MINIMUM ERROR CLASSIFICATION WITH GEOMETRIC MARGIN CONTROL

Hideyuki Watanabe\textsuperscript{1}, Shigeru Katagiri\textsuperscript{2}, Kouta Yamada\textsuperscript{2}, Erik McDermott\textsuperscript{3}, Atsushi Nakamura\textsuperscript{3}, Shinji Watanabe\textsuperscript{3}, and Miho Ohsaki\textsuperscript{2}

\textsuperscript{1} MASTAR Project, National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289 Japan
E-mail: hideyuki.watanabe@nict.go.jp
\textsuperscript{2} Graduate School of Engineering, Doshisha University
1-3 Tatara Miyakodani, Kyotanabe City, Kyoto 610-0394 Japan
\textsuperscript{3} NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237 Japan

\textbf{ABSTRACT}

Minimum Classification Error (MCE) training, which can be used to achieve minimum error classification of various types of patterns, has attracted a great deal of attention. However, to increase classification robustness, a conventional MCE framework has no practical optimization procedures like geometric margin maximization in Support Vector Machine (SVM). To realize high robustness in a wide range of classification tasks, we derive the geometric margin for a general class of discriminant functions and develop a new MCE training method that increases the geometric margin value. We also experimentally demonstrate the effectiveness of our new method using prototype-based classifiers.

\textbf{Index Terms}— discriminative training, Minimum Classification Error, MCE, margin, geometric margin

\section{1. INTRODUCTION}

In statistical pattern recognition, the ultimate goal of classifier design is to estimate the Bayes error, i.e., to attain the status of minimum classification error probability over an infinite number of input patterns [1]. Discriminative training (DT) of pattern classifiers has been extensively studied to meet this goal [2, 3, 4, 5, 6]. DT has become the main current of pattern recognition research.

Among DT embodiments, the Minimum Classification Error (MCE) training method [2, 3], which has been mainly developed in the time series processing research field, is extensively used as a training method to realize high classification accuracy. MCE directly estimates the Bayes error for various types of patterns, including variable-length patterns, and is applicable to various kinds of discriminant functions. However, to increase the classification robustness (generalization capability), MCE has no practical optimization procedures and instead relies on such indirect mechanisms as the control of loss function smoothness [3].

A key measure to directly represent classification robustness is the distance between a class decision boundary and its closest training sample, which is called the “geometric margin” [6]. In fact, Support Vector Machine (SVM) [5, 6], a standard DT technology that mainly classifies fixed-dimensional patterns, increases robustness by maximizing the geometric margin. In SVM, however, the geometric margin is derived for a limited class of linear discriminant functions. Moreover, especially in realistic cases that are not linearly separable, SVM does not directly link the loss (hinge loss) and misclassification counts; as a result, there is no clear relationship between the optimal status of the classifier and the Bayes error.

In this paper, we newly formulate the geometric margin for a general class of discriminant functions to which MCE can be easily applied and then develop a new MCE training method that increases the geometric margin. High classification accuracy for unknown patterns is expected to be realized in various classification tasks. We also experimentally demonstrate the effectiveness of the new method using prototype-based classifiers that are standard classifiers with very high versatility.

\section{2. MINIMUM CLASSIFICATION ERROR TRAINING}

\subsection{2.1. Formulation Incorporating a Decision Rule}

For discussion simplicity, we consider the task of classifying a pattern $\mathbf{x}$ that belongs to the fixed-dimensional vector pattern space $\mathcal{X}$ as one of the $J$ classes ($C_j; j = 1, ..., J$). We assume that a classifier consists of trainable parameter set $\Lambda$, and a set of $N$ training samples, $\mathcal{X}_N = \{\mathbf{x}_n\}_{n=1}^N$, is given for adjusting $\Lambda$.

The ultimate design goal here is to estimate $\Lambda$ that minimizes the following classification error count risk:

$$R(\Lambda) = \sum_{y=1}^{J} \int_{\mathcal{X}} p_Y(y|\mathbf{x}) \mathbb{1}\{y \neq \arg \max_j p_X(C_j|\mathbf{x})\} d\mathbf{x},$$

where we assume that the Bayes decision rule, which classifies an input pattern as the class with the maximum a posteriori probability, is adopted and true a posteriori probability for $C_j$ can be expressed as the form of $p_X(C_j|\mathbf{x})$ [1]. $1(\mathcal{P})$ denotes an indicator function that returns 1 when proposition $\mathcal{P}$ is true and 0 otherwise. However, since the true functional forms of the a posteriori probabilities are rarely known in reality, the estimation of Eq. (1) is unrealistic. Therefore, we need a practical method to estimate the classification error count risk even when the true functional forms of a posteriori probabilities are unknown.

MCE training provides a practical solution to the above problem based on the following expression [2]:

$$R(\Lambda) \simeq \sum_{y=1}^{J} \int_{\mathcal{X}} p_Y(y|\mathbf{x}) \ell(d_y(\mathbf{x}, \Lambda)) d\mathbf{x},$$

This work was supported in part by Grant-in-Aid for Scientific Research (B), No. 19300064.
where \(d_y(x, \Lambda)\) is called a misclassification measure and is constructed using discriminant functions \(g_j(x, \Lambda)\) \((j = 1, ..., J)\) and is not limited to a posteriori probabilities as follows \((\psi > 0)\):

\[
d_y(x, \Lambda) = -g_y(x, \Lambda) + \log \left[ \frac{1}{J} \sum_{j \neq y} e^{\psi g_j(x, \Lambda)} \right]^{1/\psi}.
\]  

(3)

In Eq. (3) we assume that \(x\) belongs to \(C_y\) and the classification decision rule maps \(x\) to the class with a maximum value of \(g_j(x, \Lambda)\). The misclassification measure in Eq. (3), which is differentiable in \(\Lambda\), is an approximated expression of the following measure:

\[
d_y(x, \Lambda) = -g_y(x, \Lambda) + \max_{j \neq y} g_j(x, \Lambda)
\]

(4)

which emulates the correctness/incorrectness of the classification decision by the sign of the scalar value\(^1\). Furthermore, \(\ell(\cdot)\) is a smoothed classification error count loss generally defined as

\[
\ell(d_y(x, \Lambda)) = \frac{1}{1 + \exp(-ad_y(x, \Lambda))} \quad (a > 0).
\]

(5)

\(\ell(\cdot)\) is differentiable in \(\Lambda\) and is an approximated expression of indicator function \(1(\cdot)\) in Eq. (1).

The approximation precision of Eq. (1) by Eq. (2) can be flexibly controlled by functional smoothness in Eqs. (3) and (5) through the adjustment of \(\psi\) and \(a\). In principle, MCE training accurately approximates the ideal classification error count risk by Eq. (2), which incorporates the whole classification decision process in a scalar functional form and is differentiable in \(\Lambda\). As a result, the status of \(\Lambda\) corresponding to the minimum risk can be easily calculated by a gradient search of variable \(\Lambda\) in Eq. (2).

The above functional smoothness in MCE training readily leads one to the use of standard gradient search procedures, and also it enables the classifier to be robust to unknown patterns by making the classification error count risk sensitive to the region around given training samples, i.e., entailing an indirect but effective increase of training samples and training robustness [3].

2.2. Insufficiency of Misclassification Measure

Equation (5) suggests that the adjustment of \(\Lambda\) by MCE training moves the values of the misclassification measure for the misclassified samples from the positive domain to zero and also to the negative domain; for correctly classified samples, the values are changed for much larger absolute values in the negative domain. Note here that the absolute value of a negative misclassification measure implies the certainty factor of decision correctness. The intuition is the larger this amount of negative misclassification error measure, the more certain is the correct decision. Therefore, we can naturally assume that the misclassification measure behaves in the more certain it is the correct decision. This property of the misclassification measure is common to functional margin [6], which has long and widely been used on pattern classification from the earliest research periods. In fact, the functional margin coincides exactly with the sign reversed misclassification measure for two-class cases.

At first glance, the above analysis suggests that MCE training not only pursues classification error minimization but also increases robustness by increasing the margin. However, the misclassification measure as well as the functional margin is insufficient to represent the certainty of classification judgment, which is clear from the fact that the scale transform of the discriminant function values does not always change the class boundary, although it does change the misclassification measure value. Therefore, we have to reformulate the misclassification measure so that it can more directly represent the strength of the robustness.

3. LARGE GEOMETRIC MARGIN MCE

3.1. Geometric Margin

A reasonable measure directly reflecting the certainty of classification judgment can be found by considering the distance between a class boundary and a training sample. For classification robustness, one key measure is the Euclidean distance between the class boundary and its closest, correctly classified sample, which is called the geometric margin [6]. If both training and unknown samples are properly extracted from the true population, setting this geometric margin to a large value seems to guarantee the accurate classification of unknown samples that will appear near the training samples. In other words, the geometric margin directly represents the strength of the classification robustness.

In fact, SVM increases robustness by geometric margin maximization. However, the geometric margin derived in SVM is limited to two-class linear discriminant functions. Furthermore, for a realistic case that is not linearly separable, SVM training attempts to minimize the hinge loss (depicted in Fig. 1 as a broken line) as well as maximize the geometric margin [1]; this loss has little consistency with the 0-1 loss (depicted as a solid line) that must be originally minimized. Recall that MCE is applicable to various types of discriminant functions for the multi-class case and also that, unlike hinge loss, the loss minimized by MCE training (depicted as a dotted line) shows strong consistency with the 0-1 loss. Motivated by this concern, we develop a new MCE training method that increases classification robustness by geometric margin maximization and essentially leads to the ultimate goal of a minimum classification error situation for all samples including unknown future samples.

3.2. Geometric Margin for a General Class of Discriminant Functions

First, we derive the geometric margin for a general class of discriminant functions. New geometric margin \(r\) is defined as the Euclidean distance between the class boundary and its closest, correctly classified training sample \(x^\circ\) (see Fig. 2). Assume that \(x^\circ\) belongs to \(C_y\) and that \(\{g_j(x, \Lambda)\}_{j=1}^J\) are differentiable in both \(x\) and \(\Lambda\) for

\[\text{Eq. (3)} \text{ becomes Eq. (4)} \text{ as } \psi \to \infty.\]
simplicity. Importantly, a set of points, where the value of the misclassification measure becomes 0:

\[ B_y(\lambda) = \{x | d_y(x, \lambda) = 0\}, \]  

(6)
is a boundary that represents whether patterns are classified as \( C_y \) or not. \( r \) is then given as the distance between \( B_y(\lambda) \) and \( x^* \) and can be obtained by solving the following constrained minimization problem:

\[
\begin{align*}
\text{minimize } & x \|x - x^*\|^2 \quad \text{subject to } d_y(x, \lambda) = 0,
\end{align*}
\]

(7)
where \( \| \cdot \| \) denotes the \( L^2 \) (Euclidean) norm. \( r \) equals \( \| x^* - x^0 \| \) where \( x^* \) is a solution to the above minimization problem. To solve this problem, we introduce Lagrange’s multiplier \( \lambda \) and define the following cost function:

\[
J(x, \lambda) = \|x - x^*\|^2 + \lambda d_y(x, \lambda).
\]

(8)

Based on Lagrange’s multipler method, \( x^* \) must satisfy the following equations:

\[
\begin{align*}
2(x^* - x^0) + \lambda \nabla_x d_y(x^*, \lambda) &= 0, \\
d_y(x^*, \lambda) &= 0,
\end{align*}
\]

(9)

(10)
where \( \nabla_x \) denotes the gradient operator with respect to \( x \) and we assume that \( \nabla_x d_y(x^*, \lambda) \neq 0 \). From Eq. (9), \( r \) becomes

\[
r = \frac{|\lambda|}{2} \|\nabla_x d_y(x^*, \lambda)\|.
\]

(11)

Here we expand \( d_y(x, \lambda) \) at point \( x^* \) as follows:

\[
\begin{align*}
d_y(x, \lambda) &= d_y(x^*, \lambda) + \nabla_x d_y(x^*, \lambda)^\top (x - x^*) \\
&\quad + o(|x - x^*|),
\end{align*}
\]

(12)
where \( o(\cdot) \) denotes Landau’s symbol. \( d_y(x^*, \lambda) \) equals zero from Eq. (10) and by substituting \( x = x^* \) in Eq. (12), we get

\[
\nabla_x d_y(x^*, \lambda)^\top (x^0 - x^*) = d_y(x^0, \lambda) + o(r).
\]

(13)
Furthermore from Eq. (9), we get

\[
\nabla_x d_y(x^*, \lambda)^\top (x^0 - x^*) = \frac{\lambda}{2} \|\nabla_x d_y(x^*, \lambda)\|^2.
\]

(14)

By solving the last two equations for \( \lambda \) and substituting the result into Eq. (11), we reach the following equation:

\[
r = \frac{|d_y(x^0, \lambda) + o(r)|}{\|\nabla_x d_y(x^0, \lambda)\|}.
\]

(15)
Here \( |d_y(x^0, \lambda)| \) equals the functional margin for \( x^0 \). As a result, when \( x^0 \) is sufficiently close to the class boundary, the geometric margin is equivalent to the functional margin for \( x^0 \) that is normalized by the norm of the gradient of the functional margin (or the misclassification measure) at \( x^* \), which is the closest point to \( x^0 \) among all points on the class boundary.

For general forms of discriminant functions, we assume that \( x^0 \) is sufficiently close to the class boundary so that we can ignore term \( o(r) \) and replace \( x^* \) by \( x^0 \). Such approximation leads to the following tractable form of the geometric margin:

\[
r = \frac{-d_y(x^0, \lambda)}{\|\nabla_x d_y(x^0, \lambda)\|}.
\]

(16)
For a classifier using linear discriminant functions or a distance-based classifier using multiple prototypes, due to the linearity, component \( o(\cdot) \) vanishes, and \( \nabla_x d_y(x, \lambda) \) becomes independent of \( x \). In this case, Eq. (16) exactly holds even if \( x^0 \) is not close to the boundary.

\section{3.3. MCE Training with Large Geometric Margin}

Equation (16) suggests that we can increase the geometric margin by increasing the functional margin and/or decreasing the norm of the gradient of the misclassification measure in the vicinity of the class boundary. Since the gradient of the misclassification measure in denominator represents a variation of the classification decision due to the input pattern variation, decreasing its norm leads to the suppression of such variation of the classification decision; high robustness is also expected from this point of view.

Now we propose a new MCE training method that directly increases the geometric margin. Note that the following measure,

\[
D_y(x, \lambda) = \frac{d_y(x, \lambda)}{\|\nabla_x d_y(x, \lambda)\|},
\]

(17)
not only corresponds to the sign-reversed geometric margin but can also behave as a misclassification measure. Based on this, we propose a new MCE training method that adopts \( D_y(x, \lambda) \) as a novel misclassification measure. It is easy to understand that this new MCE training changes \( D_y(x, \lambda) \) around the class boundary (where the slope of the loss function exceeds a certain level) in the negative direction, i.e., increases the geometric margin as well as decreases the misclassification counts.

\section{4. EXPERIMENTAL RESULTS}

Although the proposed method was originally applicable to various types of discriminant functions, we conducted experiments using prototype-based distance classifiers as exemplar implementations. Due to the close relation between distance and likelihood (probability), the experimental results can straightforwardly applied to such current major classifier structures as Hidden Markov and Gaussian Mixture models. The discriminant function for class \( C_j \) is given as \( g_j(x, \lambda) = -\|x - p_j\|^2 \), where \( p_j \) is the closest prototype of \( C_j \) to \( x \). In this case, the new misclassification measure of Eq. (17) is expressed as

\[
D_y(x, \lambda) = \frac{d_y(x, \lambda)}{2\|p_y - p_i\|},
\]

(18)
where \( C_p \) and \( C_i \) denote the true and the best-incorrect classes for \( x \), respectively, and \( d_y(x, \lambda) \) is a conventional misclassification measure given as \( d_y(x, \lambda) = \|x - p_y\|^2 - \|x - p_i\|^2 \). For all experiments, the number of prototypes was identical in all classes.

We used the Glass Identification data set presented by UCI Machine Learning Repository\(^2\). This set consists of 214 glass sample patterns of six classes. The content of nine kinds of oxides included in each glass sample is used as a nine-dimensional feature vector.

\(^2\)http://archive.ics.uci.edu/ml/
We computed the recognition rate for open data using the Leave-One-Out (LOO) method. The recognition rate for the closed data was also computed by the 213 patterns used for training; the average of the rates computed by 214 kinds of training sets in the LOO method was adopted.

In Figs. 3 and 4 we summarize the experiment results comparing the proposed method and the conventional MCE method that respectively adopted $D_{\lambda}(x, \Lambda)$ and $d_{\lambda}(x, \Lambda)$ as the misclassification measure. In Fig. 3 where the number of prototypes was changed while $a$ in Eq. (5) was fixed to 10, the recognition accuracy of the proposed method was greater than that of the conventional MCE for the open data (except for the case of one prototype) while the opposite result was obtained for the closed data. This clearly demonstrates that increasing the geometric margin leads to high robustness to unknown patterns. In Fig. 4 where $a$ was changed while the number of prototypes per class was fixed to eight, the proposed method’s accuracy was greater than that of the conventional MCE for the open data when $a \geq 5$. We also calculated the correlation coefficient between the open- and closed-data results for each method and each experiment and summarized them in Table 1. In both experiments, the proposed method yielded a much larger value than the conventional MCE. Since obtaining evaluation results for unknown samples is difficult in an actual application phase, the closed data must be an excellent approximation of the open data. Also in this sense, based on a large geometric margin, the practical usefulness of the proposed method is clear.

In addition, using the proposed method, we successfully reduced the best error rate of the conventional MCE, 4.1%, to 3.4% over the Letter Recognition data set of the same Repository, which consisted of 16000 training samples and 4000 testing samples. The experiment was done using the single Hold-Out method.

![Fig. 3. Recognition rate as function of number of prototypes](image)

![Fig. 4. Recognition rate as function of $a$ in Eq. (5)](image)

<table>
<thead>
<tr>
<th>Method</th>
<th>Fig. 3</th>
<th>Fig. 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conventional MCE</td>
<td>0.80</td>
<td>0.48</td>
</tr>
<tr>
<td>Proposed Method</td>
<td>0.95</td>
<td>0.86</td>
</tr>
</tbody>
</table>

5. CONCLUSION

In this paper we explicitly pointed out that for solving the robustness issue, the geometric margin must be increased instead of the functional margin. We then derived the geometric margin for a general class of discriminant functions and proposed a new MCE-based method that can directly increase the geometric margin. Using prototype-based classifiers, the experimental results demonstrated that the proposed method yielded more robust classification than conventional MCE training.

6. ACKNOWLEDGEMENTS

The authors would like to express gratitude for the support of Dr. Satoshi Nakamura, the project leader of the MASTAR Project.

7. REFERENCES