http://dx.doi.org/10.5392/IJoC.2015.11.2.057

Deriving a New Divergence Measure from Extended Cross-Entropy Error Function

Sang-Hoon Oh
Division of Information Communication Engineering
Mokwon University, Daejeon, 302-729, Korea

Hiroshi Wakuya
Graduate School of Science and Engineering
Saga University, Saga, 840-8502, Japan

Sun-Gyu Park
Division of Architecture
Mokwon University, Daejeon, 302-729, Korea

Hwang-Woo Noh
Department of Visual Design
Hanbat National University, Daejeon, 305-509, Korea

Jae-Soo Yoo
School of Information and Communication Engineering
Chungbuk National University, 362-763, Korea

Byung-Won Min, Yong-Sun Oh
Division of Information Communication Engineering
Mokwon University, Daejeon, 302-729, Korea

ABSTRACT

Relative entropy is a divergence measure between two probability density functions of a random variable. Assuming that the random variable has only two alphabets, the relative entropy becomes a cross-entropy error function that can accelerate training convergence of multi-layer perceptron neural networks. Also, the n-th order extension of cross-entropy (nCE) error function exhibits an improved performance in viewpoints of learning convergence and generalization capability. In this paper, we derive a new divergence measure between two probability density functions from the nCE error function. And the new divergence measure is compared with the relative entropy through the use of three-dimensional plots.

***Key words**: Cross-Entropy, The n-th Order Extension of Cross-Entropy, Divergence Measure, Information Theory, Neural Networks.*

1. INTRODUCTION

Multi-layer perceptron (MLP) neural networks can approximate any function with enough number of hidden nodes [1]-[3] and this increases applications of MLPs to wide fields such as pattern recognition, speech recognition, time series prediction, bioinformatics, etc. MLPs are usually trained with the error back-propagation (EBP) algorithm, which minimizes the mean-squared error (MSE) function between outputs and their desired values of MLP [4]. However, the EBP algorithm has drawbacks with slow learning convergence and poor generalization performance [5], [6]. This is due to the incorrect saturation of output nodes and overspecialization to training samples [6].

* *Corresponding author, Email: shoh@mokwon.ac.kr*
Manuscript received Apr. 13, 2015; revised May. 26, 2015; accepted Jun. 01, 2015

Usually, sigmoidal functions are adopted as activation functions of nodes in MLP. The sigmoidal activation function can be divided into a central linear region and two outer saturated regions. When an output node of MLP is in an extremely saturated region of the sigmoidal activation function opposite to a desired value, we say the output node is "incorrectly saturated." The incorrect saturation makes updating amount of weights small and consequently learning convergence becomes slow. Also, when MLPs are trained too much for training samples, this causes overspecialization of MLP to training samples and generalization performance for untrained test samples will be poor.

Cross-entropy (CE) error function accelerates the EBP algorithm through decreasing the incorrect saturation of output nodes [5]. Furthermore, the *n*-th order extension of cross-entropy (*n*CE) error function attains accelerated learning convergence and improved generalization capability by decreasing the incorrect saturation as well as preventing the overspecialization to training samples [6].

Information theory has done a great role in neural network community. For improved performance, information theoretic view provides many learning rules of neural networks such as minimum class-entropy, minimizing entropy, and feature extraction using information theoretic learning [7]-[11]. Also, information theory can be a basis for constructing neural networks [12]. The upper bound of probability of error was derived based on the Renyi's entropy [13]. Maximizing the information contents of hidden nodes can be developed for better performance of MLPs [14], [15]. In this paper, we focus on the relationship between relative entropy and the CE error function.

Relative entropy is a divergence measure between two probability density functions [16]. Assuming that a random variable has two alphabets, the relative entropy becomes cross-entropy (CE) error function which can accelerates the learning convergence of MLPs. Since *n*CE error function is an extension of CE error function, there must be a divergence measure corresponding to *n*CE error function as CE does. In this sense, this paper derives a new divergence measure from *n*CE error function. In section 2, the relationship between the relative entropy and CE is introduced. Section 3 derives a new divergence measure from *n*CE error function and compares the new divergence measure with the relative entropy. Finally, section 4 concludes this paper.

2. RELATIVE ENTROPY AND CROSS-ENTROPY

Consider a random variable **x** whose probability density function (p.d.f.) is $p(x)$. In the case that the p.d.f. of **x** is estimated with $q(x)$, we need to measure how accurate the estimation is. Therefore, the relative entropy is defined by

$$D(p\,||\,q)=\sum_{x}p(x)\log\frac{p(x)}{q(x)} \tag{1}$$

as a divergence measure between $p(x)$ and $q(x)$ [16]. Let's assume that the random variable **x** has only two alphabets 0 and 1, in which the probabilities are

$$p(x=1)=p \quad \text{and} \quad p(x=0)=1-p. \tag{2}$$

Also,

$$q(x=1)=q \quad \text{and} \quad q(x=0)=1-q. \tag{3}$$

Then,

$$\begin{aligned} D(p\,||\,q) &= p(0)\log\frac{p(0)}{q(0)}+p(1)\log\frac{p(1)}{q(1)} \\ &= (1-p)\log\frac{1-p}{1-q}+p\log\frac{p}{q} \\ &= E_{CE}-H(p). \end{aligned} \tag{4}$$

Here,

$$H(p)=-p\log p-(1-p)\log(1-p) \tag{5}$$

is the entropy of a random variable **x** with two alphabets and

$$E_{CE}=-p\log q-(1-p)\log(1-q) \tag{6}$$

is the cross-entropy. If we assume that 'q' corresponds to a real output value 'y' of MLP output node and 'p' corresponds to its desired value 't', we can define the cross-entropy error function as

$$E_{CE}=-t\log y-(1-t)\log(1-y). \tag{7}$$

Thus, the cross-entropy error function is one specific type of relative entropy assuming that a random variable has only two alphabets [15].

We can use the unipolar [0, 1] mode or bipolar [-1, +1] mode for describing node values of MLPs. Since 't' and 'y' corresponds to 'p' and 'q' respectively, they are in the range of [0, 1]. Thus, the relationship between relative entropy and cross-entropy error function is based on the unipolar mode of node values.

3. NEW DIVERGENC MEASURE FROM THE *n*-th ORDER EXTENSION OF CROSS-ENTROPY

The *n*-th order extension of cross-entropy (*n*CE) error function was proposed based on the bipolar mode of node values as [6]

$$E_{nCE}=-\int\frac{t^{n+1}(t-y)^{n}}{2^{n-2}(1-y^{2})}dy, \tag{8}$$

where n is a natural number. In order to derive a new divergence measure from *n*CE error function based on the relationship between relative entropy CE error function, we need an unipolar mode formulation of *n*CE error function. That is derived as

$$E_{nCE}=-\int\frac{(2t-1)^{n+1}(t-y)^{n}}{y(1-y)}dy. \tag{9}$$

We will derive new divergence measures from Eq. (9) with n=2 and 4.

When n=2, the nCE error function given by Eq. (9) becomes

$$E_{nCE}(n=2) = -\int \frac{(2t-1)^3(t-y)^2}{y(1-y)}dy = -(2t-1)^3[A+B], \quad (10)$$

where

$$A = \int\left(\frac{t^2}{y} - 2t + y\right)dy = t^2\ln y - 2ty + \frac{y^2}{2} \quad (11)$$

and

$$\begin{aligned} B &= \int\left(\frac{t^2}{1-y} - 2t\frac{y}{1-y} + \frac{y^2}{1-y}\right)dy \\ &= -t^2\ln y + 2t(\ln(1-y)+y) - \ln(1-y) - y - \frac{y^2}{2}. \end{aligned} \quad (12)$$

By substituting Eqs. (11) and (12) into Eq. (10),

$$E_{nCE}(n=2) = (2t-1)^3\left[y + (t-1)^2\ln(1-y) - t^2\ln y\right]. \quad (13)$$

In order to derive a new divergence measure corresponding to nCE(n=2), t and y are substituted to p and q, respectively. This is the reverse procedure for deriving Eq. (7) from (6) by substituting 'p' and 'q' to 't' and 'y', respectively. Then, we can get

$$E_{nCE}(n=2) = (2p-1)^3\left[q + (1-p)^2\ln(1-q) - p^2\ln q\right]. \quad (14)$$

Thus, by resembling the last equation in Eq. (4), the new divergence measure is derived by

$$\begin{aligned} F(p\,\|\,q;n=2) &= E_{nCE}(n=2) - K(p;n=2) \\ &= (1-2p)^3\left[(p-q) + (1-p)^2\ln\frac{1-p}{1-q} - p^2\ln\frac{p}{q}\right], \end{aligned} \quad (15)$$

where

$$K(p;n=2) = (2p-1)^3\left[p + (1-p)^2\ln(1-p) - p^2\ln p\right]. \quad (16)$$

When n=4, the nCE error function given by Eq. (9) is

$$\begin{aligned} E_{nCE}(n=4) &= -\int \frac{(2t-1)^5(t-y)^4}{y(1-y)}dy \\ &= -(2t-1)^5[C+D+E+F+G], \end{aligned} \quad (17)$$

where

$$C = \int\frac{y^3}{1-y}dy = -\ln(1-y) - y - \frac{y^2}{2} - \frac{y^3}{3}, \quad (18)$$

$$D = -4t\int\frac{y^2}{1-y}dy = 4t\left[\ln(1-y) + y + \frac{y^2}{2}\right], \quad (19)$$

$$E = 6t^2\int\frac{y}{1-y}dy = -6t^2[\ln(1-y)+y], \quad (20)$$

$$F = -4t^3\int\frac{1}{1-y}dy = 4t^3\ln(1-y), \quad (21)$$

and

$$G = t^4\int\frac{1}{y(1-y)}dy = t^4[\ln y - \ln(1-y)]. \quad (22)$$

Substituting Eqs. (18), (19), (20), (21), and (22) into Eq. (17),

$$E_{nCE}(n=4) = (2t-1)^5\begin{bmatrix} \ln(1-y) + y + \frac{y^2}{2} + \frac{y^3}{3} \\ -4t\left(\ln(1-y) + y + \frac{y^2}{2}\right) + 6t^2(\ln(1-y)+y) \\ -4t^3\ln(1-y) - t^4(\ln y - \ln(1-y)) \end{bmatrix}. \quad (23)$$

By substituting t and y to p and q, respectively, we can get

$$E_{nCE}(n=4) = (2p-1)^5\begin{bmatrix} \ln(1-q) + q + \frac{q^2}{2} + \frac{q^3}{3} \\ -4p\left(\ln(1-q) + q + \frac{q^2}{2}\right) + 6p^2(\ln(1-q)+q) \\ -4p^3\ln(1-q) - p^4(\ln q - \ln(1-q)) \end{bmatrix}. \quad (24)$$

Thus, by resembling the last equation in Eq. (4), the new divergence measure is derived by

$$\begin{aligned} &F(p\,\|\,q;n=4) = E_{nCE}(n=4) - K(p;n=4) \\ &= (1-2p)^5\begin{bmatrix} \ln\frac{1-p}{1-q} + (p-q) + \frac{p^2-q^2}{2} + \frac{p^3-q^3}{3} \\ -4p\left(\ln\frac{1-p}{1-q} + (p-q) + \frac{p^2-q^2}{2}\right) \\ +6p^2\left(\ln\frac{1-p}{1-q} + (p-q)\right) - 4p^3\ln\frac{1-p}{1-q} \\ -p^4\left(\ln\frac{p}{q} - \ln\frac{1-p}{1-q}\right) \end{bmatrix}, \end{aligned} \quad (25)$$

where

$$K(p;n=4) = (2p-1)^5\begin{bmatrix} \ln(1-p) + p + \frac{p^2}{2} + \frac{p^3}{3} \\ -4p\left(\ln(1-p) + p + \frac{p^2}{2}\right) \\ +6p^2(\ln(1-p)+p) - 4p^3\ln(1-p) \\ -p^4(\ln p - \ln(1-p)) \end{bmatrix}. \quad (26)$$

In order to compare the new divergence measures given by Eqs. (15) and (25) with the relative entropy given by Eq. (4), we plot them in the range that p and q are in [0,1]. Fig. 1 shows the three-dimensional plot of relative entropy $D(p||q)$. The x and y axes correspond to p and q, respectively, and z axis corresponds to $D(p||q)$. $D(p||q)$ is minimum of zero when $p=q$ and it increases when p goes far from q. Since $D(p||q)$ is a divergence measure, it is not symmetric.

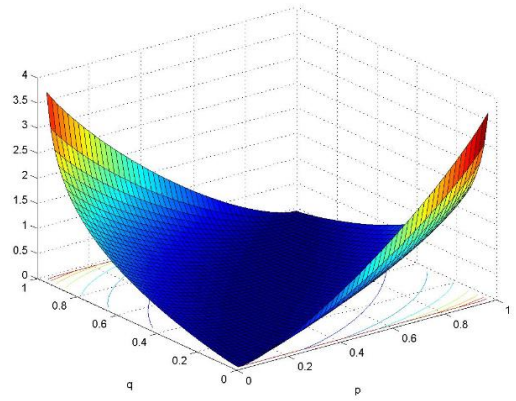

Fig. 1. The three-dimensional plot of relative entropy $D(p||q)$ with two alphabets

Fig. 2 shows the three-dimensional plot of new divergence measure $F(p||q;n=2)$ given by Eq. (15). $F(p||q;n=2)$ is minimum of zero when $p=q$ and it increases when p goes far from q as $D(p||q)$ does. Furthermore, we can find that $F(p||q;n=2)$ is more flat than $D(p||q)$. Also, the three-dimensional plot of $F(p||q;n=4)$ shown in Fig. 3 is minimum of zero when $p=q$ and more flat than $F(p||q;n=2)$ shown in Fig. 2. So, increasing the order n of the new divergence measure makes the new divergence measure more flat.

When applying MLPs to pattern classification, the optimal outputs of MLP based on various error functions were derived in [6] and [18]. We plot them in Fig. 4. The optimal output of MLP based on CE error function is a first order function of a posterior probability that a certain input sample belongs to a specific class. When using nCE error function with n=2 for training MLPs, as shown in Fig. 4, the optimal output of MLP shows more flat than the CE case. And, nCE error function with n=4 shows the optimal output more flat than CE and nCE with n=2 cases. The two-dimensional contour plots of CE and nCE error functions also show the same property [17]. So, we can argue that the property of divergence measures derived from CE and nCE coincides with the two-dimensional contour plot of CE and nCE error function in [17] and optimal outputs in [6] and [18].

4. CONCLUSIONS

In this paper, we introduce the relationship between relative entropy and CE error function. When a random variable has only two alphabets, the relative entropy becomes cross-entropy. Based on the relationship, we derive a new divergence measure form the nCE error function. Comparing the three-dimensional plot of relative entropy and new divergence measure when n=2 and 4, we can argue that the order n of new divergence measure has an effect of flatting the divergence measure. This property coincides with the previous results which comparing the optimal outputs and contour plots of CE and nCE.

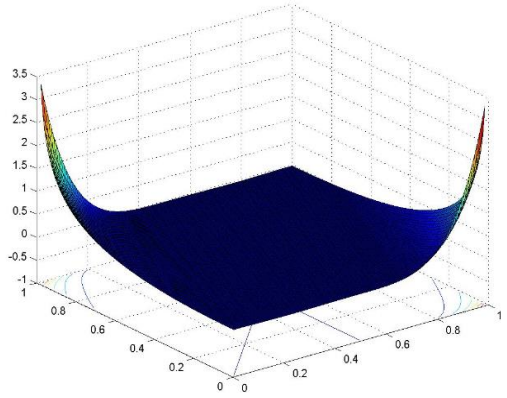

Fig. 2. The three-dimensional plot of new divergence measure with two alphabets when n=2, $F(p||q;n=2)$

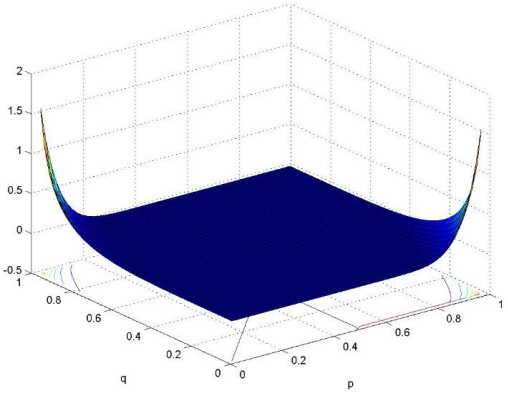

Fig. 3. The three-dimensional plot of new divergence measure with two alphabets when n=4, $F(p||q;n=4)$

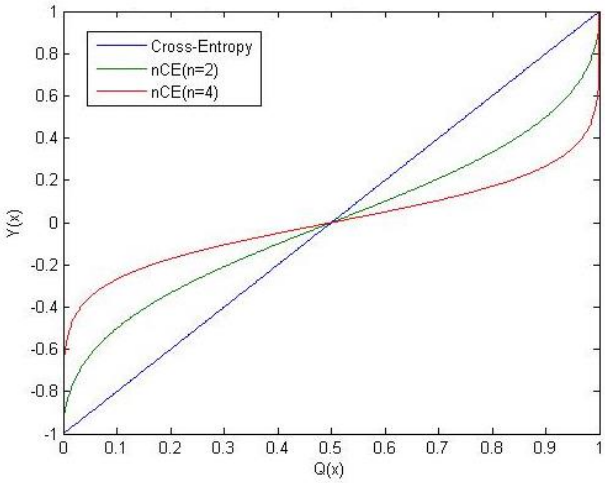

Fig. 4. Optimal outputs of MLPs. Here, Q(x) denotes a posteriori probability that a certain input x belongs to a specific class

ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2014K2A2A4001441)

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] D. E. Rumelhart and J. L. McClelland, *Parallel Distributed Processing*, Cambridge, MA, 1986.
[5] A. van Ooyen and B. Nienhuis, "Improving the Convergence of the Backpropagation Algorithm," Neural Networks, vol. 5, 1992, pp. 465-471.
[6] S.-H. Oh, "Improving the Error Back-Propagation Algorithm with a Modified Error Function," IEEE Trans. Neural Networks, vol. 8, 1997, pp. 799-803.
[7] A. El-Jaroudi and J. Makhoul, "A New Error Criterion for Posterior probability Estimation with Neural Nets," Proc. IJCNN'90, vol. III, Jun. 1990, pp. 185-192.
[8] M. Bichsel and P. Seitz, "Minimum Class Entropy: A maximum Information Approach to Layered Networks," Neural Networks, vol. 2, 1989, pp. 133-141.
[9] S. Ridella, S. Rovetta, and R. Zunino, "Representation and Generalization Properties of Class-Entropy Networks," IEEE Trans. Neural Networks, vol. 10, 1999, pp. 31-47.
[10] D. Erdogmus and J. C. Principe, "Entropy Minimization Algorithm for Multilayer Perceptrons," Proc. IJCNN'01, vol. 4, 2001, pp. 3003-3008.
[11] K. E. Hild II, D. Erdogmus, K. Torkkola, and J. C. Principe, "Feature Extraction Using Information-Theoretic Learning," IEEE Trans. PAMI, vol. 28, no. 9, 2006, pp. 1385-1392.
[12] S.-J. Lee, M.-T. Jone, and H.-L. Tsai, "Constructing Neural Networks for Multiclass-Discretization Based on Information Theory," IEEE Trans. Sys., Man, and Cyb.-Part B, vol. 29, 1999, pp. 445-453.
[13] D. Erdogmus and J. C. Principe, "Information Transfer Through Classifiers and Its Relation to Probability of Error," Proc. IJCNN'01, vol. 1, 2001, pp. 50-54.
[14] R. Kamimura and S. Nakanishi, "Hidden Information maximization for Feature Detection and Rule Discovery," Network: Computation in Neural Systems, vol. 6, 1995, pp. 577-602.
[15] K. Torkkola, "Nonlinear Feature Transforms Using Maximum Mutual Information," Proc. IJCNN'01, vol. 4, 2001, pp. 2756-2761.
[16] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, 1991.
[17] S.-H. Oh, "Contour Plots of Objective Functions for Feed-Forward Neural Networks," Int. Journal of Contents, vol. 8, no. 4, Dec. 2012, pp. 30-35.
[18] S.-H. Oh, "Statistical Analyses of Various Error Functions For Pattern Classifiers," CCIS, vol. 206, 2011, pp. 129-133.

Sang-Hoon Oh
He received his B.S. and M.S degrees in Electronics Engineering from Busan 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 a 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.

Hiroshi Wakuya
He received the B.E. degree in electronic engineering from Kyushu Institute of Technology, Kitakyushu, Japan, in 1989, and the M.E. and Ph.D. degrees in electrical and communication engineering from Tohoku University, Sendai, Japan, in 1991 and 1994, respectively. In 1994, he joined the staff of Saga University as a Research Associate. In 1995, he became a Lecturer. From 1999 to 2000, he was a Visiting Scientist in University of Louisville, Louisville, Kentucky, USA. Since 2004, he has been an Associate Professor. His research interests include neural networks, intelligent instrumentation, and biological engineering.

Sun-Gyu Park
He received the Ph.D. degree in the department of architecture from Tokyo University, Japan in September, 2004. He was employed by Mokwon University, Korea as a professor from March, 2009. His main research fields are an initial crack prediction of concrete technology development of the blast furnace slag concrete using alkali activator and a performance enhancement of the recycled aggregate concrete.

Hwang-Woo Noh

He received the B.S. and M.S. degrees from Department of Industrial Design, Hanbat National University, Daejeon, Korea, in 1996 and 2003 respectively. He has also completed his Ph.D. degree course at the Department of Industrial Design, Chungnam National University, Daejeon, Korea, in 2013. From 1997 to 2008, he worked as a representative of Design-Comphics Inc. He joined the faculty of Department of Visual Design, Hanbat National University, Daejeon, Korea, in 2009. During 2008-2009 he served as an executive director of Korea Design Industrial Association where he was nominated Head of the Daejeon-Chungcheong Branch. He is currently a professor of Hanbat National University, a Secretary-General of Daejeon Design Development Forum, and a Vice-president of Korea Contents Association. His main research interests include Visual Communication Design and its Fundamentals, Packaging Design, and Disaster Prevention Design.

Jae-Soo Yoo

He received his M.S. and Ph.D. degrees in Computer Science from the Korean Advanced Institute of Science and Technology, Korea in 1991 and 1995, respectively. He is now a professor in Information and Communication Engineering, Chungbuk National University, Korea. He has also been the president of Korea Contents Association since 2013. His main research interests include sensor data management, big data, and mobile social networks.

Byung-Won Min

He received M.S degree in computer software from Chungang University, Seoul, Korea in 2005. He worked as a professor in the department of computer engineering, Youngdong University, Youngdong, Chungbuk, Korea, from 2005 to 2008. He received Ph.D. degree in Information and Communication Engineering from Mokwon University, Daejeon, Korea., in 2010. His research interests include SaaS & Mobile Cloud, Database, and Software Engineering. Recently he is interested in Big Data processing and its applications.

Yong-Sun Oh

He received B.S., M.S., and Ph.D. degrees in electronic engineering from Yonsei University, Seoul, Korea, in 1983, 1985, and 1992, respectively. He worked as an R&D engineer at the System Development Division of Samsung Electronics Co. Ltd., Kiheung, Kyungki-Do, Korea, from 1984 to 1986. He joined the Dept. of Information & Communication Engineering, Mokwon University in 1988. During 1998-1999 he served as a visiting professor of Korea Maritime University, Busan, Korea, where he was nominated as a Head of Academic Committee of KIMICS an Institute. He returned to Mokwon University in 1999, and served as a Dean of Central Library and Information Center from 2000 to 2002, as a Director of Corporation of Industrial & Educational Programs from 2003 to 2005, as a Dean of Engineering College and as a Dean of Management Strategic Affairs from 2010 to 2013, respectively. He had been the President of KoCon from 2006 to 2012. During his sabbatical years, he worked as an Invited Researcher at ETRI from 2007 to 2008, and as a Visiting Scholar at KISTI from 2014 to 2015. His research interests include Digital Communication Systems, Information Theory and their applications. Recently he is interested in Multimedia Content and Personalized e-Learning.