IMPROVING THE GENERALIZATION PERFORMANCE OF THE MCE/GPD LEARNING

Hiroshi SHIMODAIRA *Jun ROKUI* *Mitsuru NAKAI*

School of Information Science
Japan Advanced Institute of Science and Technology
Tatsunokuchi, Ishikawa, 923–1292 JAPAN
http://www-ks.jaist.ac.jp/index.html

ABSTRACT

A novel method to prevent the over-fitting effect and improve the generalization performance of the Minimum Classification Error (MCE) / Generalized Probabilistic Descent (GPD) learning is proposed. The MCE/GPD method, which is one of the newest discriminative-learning approaches proposed by Katagiri and Juang in 1992, results in better recognition performance in various areas of pattern recognition than the maximum-likelihood (ML) based approach where a posteriori probabilities are estimated. Despite its superiority in recognition performance, it still suffers from the problem of over-fitting to the training samples as it is with other learning algorithms. In the present study, a regularization technique is employed to the MCE method to overcome this problem. Feed-forward neural networks are employed as a recognition platform to evaluate the recognition performance of the proposed method. Recognition experiments are conducted on several sorts of datasets. The proposed method shows better generalization performance than the original one.

1. INTRODUCTION

The idea of Minimum Classification Error (MCE) / Generalized Probabilistic Descent (GPD) learning was first proposed in 1992 by Katagiri and Juang [1] to establish a general learning framework for minimizing classification error of an arbitrary discriminant functions. In contrast to the maximum likelihood (ML) based learning which estimates probabilistic distributions of data based on a model, MCE/GPD learning adapts the parameters of the model on the basis of minimum classification error. Although a number of discriminative-learning algorithms have been proposed so far, the MCE/GPD learning is unique in the sense that it is applicable to arbitrary discriminant functions that are differentiable in respect to the parameters that are to be adapted. To be specific, it can be applied to discriminant functions that deal with variable record length of data like speech recognition.

The MCE/GPD learning has been applied successfully to various functions such as linear-discriminant functions, MLP (multi-layer perceptron), DTW (dynamic time warping) [2] and HMM (hidden Markov models) [3]. Since the MCE learning tries to minimize a loss function that corresponds to the number of classification error for given training data set, it still suffers from a problem of generalization ability for unseen data. In another word, over-fitting to the training data is inevitable.

In order to improve the generalization ability of the MCE learning, a regularization technique, which is widely used to solve ill-posed problems, is employed in this study.

2. MINIMUM CLASSIFICATION ERROR LEARNING

Let $g_k(\mathbf{x};\Lambda_k)$ be a discriminant function with positive value to discriminate a data of class Ω_k from the other classes, where $\mathbf{x} \in \mathcal{R}^D$ and Λ_k denote a vector in D-dimensional feature space and a set of parameters of the discriminant function, respectively. For an input vector $\mathbf{x}$, if the following equation holds

$$g_k(\mathbf{x};\Lambda_k) > g_i(\mathbf{x};\Lambda_i) \quad \text{for all } i \neq k \tag{1}$$

then $\mathbf{x}$ is classified to class Ω_k.

In the framework of MCE learning, misclassification measure for a given sample $\mathbf{x}$ of class Ω_k is defined as follows

$$d_k(\mathbf{x}) = -g_k(\mathbf{x};\Lambda_k) + \left[\frac{1}{C-1}\sum_{j,j\neq k} g_j(\mathbf{x};\Lambda_j)^\eta\right]^{1/\eta} \tag{2}$$

where C represents the number of classes and η is a positive constant. In an extreme case where η goes to infinity, the misclassification measure becomes

$$d_k(\mathbf{x}) = -g_k(\mathbf{x};\Lambda_k) + \max_{i\neq k} g_i(\mathbf{x};\Lambda_i). \tag{3}$$

Obviously $d_k(\mathbf{x}) \leq 0$ in case of correct classification, and $d_k(\mathbf{x}) > 0$ in case of misclassification.

Using the misclassification measure for a set of data $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_P\}$, the objective function to be minimized is defined as an empirical average cost function as given below

$$L_0(\Lambda|X) = \frac{1}{P}\sum_{p=1}^{P}\sum_{k=1}^{C} \ell(d_k(\mathbf{x}_p))1(\mathbf{x}_p \in \Omega_k). \tag{4}$$

Here $\Lambda = \{\Lambda_1, \Lambda_2, \cdots, \Lambda_C\}$ and $\ell(d)$ is a smooth loss function, for which the following sigmoid function is typically used

$$\ell(d) = \frac{1}{1+e^{-\xi(d+\theta)}}. \tag{5}$$

$1(\ )$ in (4) is an indicator function which has value of one when the argument is true and zero otherwise.

In order to minimize the objective function of (4), the well-known *gradient descent method* can be applied and the set of parameter of each discriminant function is adapted by the following rule:

$$\Lambda^{(t+1)} = \Lambda^{(t)} - \varepsilon \nabla L_0(\Lambda^{(t)}|X) \tag{6}$$

where $\Lambda^{(t)}$ denotes the parameter set at the t-th iteration and ε denotes the learning parameter of a positive small value.

If one employs an expected cost function $E[\ell(d(\mathbf{x}))]$ instead of the empirical cost function $L_0(\Lambda|X)$ of (4), the parameter updating rule which is called Generalized Probabilistic Descent (GPD) is given by

$$\Lambda^{(t+1)} = \Lambda^{(t)} - \varepsilon_t U \nabla \ell(d_k(\mathbf{x})). \tag{7}$$

Here U is a positive-definite matrix and ε_t is a small positive real number.

3. MODIFICATION OF THE MCE LEARNING

As is shown in (4), the MCE/GPD learning basically tries to minimize an empirical error [4]. Therefore, the MCE learning scheme suffers from the problem of over-fitting to the training dataset as it is with other training schemes. McDermott and Katagiri [2] proposed a method to adapt the slope parameter ξ in (5) to prevent the over-fitting effect. One of the drawbacks of this approach is that the relationship between ξ and the shape of decision boundary in the feature space is not clear.

In the present study, in order to improve generalization performance more directly than the previous method, a regularization technique [5] has been employed. In regularization, a penalty term $F(\Lambda)$ which is called a regularizer is added to the original objective function and the the new objective function $\tilde{L}(\Lambda)$ is give by

$$\tilde{L}(\Lambda|X) = L_0(\Lambda|X) + \gamma F(\Lambda). \tag{8}$$

The regularizer works as an constraint in the optimization problem, and it conveys a priori knowlegde about the target function that is to be learnt.

Tikhonov and Arsenin [5] proposed the class of Tikhonov regularizers, whose form is given by

$$F = \frac{1}{2}\sum_{r=0}^{R}\int_a^b h_r(x)\left(\frac{d^r y}{dx^r}\right)^2 dx \tag{9}$$

in which x, y denote the input, output variable, respectively, and $h_r(x) \geq 0$ for $r = 0, \ldots, R-1$ and $h_R(x) > 0$.

In the present study, as a simple case of the Tikhonov regularizer, we have employed the following empirical penalty term given in [6], [7], which is

$$F(\Lambda|X) = \frac{1}{2P}\sum_{k=1}^{C}\sum_{p=1}^{P}\sum_{i=1}^{D}\left(\frac{\partial^2 g_k(\mathbf{x}_p)}{\partial x_{pi}^2}\right)^2 \tag{10}$$

where $\mathbf{x}_p = (x_{p1}, x_{p2}, \ldots, x_{pD})$ represents the p-th training data in D dimensional space.

4. APPLICATION TO NEURAL NETWORKS

The proposed modified MCE (mMCE) learning given in (8) can be applied to arbitrary discriminant functions that are twice differentiable. In the present study, multi-layer perceptron type neural network is employed to evaluate the performance.

For the p-th training data $\mathbf{x}_p \in \mathcal{R}^D$, let $i_{pj}^{(m)}$ and $o_{pj}^{(m)}$ be the input and output of the j-th cell of layer m respectively. Then the input value of the j-th cell of layer m is given by

$$i_{pj}^{(m)} = \sum_{i=1}^{n_{m-1}} w_{ji}^{(m,m-1)} o_{pi}^{(m-1)} + \theta_j^{(m)}. \tag{11}$$

Here $w_{ji}^{(m,m-1)}$ is the connection weight between the j-th cell of layer m and the i-th cell of layer $m-1$, $\theta_j^{(m)}$ is a constant and n_m represents the number of cells in layer m. The output of each cell is given by

$$o_j^{(m)} = f(i_j^{(m)}) \tag{12}$$

where $f(\ )$ is a sigmoid function. In the classical error back-propagation (EBP) training [8], the object function, which is defined on the basis of least squared error (LSE), is given by

$$E_{sq} = \frac{1}{2}\sum_{p=1}^{P}\sum_{k=1}^{n_3}\left(t_{pk} - o_{pk}^{(3)}\right)^2, \tag{13}$$

in which three-layer network is assumed and t_{pk} denotes desired output for the k-th output cell against the p-th input $\mathbf{x}_p$.

On the other hand, in the proposed mMCE, the objective function is defined as follows

$$\tilde{L}(\Lambda|X) = \frac{1}{P}\sum_{p=1}^{P} L_{0p}(\Lambda|X) + \gamma\frac{1}{P}\sum_{p=1}^{P}\sum_{i=1}^{n_1} F_{pi}(\Lambda|X), \tag{14}$$

where

$$L_{0p}(\Lambda) = \sum_{i=1}^{n_3}\ell(d_i(\mathbf{x}_p))1(\mathbf{x}_p \in C_i), \tag{15}$$

$$F_{pi}(\Lambda) = \frac{1}{2}\sum_{k=1}^{n_3}\left(w_{kj}^{(32)}(w_{ji}^{(21)})^2 f''(i_{pj}^{(2)})\right)^2. \tag{16}$$

The weight updating rule is given by

$$\Delta w_{pji}^{(m,m-1)} = \frac{\partial L_{0p}(\Lambda)}{\partial w_{ji}^{(m,m-1)}} + \gamma\frac{\partial F_p}{\partial w_{ji}^{(m,m-1)}}. \tag{17}$$

In the output layer where $m = 3$,

$$\frac{\partial L_{0p}}{\partial w_{kj}^{(32)}} = \ell'(d_k(\mathbf{x}_p))\frac{\partial d_k(\mathbf{x}_n)}{\partial i_{pk}^{(3)}} o_{pj}^{(2)} 1(\mathbf{x}_p \in C^k), \tag{18}$$

$$\frac{\partial F_{pi}}{\partial w_{kj}^{(32)}} = \frac{1}{2} w_{ji}^{(21)} f''(i_{pj}^{(2)}) Q_{pki}, \tag{19}$$

where

$$Q_{pki} = \sum_{j'=1}^{n_2} w_{kj'}^{(32)} {w_{j'i}^{(21)}}^2 f''(i_{pj'}^{(2)}). \tag{20}$$

In the hidden layer where $m = 2$,

$$\frac{\partial L_{0p}}{\partial w_{ji}^{(21)}} = \sum_{k=1}^{n_3} \left(\frac{\partial L_p}{\partial i_{pk}^{(3)}} w_{kj}^{(32)} \right) \frac{\partial \ell(d_j(i_{pj}^{(2)}))}{\partial i_{pj}^{(2)}} o_{pi}^{(1)}, \tag{21}$$

$$\frac{\partial F_{pi}}{\partial w_{ji'}^{(21)}} = \frac{1}{2} \Big(2\delta_{ii'} f''(i_{pj}^{(2)}) w_{ij}^{(21)} + i_{pi'}^{(1)} {w_{ij}^{(21)}}^2 \Big[(1 - 2f(i_{pj}^{(2)})) f''(i_{pj}^{(2)}) - 2{f'(i_{pj}^{(2)})}^2 \Big] \Big) \sum_{k=1}^{n_3} w_{kj}^{(32)} Q_{pki}. \tag{22}$$

Here $\delta_{ii'}$ is the Kronecker delta.

5. EXPERIMENTS

Performance evaluation was conducted on several types of datasets in UCI machine learning repository [9] and ATR speech database [10].

In order to compare the performance of the proposed method with other learning methods, the EBP based neural networks, the original MCE based neural networks, and Bayes discriminant functions where a single normal distribution (full covariance) is assumed for each category were applied on the same datasets. Three-layer feed-forward neural networks were employed for the experiments, the parameter γ in (8) was set to 0.01 and the slope parameter ξ in (5) was set to 1.0.

Since the MCE and mMCE learning are computationally expensive, the initial parameters used in the parameter updating rule of (6) were set to the one obtained by the EBP learning.

A. Results for Two-Class Problems

Preliminary experiments were, at first, performed for two-class problems on the UCI datasets "cancer", "house" and "sonar". Each dataset was divided into two groups, one was used for training and the other was used for testing.

The experimental results are summarized in Table 1. It can be seen that mMCE gives better test-set performance than the original MCE for each dataset.

Fig. 1 shows the correct classification rates in terms of the slope parameter ξ in (5). Although ξ influences the correct classification rate, mMCE performs better than MCE for any value of ξ.

In the framework of regularization, it is still an open problem to determine the appropriate weighting parameter γ in (8). As it can be seen in Fig. 2 where classification performance in terms of the parameter γ is shown, the classification performance is not sensitive to the the parameter γ.

Table. 1: Correct classification rates [%] in two-class problems

Dataset			Method			
	#samples		Bayes/ ML	NN/ EBP	NN/ MCE	NN/ mMCE
Cancer	training	420	95.0	99.3	97.8	96.3
	testing	279	95.7	91.8	92.5	94.6
House	training	265	98.3	99.6	98.5	98.5
	testing	170	96.4	95.3	98.2	99.4
Sonar	training	141	100.0	98.6	98.6	90.1
	testing	67	74.6	79.1	86.6	92.5

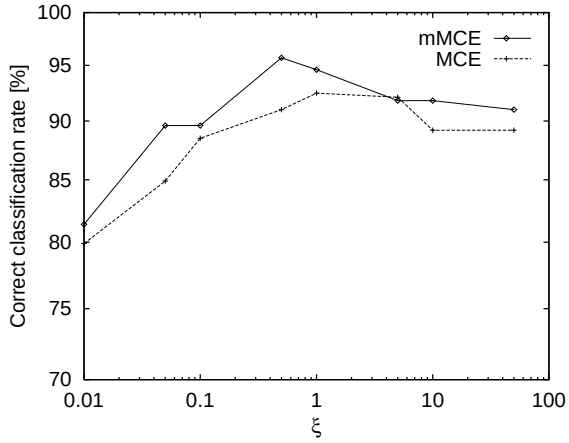

Figure. 1: Classification performance for the test-set "cancer" in terms of the slope parameter ξ in (5)

In multi-layer neural networks, it is well-known that the number of hidden nodes affects the generalization performance. The classification performance for the test-set "cancer" with respect to the number of hidden nodes is shown in Fig. 3. Although the performance varies with the number of hidden nodes, mMCE always shows better performance than the original MCE.

B. Results for speech data

In order to evaluate the performance for speech recognition, speech database "isolet" (isolated alphabet letters) of the UCI repository, and "vowels" (Japanese five vowels) made of the ATR continuous speech database "Set-B" were collected. In the "iso-

Table. 2: Speech datasets and network architecture

Dataset	#classes	#attributes	#hidden nodes
isolet(UCI)	26	617	32
vowels(ATR)	5	12	12

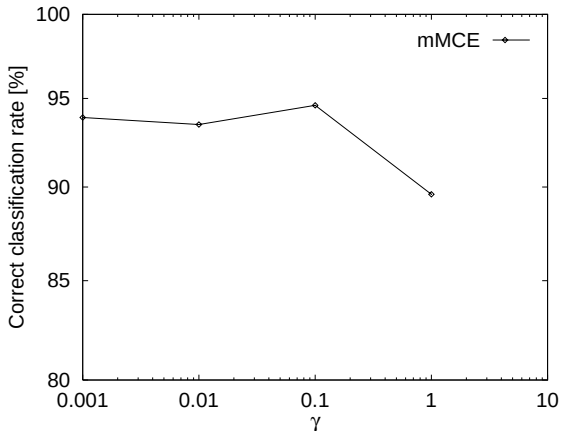

Figure. 2: Classification performance for the test-set "cancer" as a parameter of γ

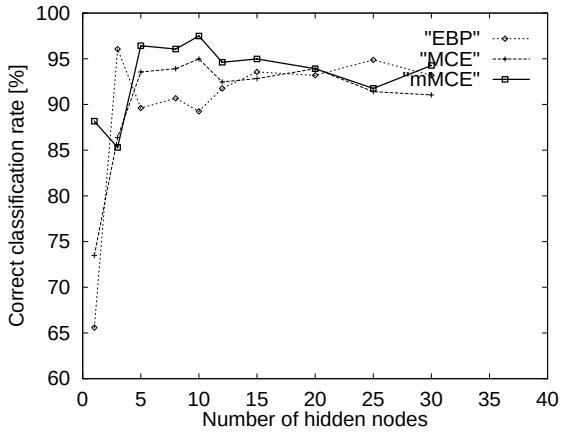

Figure. 3: Correct classification rate as a parameter of the number of hidden nodes ("cancer")

let" database, the data file "isolet1+2+3+4" was used for training and "isolet5" was used for testing. The database "vowels" was created for this research purpose by extracting 100 samples of each vowel uttered by each subject from the ATR database containing the uttered voice of six subjects. The data of four subjects (msh, mmy, mht, mho) were used for training and the data of remained two subjects (myi, mtk) were used for testing. Some information about the datasets and the number of hidden nodes used in the experiment is shown in Table 2.

Table 3 shows the correct classification rate for both the training and test sets. The proposed mMCE gives better test-set classification performance than the original MCE.

6. CONCLUSION

A regularization technique has been proposed to improve the generalization performance of the MCE/GPD learning. The simplified Tikhonov type regularizer, which takes the power of the second derivatives of the discriminant functions, has been employed as a regularizer in the present study. It should be noted that the employed regularizer is not case specific but general, apart from neural networks, the proposed modified MCE (mMCE) learning can be applied to various type of recognizers like HMM (hidden Markov models) and so on.

Table. 3: Correct classification rates [%] for speech datasets

Dataset			Method			
	#samples		Bayes/ ML	NN/ EBP	NN/ MCE	NN/ mMCE
isolet(UCI)	training	6238	-	93.4	96.9	96.2
	testing	1559	-	94.3	95.5	96.4
vowels(ATR)	training	4000	86.3	89.0	92.7	91.7
	testing	1000	79.3	82.1	88.6	89.1

7. REFERENCES

1. B-H. Juang and S. Katagiri. Discriminative learning for minimum error classification. *IEEE Trans. Signal Processing*, 40(12):3043–3054, 1992.
2. Eric McDermott and Shigeru Katagiri. Prototype-based minimum classificatikn error / generalized probabilistic descent training for various speech units. *Computer Speech and Language*, pages 351–368, August 1994.
3. Biing-Hwang Juang, Wu Chooud, and Chin-Hui Lee. Minimum classification error rate methods for speech recognition. *IEEE Trans. Speech and Audio Processing*, 5(3):257–265, 1997.
4. Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.
5. A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill-Posed Problems*. V. H. Winston, 1977.
6. Christopher M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.
7. Christopher M. Bishop. Curvature-Driven Smoothing: A Learning Algorithm for Feed-forward Networks. *IEEE Trans. Neural Networks*, 4(5):882–884, 1993.
8. D. E. Rumelhart, G. E. Hinton, and R. J. Willams. Learning representations by back-propagation errors. *Nature 323 9*, 323(9):533–536, October 1986.
9. C. J. Merz and P. M. Murphy. UCI repository of machine learning databases, 1996. http://www.ics.uci.edu/∼mlearn/MLRepository.html.
10. H. Kuwabara, K. Takeda, Y. Sagisaka, S. Katagiri, S. Morikawa, and T. Watanabe. Construction of a large-scale Japanese speech database and its management system. *Proc. of Intl. Conferece on Acoustics, Speech, and Signal Processing (ICASSP-89)*, pages 560–563, 1989.