Then, the objective function for so called "target node method" is[11]

$$E_{TN} = -\left[\int \frac{t_1^{n+1}\left(t_1 - y_1\right)^n}{2^{n-2}\left(1-y_1^2\right)} dy_1 + \int \frac{t_2^{m+1}\left(t_2 - y_2\right)^m}{2^{m-2}\left(1-y_2^2\right)} dy_2\right], \quad (7)$$

where *n* and *m* (*n*<*m*) are natural numbers.

3. CONTOUR PLOTS

In this section, we draw contours of the various objective functions introduced in section 2. For contour plots in the two-dimensional space, we assume that FNNs have two output nodes (*M=2*) whose desired values are $t_1 = 1$ and $t_2 = -1$. This means that an input sample belongs to the class 1. In this case, using the max rule for classification, y_1 should be greater than y_2 for correct classification. That is, the region $y_1 > y_2$ corresponds to a correct classification referred to as a "hit"[7]. On the contrary, the region $y_1 < y_2$ corresponds to an incorrect classification referred to as a "miss". This is shown in Fig. 1. Here, the straight line with slope 1 is the border line.

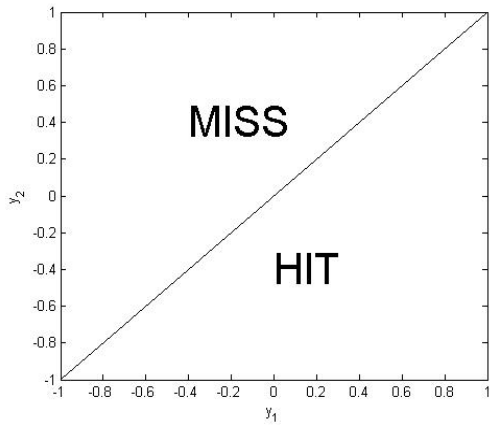

Fig. 1. A Hit-Miss diagram for a two-output FNN with class 1 input samples.

Using the assumption that

$$M = 2,\ t_1 = 1, \text{ and } t_2 = -1, \quad (8)$$

MSE in Eq. (1) is given by

$$E_{MSE} = \left(1 - y_1\right)^2 + \left(1 + y_2\right)^2, \quad (9)$$

whose contour is plotted in Fig. 2. Here, point (1,-1) corresponds to $E_{MSE} = 0$ and point (-1,1) has $E_{MSE} = 4$. In this figure, we can find that MSE at the point 'B'

($E_{MSE}(B)$) is greater than MSE at the point 'A' ($E_{MSE}(A)$) although the point 'A' is in the MISS region and the point 'B' is in the HIT region. During training of the FNN, a point (y_1, y_2) moves to a direction of decreasing MSE. However, decreasing MSE does not always give us a better classification as explained using the relationship $E_{MSE}(A) < E_{MSE}(B)$ [7]. Also, the movement of (y_1, y_2) from the MISS region to the HIT region may take long period, since the gradient of E_{MSE} in the MISS region is not steep. Here, the steepness is inversely proportional to the interval between lines of contour. The contour of E_{AN} given by Eq. (2) can be taken by adding noises to the contour of E_{MSE}. So, the weakness of E_{MSE} is inherited to E_{AN}.

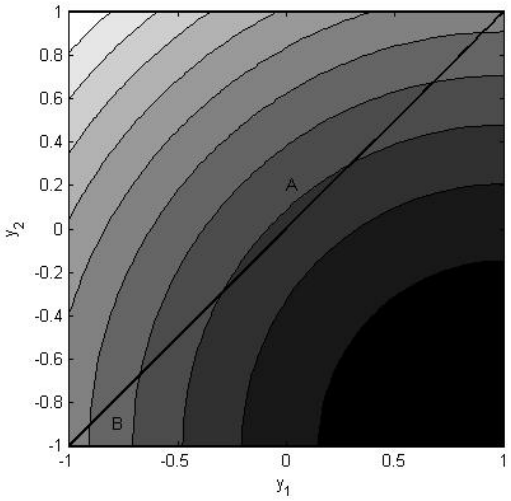

Fig. 2. A contour plot of MSE for a two-output FNN.

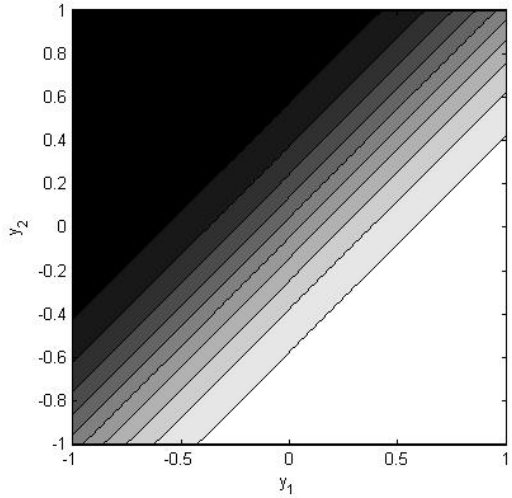

Fig. 3. A contour plot of CFM for a two-output FNN.

With $\beta = 4$, CFM in Eq. (3) is given by

$$O_{CFM} = \frac{1}{1+\exp\left(-4\left(y_1 - y_2\right)\right)} \qquad (10)$$

and its contour plot is in Fig. 3. On the contrary to MSE, the contour plot for CFM is composed of straight lines. Point (1,-1) is the highest point and (-1,1) is the lowest point. Using CFM, parameters of FNN are updated to the direction of increasing O_{CFM}. The gradient of O_{CFM} near the border line is very steep and point (y_1, y_2) moves to the HIT region rapidly. Also, the contour lines are parallel to the border line. However, near the point (-1,1), the gradient has a very gentle slope and it will take very long time to move from point (-1,1) to the region near the border line. This slow convergence of learning is the weakness of CFM.

Applying the same assumption of Eq. (8) to Eq. (4), MLS is given by

$$E_{MLS} = \ln\left(1+\frac{1}{2}\left(1-y_1^2\right)\right)\left(1+\frac{1}{2}\left(1+y_2^2\right)\right). \qquad (11)$$

Fig. 4 shows the contour of Eq. (11) which consists of straight lines in the "miss" region like CFM and resembles E_{MSE} in the "hit" region.

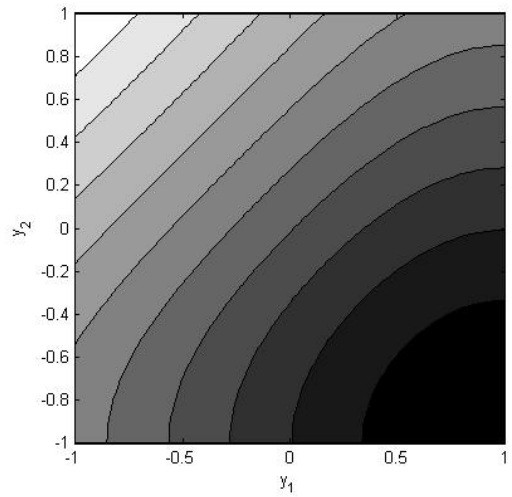

Fig. 4. A contour plot of MLS for a two-output FNN.

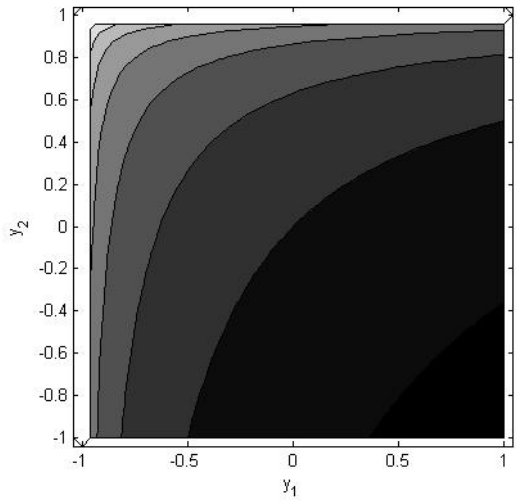

Fig. 5. A contour plot of CE for a two-output FNN.

Now, we consider the contour of CE given by

$$E_{CE} = -2\ln(1+y_1)(1-y_2), \tag{12}$$

which is derived from Eq. (5) using the assumption Eq. (8). Fig. 5 shows the contour of Eq. (12) that has steep slope in the "miss" region especially near the point (-1,1). Thus, FNN can escape the "miss" region rapidly during training based on E_{CE}. However, the gradient near the point (1,-1) has the effect of overspecialization for training samples[10].

The contour of *n*CE with *n*=2 is derived by substituting Eq. (8) into Eq. (6) as

$$E_{nCE}(n=2) = -\int \frac{1-y_1}{1+y_1} dy_1 + \int \frac{1+y_2}{1-y_2} dy_2 . \tag{13}$$

Using

$$\int \frac{1}{x} dx = \ln|x|, \tag{14}$$

we can get

$$\begin{aligned} E_{nCE}(n=2) &= -\left[2\ln|1+y_1| - (1+y_1)\right] + \left[(1-y_2) - 2\ln|1-y_2|\right] \\ &= u + v + 2\ln|uv| \end{aligned} \tag{15}$$

Here,

$$u = 1+y_1 \quad \text{and} \quad v = 1-y_2 . \tag{16}$$

Fig. 6 shows the contour of Eq. (15), which has steeper slope in the "miss" region. This has the acceleration effect of learning in the "miss" region. Also, the gentle slope in the "hit" region can prevent the overspecialization of learning. This gentle slope characteristic in the "hit" region is in the contour of CFM, too.

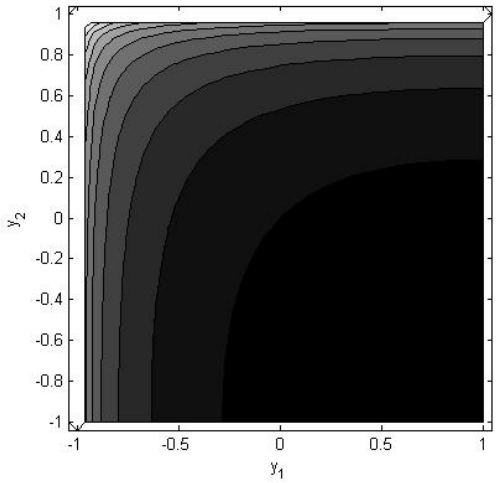

Fig. 6. A contour plot of *n*CE (*n*=2) for a two-output FNN.

In order to investigate the effect of parameter *n* in the *n*CE, we derive the contour of *n*CE with *n*=4 as

$$\begin{aligned} E_{nCE}(n=4) &= -\int \frac{(1-y_1)^4}{4(1-y_1^2)} dy_1 + \int \frac{(1+y_2)^4}{4(1-y_2^2)} dy_2 \\ &= -\frac{1}{4}\int \frac{(1-y_1)^3}{(1+y_1)} dy_1 + \frac{1}{4}\int \frac{(1+y_2)^3}{(1-y_2)} dy_2 . \end{aligned} \tag{17}$$

Using

$$\int \frac{x^3}{ax+b} dx = \frac{(ax+b)^3}{3a^4} - \frac{3b(ax+b)^2}{2a^4} + \frac{3b^2(ax+b)}{a^4} - \frac{b^3}{a^4}\ln(ax+b), \tag{18}$$

the contour is given by

$$E_{nCE}(n=4) = \frac{\frac{u^3+v^3}{3} - 3\left(u^2+v^2\right) + 12(u+v) - 8\ln|uv|}{4} . \tag{19}$$

Here, u and v are defined by Eq. (16) and the contour of *n*CE with *n*=4 is in Fig. 7. By comparing Fig. 7 with Fig. 6, we can find that increasing of *n* in *n*CE has effects of steeper gradient in the "miss" region and gentler gradient in the "hit" region. Thus, nCE accelerates the training of FNN by the steep gradient in the "miss" region and prevents the overspecialization to training samples by the gentle gradient in the "hit" region.

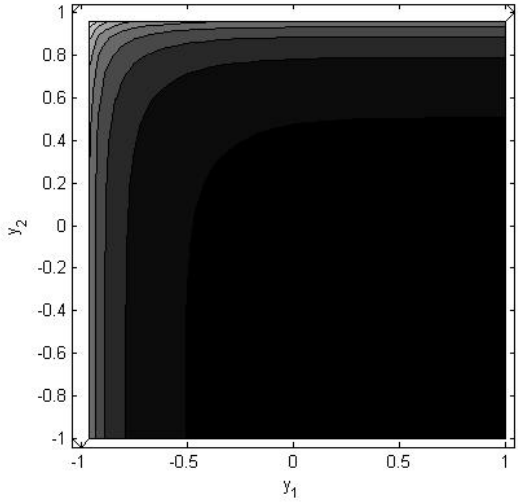

Fig. 7. A contour plot of *n*CE (*n*=4) for a two-output FNN.

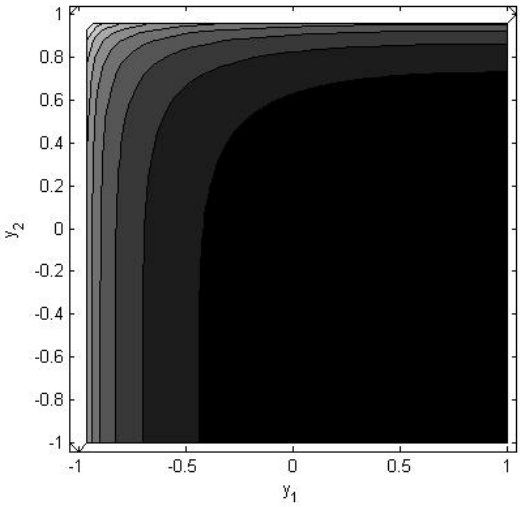

Fig. 8. A contour plot of the target node method (*n*=2, *m*=4).

Assuming Eq. (8), E_{TN} with *n*=2 and *m*=4 is given by

$$\begin{aligned} E_{TN}(n=2, m=4) &= -\int \frac{(1-y_1)^2}{(1-y_1^2)} dy_1 + \int \frac{(1+y_2)^4}{4(1-y_2^2)} dy_2 \\ &= -\int \frac{(1-y_1)}{(1+y_1)} dy_1 + \frac{1}{4}\int \frac{(1+y_2)^3}{(1-y_2)} dy_2 . \end{aligned} \tag{20}$$

By using Eqs. (14) and (18),

$$E_{TN} = u + \frac{v^3}{12} - \frac{3v^2}{4} + 3v - 2\ln|uv| . \qquad (21)$$

Here, u and v are the same with Eq. (16). The contour of Eq. (21) is plotted in Fig. 8. The gradient along to the horizontal direction from (-1,-1) to (1,-1) is gentler than that along to the vertical direction from (1,1) to (-1,1). This means that the error component of y_1 (the first term in Eq. (20)) is greater than the error component of y_2 (the second term in Eq. (20)). Therefore, the target node method strengthens updating amount of parameters related to the minority class target node y_1. This can prevent the boundary distortion from the majority class to the minority class[11].

Experimental comparisons of the objective functions for multi-class problems can be found in [7], [9], and [10]. Also, the comparisons for imbalanced two-class problems are in [11]. These experimental results can be complementary with the contour analysis for elucidating the characteristics of objective functions.

5. CONCLUSIONS

In this paper, we investigate the contours of various objective functions which were proposed to improve performances of FNNs. Decreasing of MSE objective function cannot guarantee the better classification performance. CFM shows steeper gradient near the border line between 'hit' and 'miss' regions. Although updating parameters to the direction of increasing CFM coincides the increasing of classification ratio, CFM shows slow convergence of learning due to the very little gradient in the 'miss' region. MLS shows straight contour lines in the "miss" region. CE shows stepper gradient in the "miss" region than MSE and CFM. In *n*CE, increasing *n* makes steeper gradient in the "miss" region for acceleration of learning and gentler gradient in the "hit" region for preventing overspecialization to training samples. In the target node method, learning related to the minority class is strengthened for preventing boundary distortion due to imbalanced data.

REFERENCES

[1] K. Hornik, M. Stinchcombe, and H. White, "Multilayer Feed-forward Networks are Universal Approximators," Neural Networks, vol.2, 1989, pp. 359-366.

[2] K. Hornik, "Approximation Capabilities of Multilayer Feedforward Networks," Neural Networks, vol.4, 1991, pp. 251-257

[3] S. Suzuki, "Constructive Function Approximation by Three-Layer Artificial Neural Networks," Neural Networks, vol.11, 1998, pp. 1049-1058

[4] Y. Liao, S. C. Fang, H. L. W. Nuttle, "Relaxed Conditions for Radial-Basis Function Networks to be Universal Approximators," Neural Networks, vol.16, 2003, pp. 1019-1028

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

[6] C. Wang and J. C. Princope, "Training Neural Networks with Additive Noise in the Desired Signal," IEEE Trans. Neural Networks, vol.10, 1999, pp. 1511-1517.

[7] J. B. Hampshare and A. H. Waibel, "A Novel Objective Function for Improved Phoneme Recognition Using Time-Delay Neural Networks," IEEE Trans. Neural Networks, vol.1, 1990, pp. 216-228.

[8] K. Liano, "Robust Error measure for Supervised Neural Network Learning with Outliers," IEEE Trans. Neural Networks, vol.7, 1996, pp. 246-250.

[9] A. van Ooyen and B. Nienhuis, "Improving the Convergence of the Backpropagation Algorithm," Neural Networks, vol.5, 1992, pp. 465-471.

[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] M. D. Richard and R. P. Lippmann, "Neural Network Classifier Estimate Bayesian a Posteriori Probabilities," Neural Computa., vol.3, 1991, pp. 461-483.

[14] S.-H. Oh, "Statistical Analyses of Various Error Functions for Pattern Classifiers," *Proc. Convergence on Hybrid Information Technology*, CCIS vol. 206, 2011, p. 129-133.

[15] S.-H. Oh, "A Statistical Perspective of Neural Networks for Imbalanced Data problems," Int. Journal of Contents, vol.7, 2011, pp. 1-5.

[16] N. Baba, "A New Approach for Finding the Global Minimum of Error Function of Neural Networks," Neural Networks, vol.2, 1989, pp. 367-373.

[17] Y. Lee, S.-H. Oh, and M. W. Kim, "An Analysis of Premature Saturation in Back-Propagation Learning," Neural Networks, vol.6, 1993, pp. 719-728.

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 and music signal processing, pattern recognition, and bioinformatics.