A NEW FORMALIZATION OF MINIMUM CLASSIFICATION ERROR USING A PARZEN ESTIMATE OF CLASSIFICATION CHANCE

Erik McDermott & Shigeru Katagiri

NTT Communication Science Laboratories, NTT Corporation
Hikari-dai 2-4, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
mcd@cslab.kecl.ntt.co.jp

ABSTRACT

In recent work, we showed that the Minimum Classification Error (MCE) criterion function commonly used for discriminative design of pattern recognition systems is equivalent to a Parzen window based estimate of the theoretical classification risk. In this analysis, each training token is mapped to the center of a Parzen kernel in the domain of a suitably defined random variable; the kernels are then summed and integrated over the domain of incorrect classifications, yielding the risk estimate. Here, we deepen this approach by applying Parzen estimation at an earlier stage of the overall definition of classification risk. Specifically, the new analysis uses all incorrect categories, not just the single best incorrect category, in deriving a "correctness" function that is a simple multiple integral of a Parzen kernel over the region of correct classifications. The width of the Parzen kernel determines how many competing categories to use in optimizing the resulting overall risk estimate. This analysis uses the classic Parzen estimation method to support the notion that using multiple competing categories in discriminative training is a type of smoothing that enhances generalization to unseen data.

1. INTRODUCTION

The Minimum Classification Error (MCE) framework is an approach to discriminative training for pattern classification that explicitly incorporates classification performance into the training criterion. Given discriminant functions for each category, MCE defines a loss function that is a smoothed approximation of the recognition error rate, and then uses this function as the criterion function for optimization [3, 4, 5, 6]. Through minimization of this criterion function, MCE is aimed directly at minimizing classification error rather than at learning the true data probability distributions, the target of Maximum Likelihood Estimation (MLE). MCE has been used succesfully in various pattern recognition tasks [1, 7].

In recent work, a new theoretical perspective on MCE was presented [8, 9]. This addressed the nature of the smoothness of the MCE loss function, as well as the relationship between minimization of an overall MCE loss summed over a finite set of training data and minimization of the theoretical classification risk measured over the continuous densities underlying the classification problem. It was shown that the continuous, 0-1 MCE loss function can be derived from an estimate of the theoretical classification risk, using Parzen estimation of the density of a suitably defined misclassification measure. In this analysis, the specific kernel type used for Parzen estimation leads to a specific type of MCE loss function, and vice versa; the width of the Parzen kernel directly corresponds to the steepness of the MCE loss function, and vice versa. Minimization of the MCE loss function corresponds to the minimization of a Parzen estimate of the theoretical classification risk. This derivation used a Parzen estimate of the density of a random variable corresponding to a comparison of the correct category's score against the best incorrect category's score.

Here, we deepen this approach by applying Parzen estimation to the density of a vector of jointly distributed random variables, each corresponding to a pair-wise comparison between the correct category and all incorrect categories. By using Parzen estimation at an earlier level than that used in the previous analysis, a finer grained model of classification risk is obtained. In the new analysis, larger kernel widths, used in Parzen estimation to obtain smoother estimates and better generalization to unseen data, directly result in more categories being used in gradient-based optimization of the risk estimate. This analysis shows that the use of multiple competing categories, intuitively appealing for the sake of generalization to unseen data, and commonly used in existing MCE and Maximum Mutual Information (MMI) speech recognition studies, can be rigorously linked to the theoretical classification risk.

2. THE MINIMUM CLASSIFICATION ERROR FRAMEWORK

The MCE framework has been described in several publications [3, 5, 6]. For each training token, MCE maps a training pattern token $\mathbf{x}$ and the system parameters $\mathbf{\Lambda}$ (e.g., all the

hidden Markov model means and covariances) to a 0-1 loss function reflecting classification error. The pattern $\mathbf{x}$ could be a single pattern vector or a sequence of, e.g., speech-derived feature vectors, $\mathbf{x} = \mathbf{x}_1^T = (\mathbf{x}^1, ..., \mathbf{x}^t, ..., \mathbf{x}^T)$. The formalism assumes that discriminant functions $g_j(\mathbf{x}, \mathbf{\Lambda})$ can be defined for each string category C_j, and uses a misclassification measure $d_k(\mathbf{x}, \mathbf{\Lambda})$ to compare the match between the training token to the correct category C_k with the match to the best incorrect categories. The loss function is typically a sigmoid,

$$\ell(d_k(\mathbf{x}, \mathbf{\Lambda})) = \frac{1}{1 + e^{-\alpha d_k(\mathbf{x}, \mathbf{\Lambda})}}. \quad (1)$$

The total loss $L(\mathbf{X}, \mathbf{\Lambda})$ is the local loss summed over the M categories and N_k tokens in each category C_k making up the training data $\mathbf{X}$:

$$L(\mathbf{X}, \mathbf{\Lambda}) = \frac{1}{N} \sum_{k}^{M} \sum_{j=1}^{N_k} \ell(d_k(\mathbf{x}_{k,j}, \mathbf{\Lambda})), \quad (2)$$

where $\mathbf{x}_{k,j}$ denotes the j-th training token in category C_k, and $N = \sum_{k=1}^{M} N_k$. The overall loss function can be minimized using several different approaches [6].

3. A NOVEL ANALYSIS OF THE SMOOTHNESS OF THE MCE LOSS FUNCTION

3.1. Theoretical classification risk

The starting point for our new formalization of Minimum Classification Error is the theoretical classification risk, integrated over the entire pattern space $\mathcal{X}$ [2]:

$$\mathcal{R} = \sum_{k=1}^{M} \int_{\mathcal{X}} \lambda(\alpha(\mathbf{x}, \mathbf{\Lambda})|C_k) p(C_k, \mathbf{x}) d\mathbf{x}. \quad (3)$$

This represents the probability of error, optimal or not, for a given pattern classification system in the theoretical situation where the joint densities $p(C_k, \mathbf{x})$ are known. Here $\alpha(\mathbf{x}, \mathbf{\Lambda})$ represents the classification decision (the choice of one out of M categories) made for the input pattern $\mathbf{x}$, given the system parameters $\mathbf{\Lambda}$, while $\lambda(\alpha_i|C_k)$ denotes the cost of mis-classifying a member of category C_k as category C_i. Typically this cost is 1 whenever i and k are different. In the following it is assumed that the system decisions $\alpha(\mathbf{x}, \mathbf{\Lambda})$ are taken by choosing the category with the largest discriminant function value $g_j(\mathbf{x}, \mathbf{\Lambda})$.

3.2. Defining classification risk in a new domain

The new approach defines a *pair-wise misclassification measure* for each of the $M - 1$ incorrect categories C_i,

$$m_{k,i} = d_{k,i}(\mathbf{x}, \mathbf{\Lambda}) = -g_k(\mathbf{x}, \mathbf{\Lambda}) + g_i(\mathbf{x}, \mathbf{\Lambda}), \quad (4)$$

where i ranges from 1 to M but skips the correct category C_k. The overall risk can then be written as

$$\mathcal{R} = 1 - \sum_{k=1}^{M} \int_{\mathcal{X}} 1(\forall i : d_{k,i}(\mathbf{x}, \mathbf{\Lambda}) < 0) p(C_k, \mathbf{x}) d\mathbf{x}, \quad (5)$$

where the indicator function $1(\forall i : d_{k,i}(\mathbf{x}, \mathbf{\Lambda}) < 0)$ is 1 when all the misclassification measures are negative, and 0 otherwise. In turn, this is equivalent to integrating over the correctly classified part of the pattern space $\mathcal{X}$:

$$\mathcal{R} = 1 - \sum_{k=1}^{M} \int_{\mathcal{X}_k} p(C_k, \mathbf{x}) d\mathbf{x}, \quad (6)$$

where $\mathcal{X}_k = \{\mathbf{x} \in \mathcal{X} \mid \forall i : d_{k,i}(\mathbf{x}, \mathbf{\Lambda}) < 0\}$. For convenience, in the following we use the classification *chance* $\mathcal{C} = 1 - \mathcal{R}$ rather than the risk $\mathcal{R}$, but use the two terms interchangeably, with the understanding that the one is just 1 minus the other.

If X and all the $g_i(X, \mathbf{\Lambda})$ are continuous random variables, then all the $M_{k,i} = d_{k,i}(X, \mathbf{\Lambda})$ are also continuous random variables. We can then express the integral for each category C_k over the reduced space $\mathcal{X}_k$ with a multiple integral over the negative domains of the misclassification measures:

$$\begin{aligned} &\int_{\mathcal{X}_k} p(C_k, \mathbf{x}) d\mathbf{x} \quad (7) \\ &\quad = P[M_{k,1} < 0, ..., M_{k,M} < 0, C_k] \\ &\quad = \int_{-\infty}^{0} ... \int_{-\infty}^{0} p(C_k, m_{k,1}, ..., m_{k,M}) dm_{k,1} ... dm_{k,M}. \end{aligned}$$

Let $\mathbf{m}_k = (m_{k,1}, ..., m_{k,i \neq k}, ..., m_{k,M})$ denote an arbitrary $M - 1$ dimensional vector of jointly distributed pair-wise misclassification measures, and $\mathcal{M}_k = \{\mathbf{m}_k | \ \forall i : m_{k,i} < 0\}$ be the jointly negative part of the misclassification measure space. We can now express the overall chance corresponding to Equ. (6) as:

$$\mathcal{C} = \sum_{k=1}^{M} P(C_k) \int_{\mathcal{M}_k} p(\mathbf{m}_k|C_k) d\mathbf{m}_k. \quad (8)$$

This is equivalent to (1 minus) the original expression of risk given in Equ. (3). In contrast to the previous approach [8, 9] where the classification risk was rewritten using just one variable, comparing the correct category with the best incorrect category, here the classification risk has been rewritten using the vector of comparisons between the correct category and all incorrect categories.

A new approach to pattern classifier design is suggested by Equ. (8). We can try to estimate the density $p(\mathbf{m}_k|C_k)$ given the available training tokens $\mathbf{x}_{k,j}$ and the system parameters $\mathbf{\Lambda}$, plug that estimate into Equ. (8) to obtain an

estimate of classification chance, and then adjust the system parameters so as to maximize the estimate of chance / minimize the estimate of risk. The following sections outline this new approach.

3.3. Parzen estimate of classification chance

Let $\mathbf{d}_k(\mathbf{x}_{k,j}, \mathbf{\Lambda})$ denote the $M-1$ dimensional vector of pair-wise misclassification measures resulting from presenting training token $\mathbf{x}_{k,j}$ to a recognition system with parameters $\mathbf{\Lambda}$. In order to explicitly relate a finite body of training data to the theoretical cost, we use a Parzen estimate of the density $p(\mathbf{m}_k|C_k)$:

$$p_{N_k}(\mathbf{m}_k|C_k) = \frac{1}{N_k}\sum_{j=1}^{N_k}\frac{1}{h^{M-1}}\phi(\frac{\mathbf{m}_k - \mathbf{d}_k(\mathbf{x}_{k,j}, \mathbf{\Lambda})}{h}). \tag{9}$$

Here $\phi((\mathbf{m}_k - \mathbf{d}_k(\mathbf{x}_{k,j}, \mathbf{\Lambda}))/h)$ is a kernel of uniform width h along each dimension, centered on the transformed data point $\mathbf{d}_k(\mathbf{x}_{k,j}, \mathbf{\Lambda})$. With $p_{N_k}(\mathbf{m}_k|C_k)$ defined for any value of $\mathbf{m}_k$, we can now define an estimate of the theoretical classification chance expressed in Equ. (8):

$$\mathcal{C}_N = \sum_{k=1}^{M} P(C_k)\int_{\mathcal{M}_k} p_{N_k}(\mathbf{m}_k|C_k)d\mathbf{m}_k. \tag{10}$$

Expanding this using the estimate $p_{N_k}(\mathbf{m}_k|C_k)$ given by Equ. (9) yields

$$\mathcal{C}_N = \sum_{k=1}^{M}\frac{P(C_k)}{N_k}\int_{\mathcal{M}_k}\sum_{j=1}^{N_k}\frac{1}{h^{M-1}}\phi(\frac{\mathbf{m}_k - \mathbf{d}_k(\mathbf{x}_{k,j}, \mathbf{\Lambda})}{h})d\mathbf{m}_k. \tag{11}$$

Rearranging, and using N_k/N for $P(C_k)$, gives:

$$\mathcal{C}_N = \frac{1}{N}\sum_{k=1}^{M}\sum_{j=1}^{N_k}\frac{1}{h^{M-1}}\int_{\mathcal{M}_k}\phi(\frac{\mathbf{m}_k - \mathbf{d}_k(\mathbf{x}_{k,j}, \mathbf{\Lambda})}{h})d\mathbf{m}_k. \tag{12}$$

Defining the "correctness" function for a single training token,

$$\mathcal{V}_k(\mathbf{x}_{k,j}, \mathbf{\Lambda}) = \frac{1}{h^{M-1}}\int_{\mathcal{M}_k}\phi(\frac{\mathbf{m}_k - \mathbf{d}_k(\mathbf{x}_{k,j}, \mathbf{\Lambda})}{h})d\mathbf{m}_k, \tag{13}$$

we can express the Parzen estimate of classification chance (Equ. (12)) as

$$\mathcal{C}_N = \frac{1}{N}\sum_{k=1}^{M}\sum_{j=1}^{N_k}\mathcal{V}_k(\mathbf{x}_{k,j}, \mathbf{\Lambda}). \tag{14}$$

This analysis has arrived at an estimate of classification chance, defined in terms of an integral over jointly negative values of the misclassification measures, and using a Parzen estimate of the joint density of the misclassification measures. The interpretation of the integral that $\mathcal{V}_k(\mathbf{x}_{k,j}, \mathbf{\Lambda})$ is based on is that it expresses the hyper-volume of the joint density in the all-negative region of the misclassification measure vector space corresponding to correct classifications, for a single Parzen window, centered on the transformed data point $\mathbf{d}_k(\mathbf{x}_{k,j}, \mathbf{\Lambda})$. Figure 1 illustrates a Parzen kernel in this scenario. This hyper-volume ranges from 0 to h^{M-1}, depending on the center $\mathbf{d}_k(\mathbf{x}_{k,j}, \mathbf{\Lambda})$, and, significantly, on the kernel width h. Normalizing by the maximum volume h^{M-1} yields a 0-1 density; summing over all data tokens and all categories yields the overall estimate of classification chance.

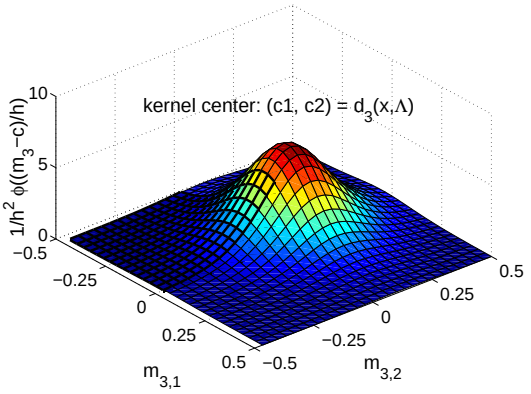

Fig. 1. Example of a Parzen kernel for use in the estimation of $p(\mathbf{m}_3|C_3)$ in a 3-category problem. The kernel center and width determine what fraction of the kernel distribution falls in the (all-negative) part of $\mathbf{m}$-space corresponding to correct classifications.

Parzen estimation theory [2] tells us that this estimate converges to the theoretical chance as the number of data tokens approaches infinity and the kernel width is reduced. Furthermore, we have some control over the value of $\mathcal{C}_N$ via the system parameters $\mathbf{\Lambda}$ that are implicit in the definition of the misclassification measures, and that therefore affect the specific values of the window anchor points $\mathbf{d}_k(\mathbf{x}_{k,j}, \mathbf{\Lambda})$.

3.4. Parzen kernel leads to a multi-category sigmoid "correctness" function

For simple Parzen kernels, we can easily find the closed form of the multiple integral in Equ. (13). In previous work [8, 9], it was shown that the sigmoid loss function commonly used in MCE studies can be seen as the result of using a particular type of Parzen kernel. Here we examine the corresponding multi-variate kernel:

$$\phi(\mathbf{u}) = \prod_{i\neq k}^{M-1}\frac{e^{u_i}}{(1+e^{u_i})^2}. \tag{15}$$

As above, the kernel is centered on a given (transformed) data point $\mathbf{d}_k(\mathbf{x}_{k,j}, \mathbf{\Lambda})$ and its width is controlled by a scalar h. Using this kernel in Equ. (13) gives:

$$\mathcal{V}_k(\mathbf{x}_{k,j}, \mathbf{\Lambda}) = \prod_{i \neq k}^{M-1} \frac{1}{1 + e^{\frac{1}{h} d_{k,i}(\mathbf{x}_{k,j}, \mathbf{\Lambda})}}, \qquad (16)$$

illustrated for a 3-category problem in Figure 2. As the training set grows larger, the kernel width h should be decreased as per Parzen estimation theory, resulting in an increasingly steep $\mathcal{V}_k(\mathbf{x}_{k,j}, \mathbf{\Lambda})$. As the training set size approaches infinity, h approaches zero, and $\mathcal{V}_k(\mathbf{x}_{k,j}, \mathbf{\Lambda})$ approaches a binary 1-0 step function.

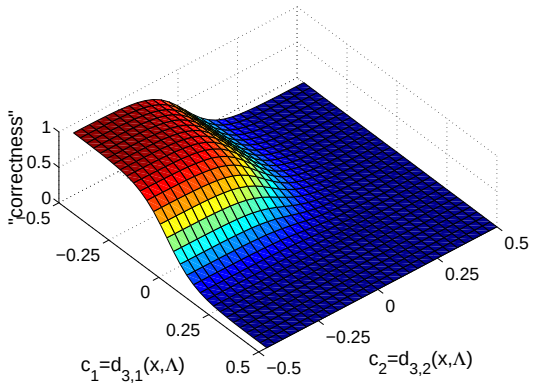

Fig. 2. Estimate of classification correctness as a function of Parzen kernel center, for a single training token (belonging to category 3) in a 3-category problem

3.5. Minimizing the risk estimate

The task of any optimization procedure applied to the risk estimate corresponding to Equ. (14) is to choose the system parameters $\mathbf{\Lambda}$ so as to position the kernel centers in such a way as to minimize the estimated risk. Whether the resulting minimization corresponds to the theoretically optimal Bayes risk rests on the effectiveness of the optimization procedure (and on the model structure). Though the new risk estimate derived here is different from previous MCE losses, it is closely related; the techniques already in use for MCE optimization can be applied with few difficulties.

The new formalism has appealing consequences for gradient based optimization. The gradient of Equ. (16) with respect to $\mathbf{\Lambda}$ will be negligible for incorrect categories whose misclassification measure is extreme, i.e. who have very good or very bad scores. This effect is controlled by the kernel width h: narrower (wider) kernels will result in fewer (more) categories having significant gradients. In particular, narrow kernels will lead the optimization procedure to use mainly the top incorrect categories. Parzen estimation theory in this context effectively suggests that when the training set is small, more incorrect categories should be used in gradient-based optimization of the classification risk estimate, but that as the training set increases, only the top few categories need be used.

4. SUMMARY

Here we have shown that an MCE-like "correctness" function using pair-wise comparisons between the correct category and all incorrect categories can be derived from a smooth Parzen window based estimate of the true theoretical classification risk. The risk estimate is found by simple integration over the domain of misclassifications in a vector space that is a system-dependent transformation of the input pattern space. Gradient-based methods can then be used to find system parameters that minimize the resulting risk estimate. The smoothness of the estimate is controlled by the Parzen kernel width, which also determines the number of incorrect categories deemed "competitive" to the correct category during optimization. This type of margin control, typically effective for generalization to unseen data, is here shown to be the direct consequence of a novel Parzen window based approach to estimating the theoretical probability of classification error.

5. REFERENCES

[1] Biem, A. (2001). *Minimum Classification Error Training of Hidden Markov Models for Handwriting Recognition*. Proceedings of the IEEE ICASSP.

[2] Duda, R.O. and Hart, P.E. (1973). *Pattern Classification and Scene Analysis*. John Wiley & Sons.

[3] Juang, B.-H. and Katagiri, S. (1992). Discriminative Learning for Minimum Error Classification. *IEEE Transactions on Signal Processing*, Vol. 40, No. 12, pp. 3043-3053.

[4] Katagiri, S., Lee, C.-H. and Juang, B.-H. (1991). New Discriminative Training Algorithms Based on the Generalized Descent Method. *Proc. 1991 IEEE Workshop on Neural Networks for Signal Processing*, pp. 299-308.

[5] Katagiri, S., Juang, B.-H. and Lee, C.-H. (1998). Pattern Recognition Using a Family of Design Algorithms based upon the Generalized Probabilistic Descent Method. *Proc. IEEE*, Vol. 86, No. 11, pp. 2345-2373.

[6] McDermott, E. (1997). *Discriminative Training for Speech Recognition*. Doctoral thesis, Waseda University, Tokyo.

[7] McDermott, E., Biem, A., Tenpaku, S., and Katagiri, S. (2000). *Discriminative Training for Large Vocabulary Telephone-based Name Recognition*. Proc. of the IEEE ICASSP.

[8] McDermott, E. and Katagiri, S. (2002). *A Parzen window based derivation of Minimum Classification Error from the theoretical Bayes classification risk.* Proc. of ICSLP.

[9] McDermott, E. and Katagiri, S. (2002). *Minimum Classification Error via a Parzen window based estimate of the theoretical Bayes classification risk* Proc. of the IEEE Workshop on Neural Networks for Signal Processing, pp. 415-424.