

Paper:

Rapid Discriminative Learning

Jun Rokui

Department of the Interdisciplinary Faculty of Science and Engineering, Shimane University
1060 Nishikawatsu-cho, Matsue-shi 690-8504, Japan
E-mail: rokui@cis.shimane-u.ac.jp

[Received July 29, 2003; accepted December 1, 2003]

This paper presents MCE/GPD using GPD that is known as a highly effective discriminative learning method. MCE/GPD is an excellent recognition method that is applicable especially to speech recognition, since it excels in recognizing performance and can be used to deal with variable-length vectors. MCE/GPD involves a problem of calculation resulting from complicated algorithms making it impractical. In this paper, we propose a learning method to increase speed at learning based on a hierarchical model. We used a hierarchical neural network to evaluate the method's performance.

Keywords: MCE/GPD, MBL, generalized ability, margin, neural networks

1. Introduction

For subjects in statistical pattern recognition learning, the two major standards for learning are the likelihood maximization standard and the recognition error minimization standard. In classic pattern recognition, Bayes' discrimination is the best likelihood maximization standard in error loss minimization if the probability statistic of the recognition subject category is already known. Accordingly, a standard used when performing a course of learning for preparing a distribution that represents the shape of the subject category most satisfactorily is said to be the likelihood maximization standard.

In contrast, the recognition error minimization standard proposed in 1992 by Juang and Katagiri [1, 2]. In discriminative learning based on the recognition error minimization standard, we propose minimum classification error learning (MCE/GPD). Introducing MCE/GPD enables us to perform parameter learning of discriminative functions by descent, providing that the differential loss function is defined and that a problem of pattern recognition is formulated by the framework for a problem of optimization in continuous functions. It has also enabled us to deal with variable-length characteristic vectors. MCE/GPD is therefore used widely, especially in the field of speech recognition. Compared to conventional learning methods based on least square error or the likelihood maximization standard, MCE/GPD has proved to show a higher recog-

niton rate.

In terms of efficiency, MCE/GPD has a problem with computational complexity. In conventional pattern recognition, recognition performance is been improved by clustering based on a learning hypothesis, complicating models and increasing computational complexity. Studies have been promoted on boosting learning [3] and bagging learning [4] to expand flexibility or improve efficiency. These methods are intended to improve recognition performance by dividing learning data into two groups, those recognized easily and those recognized with difficulty. Their common features are that they attain high recognition performance and that they involve less computational complexity.

We propose classifiers based on MCE/GPD. In the first stage, models are formed with recognition error data. In the evaluation stage, these models are evaluated simultaneously with those prepared with correct data. Models that contain information concerning recognition error data are then recognized. Correct data and recognition error data must be organized into models hierarchically. The major concern is setting up class distinction standards for mutually connecting hierarchically clustered models. To narrow down the degrees of freedom for clustering distinction, we propose connecting hierarchical models upon a hypothesis for regular distribution.

2. MCE/GPD Configuration

For conventional discriminative learning, the recognition rate represents a discontinuous function in connection with set A of parameters. Accordingly, no effective solution is furnished even if a learning problem is formulated as an optimization problem.

In MCE/GPD, an evaluation function representing an error recognition rate is given as a continuous function for a parameter set, solving the problem of optimization by ordinary descent. We assume that the discriminative function used in discriminating pattern vector $\mathbf{x} \in R^D$, which belongs to class i , ($i = 1, \dots, M$), is described with $g_i(\mathbf{x}, \Lambda_i)$. Then Λ_i is a parameter set of discriminative functions.

For function representation of the error recognition rate in MCE/GPD, the error classification standard below is defined as a standard representing excellent recognition

against pattern vector \mathbf{x} .

$$h_i(\mathbf{x}, \Lambda_i) = -g_i(\mathbf{x}, \Lambda_i) + \left(\frac{1}{M-1} \sum_{j, j \neq i} g_j(\mathbf{x}, \Lambda_j) \right)^{1/\eta} \dots \dots \dots (1)$$

Here, $\eta > 0$ is a smoothing parameter. Especially when $\eta \rightarrow \infty$, the above equation becomes

$$h_i(\mathbf{x}, \Lambda_i) = -g_i(\mathbf{x}, \Lambda_i) + g_{j^*}(\mathbf{x}, \Lambda_{j^*}) \dots \dots \dots (2)$$

where j^* stands for the class number of the discriminative function having a maximum of all discriminative functions, with the exception of class i . As seen from the definition formulated here, $h_i(\cdot) < 0$ when recognition is correct; $h_i(\cdot) \geq 0$ when recognition is erroneous.

To represent loss from erroneous recognition, the error classification standard is converted to a loss function. Loss function $l(h)$ is an arbitrary function expressed as $l: R \rightarrow [0, 1]$. The sigmoid function shown in the equation below is generally used.

$$l(h) = \frac{1}{1 + e^{-\xi(h+\alpha)}}, \xi > 0 \dots \dots \dots (3)$$

Accordingly, the experimentally expected value of $L_0(\Lambda)$ of the loss function against $X = \{\mathbf{x}_1, \dots, \mathbf{x}_p\}$, i.e., a set of learning material consisting of p -pattern vectors, can be given by the following equation.

$$L_0(\Lambda) = \frac{1}{p} \sum_{p=1}^p \sum_{i=1}^M l(h_i(\mathbf{x}_p, \Lambda_i)) 1(\mathbf{x}_p \in C^i) \dots \dots (4)$$

where C^i is the pattern set of class i and $1(\cdot)$ is a characteristic function defined by the equation below.

$$1(\varepsilon) = \begin{cases} 1, & \text{phenomenon } \varepsilon \text{ is true} \\ 0, & \text{others} \end{cases} \dots \dots (5)$$

Accordingly, the minimization problem of the error recognition rate is defined as a minimization problem of evaluation function $L_0(\Lambda)$, which is smooth against parameter Λ , and the method of the steepest descent or a general solution for the minimization problem is used as shown in the equation below.

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

3. Higher-speed Learning Based on Recognition Error Minimization Standards

For the conventional pattern recognition method, recognition performance has been improved by providing an objective class and by clustering based on correct data upon a learning hypothesis. In this section, we propose adopting recognition error data as an essential part of recognition without differentiating it specifically as learning correct data. Thus MCE/GPD is introduced as a framework for learning with multiple classifiers, not with a single classifier.

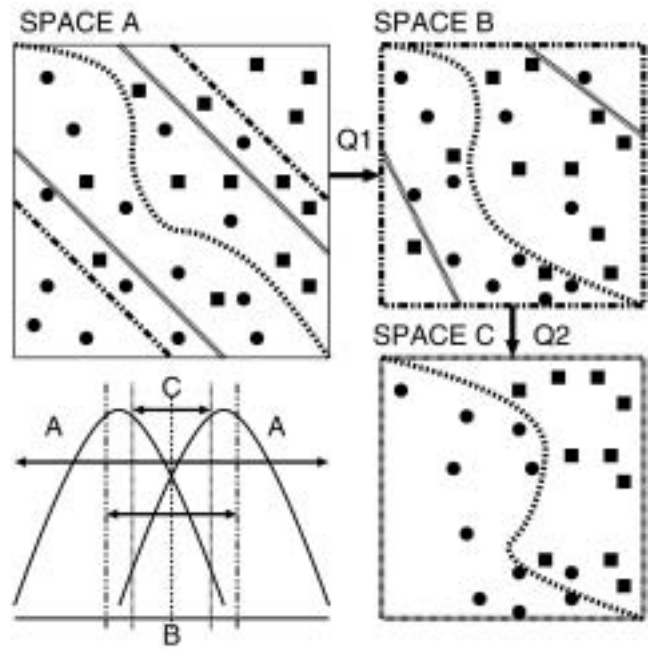


Fig. 1. Basic structure of MBL

3.1. Model Classification Algorithms

Discriminative learning such as MCE/GPD may involve the problem of overlearning, in which the recognition rate decreases by the effect of evaluation data since the discriminative boundary is overdifferentiated into learning data specifically. This is due to poor convergence and to a decrease in the region for generalization ability, in which the region for sampling that can be easily separated is affected by the region for sampling that is separated only with difficulty. To avoid the undesirable influence of overlearning, a number of cases have been made for a regular learning. For the improvement of generalization performance, studies related to mathematical planning theory and to the Support Vector Machine (SVM) based on functional analysis [5] have attracted particular attention for the last several years. SVM improves generalization performance by providing an appropriate margin or minimum distance between the hyperplane and the learning sample for linear separation. Space between the optimum discriminative boundary and the margin are assumed to be separated with difficulty, while space other than the above can be separated with ease. It is then possible to increase the speed in the total learning process by learning easy space at a comparatively higher speed and difficult space rather closely. The margin maximization in SVM is effective for high generalization performance, but involves a problem related to low learning efficiency resulting from the effect of a Kernel trick. To increase speed at learning in this study, recognition models are formed as shown in Fig.1 based on recognition error minimization standards, and models thus formed are arranged hierarchically. A series of procedures such as those above results in a