Large Margin - Minimum Classification Error Using Sum of Shifted Sigmoids as the Loss Function

Madhavi V. Ratnagiri, Biing-Hwang (Fred) Juang, Lawrence Rabiner*

Department of Electrical and Computer Engineering, Rutgers University, Piscataway, New Jersey
*School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, Georgia

10.21437/Interspeech.2011-193

Abstract

We have developed a novel loss function that embeds large-margin classification into Minimum Classification Error (MCE) training. Unlike previous efforts this approach employs a loss function that is bounded, does not require incremental adjustment of the margin or prior MCE training. It extends the Bayes risk formulation of MCE using Parzen Window estimation to incorporate large–margin classification and develops a loss function that is a sum of shifted sigmoids. Experimental results show improvement in recognition performance when evaluated on the TIDigits database.

Index Terms— *large margin, minimum classification error, bayes risk, loss function, pattern classification, Parzen window estimation.*

1. Introduction

The classical Bayes Decision theory, using a 0-1 loss function, defines the optimal classifier that minimizes the risk of mis-classification (called "Bayes Risk"). Such a classifier cannot be practically implemented because it requires knowledge of the true aposteriori probability distribution of the data. Hence other classification techniques like MCE [1] have been developed. MCE uses a sigmoid loss function and minimizes the empirical error rate and guarantees convergence to the global minima. MCE, like all classification algorithms, requires large sets of training data and guarantees improvement in recognition accuracy on the training sets. However to fairly evaluate the classifier its performance is tested on an independent test set, and for high recognition performance on the test set, the training set is made akin to the test set. However differences exist between training and test sets, and the generalization ability of a system to perform just as well on a test set has been shown to improve with discriminative training [8], where the 'distance' between competing classes is maximized.

When using MCE training it has been found that increasing the margin between the correctly classified tokens and the decision boundary leads to further improvement in recognition accuracy [3], [4], [5], [9]. X. Li et. al. [3], in their work on large-margin (LM) classification, showed that constrained optimization of the margin between competing models for correctly classified tokens results in better recognition accuracy compared to MCE based discriminative training. Since they were maximizing the margin for correctly classified tokens only, they initialized the Hidden Markov Models (HMMs) using the standard MCE training to obtain very low training set error rates and thus could discard the relatively low number of wrongly classified tokens. Yu et. al. [4], using the Bayes risk formulation incorporated the margin (defined as the distance between correctly classified tokens and the decision boundary) into the loss function, thus eliminating the need to initialize with MCE training. However in this approach the margin had to incremented manually. Another study by J. Li et. al. [5] incorporated the margin into the loss function based on the hinge loss most commonly used in Support Vector Machines (SVMs). The function they developed was unbounded, hence susceptible to outliers (that may be caused by mislabeling, mis-pronunciation etc.).

In this paper, we develop a LM-MCE approach using a bounded loss function that minimizes the empirical risk while also maximizing the margin of the correctly classified samples. By incorporating the margin into the loss function prior MCE training would not be required as in the case of X. Li et. al's work [3]. This new loss function is based on the Bayes' risk formulation and we have eliminated the need for incremental adjustment of the margin as required in Yu et. al's work [4]. Once the parameters have been initialized, no further adjustment of the margin is required and optimization can be done similar to the MCE approach using gradient probability descent (GPD).

The development of the new sum of the shifted sigmoids loss function using the Parzen Window approximation of the Bayes risk is explained in Section 2. Section 3 presents the experimental results. Section 4 concludes with a summary of the contributions of this work.

2. Sum of Shifted Sigmoid Loss

For the case of an M-class classification task, the goal is to accurately map every feature vector $x \in X$, where X is the feature space to a class label $y \in \{C_1, C_2, ...C_M\}$. The classical Bayes theory defines a cost (1) which when minimized gives the best theoretical recognition accuracy.

$$R = \sum_{j=1}^{M} \int_{X} e_{ji} P(C_j/x) p(x) dx \qquad (1)$$

Copyright © 2011 ISCA

where $p(x)$ is the probability density of the feature vectors, $P(C_j/x)$ is the aposteriori probability and e_{ji} is the cost of wrongly classifying an observation x from class C_j into class C_i. The loss function for the classical Bayes risk is defined as

$$e_{ji} = \begin{cases} 1, & i \neq j \\ 0, & i = j \end{cases} \tag{2}$$

This loss function is discontinuous and cannot be minimized using an optimization routine. Hence a loss function as shown in (3) is used for MCE estimation.

$$e_{ji} = 1\left(d_j(x,\Lambda) \geq 0\right) \tag{3}$$

where 1(.) is the indicator function and $d_j(x,\Lambda)$ is a misclassification measure that exploits the discriminant information. 'd' is a measure of class separability that has a positive value for incorrect classification and is defined as

$$d_j(x,\Lambda) = \left(-g_j(x,\Lambda) + \max_{i \neq j} g_i(x,\Lambda)\right) \tag{4}$$

If the discriminant function $g_i(x,\Lambda)$ of class C_i has a higher value than $g_j(x,\Lambda)$, the token x from class C_j is classified to class C_i, and $d_j(x,\Lambda)$ is positive for an incorrect classification. $\log\left(P\left(C_i/x,\Lambda\right)\right)$ is the most commonly used discriminant function, where Λ is the set of model parameters and $P\left(C_i/x,\Lambda\right)$ defines the model density. Rewriting the risk (1) in terms of the misclassification measure 'd' we have:

$$\begin{aligned} R &= \sum_{j=1}^{M} \int_X 1\left(d_j(x,\Lambda) \geq 0\right) p\left(C_j, x\right) dx \\ &= \sum_{j=1}^{M} P(C_j) \int_{X_j} p\left(x/C_j\right) dx \\ &= \sum_{j=1}^{M} P(C_j) \int_0^{\infty} p(m_j/C_j) dm_j \end{aligned} \tag{5}$$

where $X_j = \left\{x \in X \mid d_j(x,\Lambda) \geq 0\right\}$ is the set of feature vectors that are likely to be mis-classified, $P(C_j)$ is the apriori probability and $m_j = d_j(x,\Lambda)$. The third equality in (5) is obtained by converting the probability density from the feature domain x to the misclassification measure domain m_j. For details the reader is referred to the papers by McDermott et. al. [6] or Yu et. al. [4]. The conversion to the measure domain facilitates the Parzen window estimation of the density $p(m_j/C_j)$ as shown below.

$$p_{N_j}\left(m_j/C_j\right) = \frac{1}{N_j} \sum_{r=1}^{N_j} \frac{1}{h} K\left(\frac{m_j - d_j(x_r,\Lambda)}{h}\right) \tag{6}$$

where $K\left(\frac{m_j - d_j(x_r,\Lambda)}{h}\right)$ is the Parzen window or the kernel function of bandwidth h, centered at every data point $d_j(x_r,\Lambda)$ and x_r is among the data samples hypothesized as belonging to class C_j. The resulting risk, expressed in terms of the Parzen window estimate, is :

$$\begin{aligned} R_N(L) &= \frac{1}{N} \sum_{j=1}^{M} \int_{-\rho}^{\infty} \sum_{r=1}^{N_j} \frac{1}{h} K\left(\frac{m_j - d_j(x_r,\Lambda)}{h}\right) dm_j \\ &= \frac{1}{N} \sum_{j=1}^{M} L\left(d_j(x_r,\Lambda)\right) \end{aligned} \tag{7}$$

where L(.) is the loss function. This empirical risk will approach Bayes risk as the Parzen window estimates converge to the true density, which can be obtained by increasing the number of training samples to infinity. The convergence to Bayes' risk also depends on making the right assumption about the aposteriori probability of the class and on the optimization algorithm converging to the global minimum [2], [6].

Many functions can be used as kernels for the Parzen window, provided they satisfy the constraints defined in (8).

$$\begin{aligned} &K(u) \geq 0 \\ &\text{and} \\ &\int K(u)du = 1 \qquad \forall u \end{aligned} \tag{8}$$

When a kernel function defined in (9) is used with the MCE cost function (3) the standard sigmoid loss function (10) is obtained.

$$K(u) = \frac{e^{u/h}}{h * \left(1 + e^{u/h}\right)^2} \tag{9}$$

$$L_{\text{standard MCE}}(d) = \frac{1}{\left(1 + e^{-\gamma(d)}\right)} \tag{10}$$

where $u = \left(\frac{m_j - d_j(x_r,\Lambda)}{h}\right)$ and $\gamma = 1/h$.

Other loss functions that satisfy the kernel conditions can be developed, but not all of them will necessarily lead to better classification accuracy. We found from our previous study

that functions which extend the range of the loss function on the side of the correctly classified perform better in terms of recognition accuracy [9].

To get an intuitive understanding of error minimization using the sigmoid loss function and the new sum of shifted sigmoid loss function, recall that tokens which have d>0 are considered wrongly classified, while d<0 are the tokens that have been correctly classified and d=0 is the boundary. When optimization is done using GPD, the parameter updates are controlled by the slope of the loss function and from the slope of sigmoid loss function in Figure 1 it can be observed that maximum weight is assigned to training tokens that are at the boundary i.e. d=0 and that tokens that are further away from the boundary (either correctly or incorrectly classified) are not considered in the training process. In this paper we introduce a novel and very practical approach to incorporating large margin classification and extending the range of the loss function, by using a sum of multiple sigmoid functions (11) that have been shifted to include the correctly classified tokens. As observed in Figure 1 the kernel of the new loss function represented by the thick line gives maximum weight not only to tokens that are close to the decision boundary, but also to tokens that have been correctly classified but are removed from the decision boundary.

$$K = \frac{1}{N}\sum_{i=0}^{N} K_i$$

$$\text{where } N = 19,\ K_i = \frac{\gamma e^{\gamma(u+\rho_i)}}{\left(1+e^{\gamma(u+\rho_i)}\right)^2} \qquad (11)$$

$$\text{and } \rho_i = \{-3,-2,-1,0,1,2,\ldots 15\}$$

where ρ_i is the margin shift.

The curve has been shifted more to the left to consider only the correctly classified samples and not the wrongly classified outlier samples to the far right of the decision boundary. The kernel (11) satisfies the kernel conditions (8) for Parzen Window Estimation and hence the sum of shifted sigmoids loss function could approach the Bayes risk, provided the assumptions are satisfied. The second condition of the kernel constraints (8) requires the area under the kernel be unity, hence widening the width of the kernel so as to weigh more tokens, decreases the absolute value of the kernel. We found that it is the relative weight associated with the tokens across the range of d that affects the recognition accuracy and not the absolute value of the kernel function. The loss function for such a kernel is given in (12) and a plot of the loss function is shown in Figure 2. We see from (12) that the loss function is bounded and hence not susceptible to outliers and can be minimized efficiently using optimization techniques previously developed for MCE, like Gradient Probability Descent. The range of ρ_i can be chosen based on the range of 'd' for the given training data.

$$L = \frac{1}{N}\sum_{i=0}^{N} L_i$$

$$\text{where } N = 15,\ L_i = \frac{1}{\left(1+e^{-\gamma(d+\rho_i)}\right)} \qquad (12)$$

$$\text{and } \rho_i = \{-3,-2,-1,0,1,2,\ldots 15\}$$

3. Results

The new loss function for pattern classification was evaluated on a speech recognition task using the TIDigits database. This database consists of 8623 utterances for training and 8700 test utterances spoken by men and women. 11 whole word Hidden Markov Models (HMMs) for the 9 digits ("one", "two",…."nine") along with "oh" and "zero" were trained, with 32 mixtures/state. These models were trained using 12 Mel frequency Cepstral Coefficients (MFCCs) plus energy features including their first and second order time derivatives. The models were initialized using ML estimation, before applying the MCE or LM-MCE estimation. A comparison of the string accuracies of the recognition system evaluated on the TIDIGIT test set using the ML estimation and LM-MCE with sum of shifted sigmoids loss is presented in Table 1. As can be seen the LM-MCE with the newly introduced loss function gives the best performance. MCE with standard loss is the conventional MCE approach [1] that uses the loss function shown in (10). The sum of shifted sigmoids loss as defined by (12) was minimized just like the standard MCE using 20 iterations of GPD optimization with $\gamma = 1.3$. The results presented in Table 1 used a margin range of (-3 to 7). When different ranges were considered (Table 2) it was found that decreasing the range to (-3 to 4) as well as increasing the range to (-3 to 15) reduced the accuracy. This suggests that including more of the correctly classified tokens significantly improves the recognition accuracy but not all the correctly classified tokens need be included. The string error rate is a measure of the relative improvement and comparing this measure with the LM-MCE approach developed by Yu et. al. [4] we found that the string error rate (SER) obtained using this approach (24.5%) exceeds the best SER obtained by Yu et. al. [4]. It also exceeds the SER we obtained in our previous study using other sigmoid loss functions [9].

ML Estimation	MCE w/standard loss	LM-MCE w/ sum of shifted sigmoids loss
97.77%	98.55%	98.92%

Table 1. *Comparison of string accuracy using ML, MCE w/ standard los and LM-MCE w/ new sum of shifted sigmoids loss function.*

MCE w/ standard loss	LM-MCE w/ new loss w/ margin range (-3 to 4)	LM-MCE w/ new loss w/ margin range (-3 to 7)	LM-MCE w/ new loss w/ margin range (-3 to 15)
98.57%	98.79%	98.92%	98.85%

Table 2. *Comparison of string accuracy using different ranges of margin for the new sum of shifted sigmoids loss function*

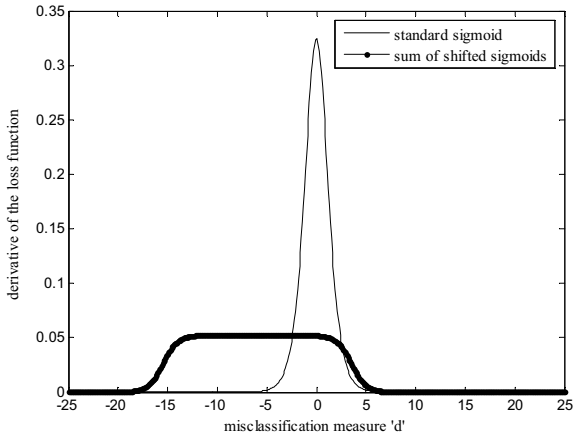


Figure 1. *Plot comparing the derivatives of the standard sigmoid, and the new, sum of shifted sigmoids, loss functions.*

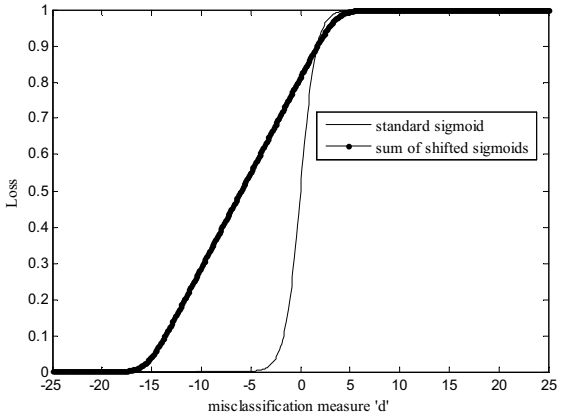


Figure 2. *Plot comparing the standard sigmoid and the new sum of shifted sigmoids loss function.*

4. Conclusions

We have introduced a new approach to Large Margin Classification that uses a bounded loss function which is a sum of shifted sigmoids. It maximizes the margin of the tokens that are correctly classified, while also minimizing the empirical error rate. The margin is optimally maximized without the need for manual increments. This new loss function was developed from a formulation of the Bayes risk using Parzen window estimation and has been shown to give improved performance on the TIDigits database. It uses GPD to do the minimization just like the conventional MCE and does not require any more iterations than the standard MCE approach and can be easily extended to large vocabulary databases.

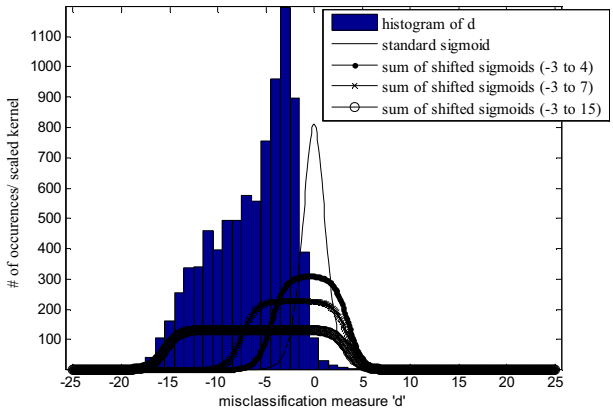


Figure 3. *Plot of kernel of the sum of shifted sigmoids functions for different margin ranges in comparison to the histogram of 'd'.*

5. References

[1]. B-H. Juang, W. Chou, C-H Lee, "Minimum Classification Eror Rate Methods for Speech Recognition", IEEE Transactions on Speech and Audio, Vol. 5, No. 3, May 1997.

[2] J. C. Burges,"A Tutorial on Support Vector Machines for Pattern Recognition", Data Mining and Knowledge Discovery, 2(2), 121-167.

[3] X. Li, H.Jiang, C. Liu,"Large Margin HMMs for Speech Recognition", IEEE 2005

[4] D. Yu and L. Deng, "Large-Margin Discriminative Training of Hidden Markov Models for Speech recognition", International Conference on Semantic Computing, IEEE 2007.

[5] J. Li, M. Yuan and C-H Lee, "Approximate Test Risk Bound Minimization Through Soft Margin Estimation", IEEE transactions on Audio, Speech and Language Processing, Vol.15, No. 8, November 2007.

[6] E. McDermott and S. Katagiri, "Minimum Classification Error Via a Parzen Window Based Estimate of the Theoretical Bayes Classification Risk" Proc. ICSLP, 2002.

[7] Duda and P. Hart, "Pattern Classification and Scene Analysis", John Wiley and Sons.

[8] E. McDermott, T.J Hazen, J.L.Roux, A. Nakamura and S. Katagiri, "Discriminative Training for Large-Vocabulary Speech Recognition Using Minimum Classification Error", ICASSP 2007.

[9] M.V. Ratnagiri, L. Rabiner and B-H. Juang, "Multi-Class Classification Using a New Sigmoid Loss Function for Minimum Classification Error (MCE)", ICMLA 2010.