

Model-Based MCE Bound to the True Bayes' Error

Ralf Schlüter and Hermann Ney

Abstract—In this letter, we show that the minimum classification error (MCE) criterion gives an upper bound to the true Bayes' error rate independent of the corresponding model distribution. In addition, we show that model-free optimization of the MCE criterion leads to a closed form solution in the asymptotic case of infinite training data. While leading to the Bayes' error rate, the resulting model distribution differs from the true distribution. This suggests that the structure of model distributions trained with the MCE criterion should differ from the structure of the true distributions, as they are usually used in statistical pattern recognition.

Index Terms—Error bounds, maximum mutual information (MMI), minimum classification error (MCE), speech recognition, statistical pattern recognition, training criteria.

I. INTRODUCTION

As starting point for this letter, we will summarize the state of the art for discriminative training. The minimum classification error (MCE) training criterion [5], as well as other discriminative training criteria like the maximum mutual information (MMI) criterion¹ ([2], pp. 785), ([4], pp. 262) are important alternatives to the standard maximum likelihood (ML) criterion for training the model parameters in statistical pattern recognition. For various classification tasks such as speech and image object recognition, improvements in error rate, often combined with a reduction in the number of parameters, have been reported for MCE and MMI training when compared to ML training [3], [6]–[8].

The MCE criterion represents a smoothed version of the empirical error rate on the training data for a given classifier. Because of this direct relation between the training criterion and the final goal to reduce the error rate, it might be expected that an MCE trained classifier should not heavily depend on the quality of certain model assumptions, as is the case for both ML and MMI training.

In this letter, proofs for the following properties for the MCE criterion and the *generalized Gini* criterion will be given.

- The MCE criterion represents an upper bound to the true (optimal) Bayes' error rate, independent of the underlying model distribution.
- Model-free optimization of the MCE criterion with sufficient training data leads to a closed form solution. The

Manuscript received April 6, 2000. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. A. S. Spanias.

The authors are with the Lehrstuhl für Informatik VI, RWTH Aachen University of Technology, Aachen, Germany (e-mail: schlüter@informatik.rwth-aachen.de).

Publisher Item Identifier S 1070-9908(01)02796-1.

¹The MMI criterion, as it is usually used in speech recognition, should correctly be referred to as the *equivocation* criterion. Both criteria are equivalent only in the case of given class priors.

corresponding model distribution leads to the true Bayes' error rate.

- A new discriminative training criterion similar to the MCE criterion, the generalized Gini (GG) criterion is defined. The GG criterion gives an closer upper bound to the Bayes' error rate than the MCE criterion. Model-free optimization of the GG criterion with infinite training data leads to the same optimal model distribution as the MCE criterion.

The model distributions resulting from model-free optimization of the MCE and the GG criterion differ considerably from the true distributions. This result suggests modifications to the structure of the model distributions applied in MCE training.

It should be noted that in this letter, we strictly distinguish between true distributions (representing the training data) and the corresponding model distributions. This strict distinction is usually not found in literature.

II. ERROR BOUNDS

Let x_n be an observation and k_n the corresponding class label with $n = 1, \dots, N$. True distributions are denoted by p , e.g., $p(x, k)$ and $p(k|x)$. The true distributions are distinguished from the corresponding model distributions by the parameter θ , e.g., $p_\theta(x, k)$ and $p_\theta(k|x)$. For both the true distributions and the model distributions, the usual normalization constraints are assumed. Then the expectation of the true Bayes error rate $p_{\text{Bayes}}(e)$ is given by

$$\begin{aligned} p_{\text{Bayes}}(e) &= \lim_{N \rightarrow \infty} \frac{1}{N} \cdot \sum_{n=1}^N \left[1 - \max_k p(k|x_n) \right] \\ &= \int dx \cdot p(x) \cdot \underbrace{\left[1 - \max_k p(k|x) \right]}_{:= p_{\text{Bayes}}(e|x)} \end{aligned}$$

with the local Bayes error $p_{\text{Bayes}}(e|x)$.

A. MMI Criterion

The model based posterior probability for class k given observation x will be denoted by $p_\theta(k|x)$. In the limit of an infinite amount of training data, the (model based) MMI criterion then is defined by

$$\begin{aligned} F_{\text{MMI}} &= \lim_{N \rightarrow \infty} \frac{1}{N} \cdot \sum_{n=1}^N \log p_\theta(k_n|x_n) \\ &= \int dx \cdot \sum_k p(x, k) \cdot \log p_\theta(k|x) \\ &= \int dx \cdot p(x) \cdot \underbrace{\sum_k p(k|x) \cdot \log p_\theta(k|x)}_{:= f_{\text{MMI}}(e|x)} \end{aligned}$$

where $f_{\text{MMI}}(e|x)$ denotes the local MMI score. The notation here reflects the fact that $f_{\text{MMI}}(e|x)$ gives an upper bound to the local true Bayes' error rate. Utilizing the inequality $\log y \leq y - 1$ for the natural logarithm, we obtain the following inequality:

$$\begin{aligned} -f_{\text{MMI}}(e|x) &= -\sum_k p(k|x) \log p_\theta(k|x) \\ &\geq \sum_k p(k|x) \cdot (1 - p_\theta(k|x)) \\ &= 1 - \sum_k p(k|x) \cdot p_\theta(k|x) \\ &\geq 1 - \max_k p(k|x) \\ &= p_{\text{Bayes}}(e|x). \end{aligned} \quad (1)$$

Hence, except for the negative sign, the MMI criterion is an upper bound to the true Bayes error rate independent of the model $p_\theta(k|x)$ used in the MMI criterion.

B. MCE Criterion

The model distribution for class k and observation x will be denoted by $p_\theta(k|x)$. In the limit of an infinite amount of training data, the MCE criterion [5] then is given by

$$\begin{aligned} F_{\text{MCE}} &= \lim_{N \rightarrow \infty} \frac{1}{N} \cdot \sum_{n=1}^N \frac{1}{1 + \left(\frac{p_\theta(k_n|x_n)}{r \sqrt{\sum_{k' \neq k_n} p_\theta^r(k'|x_n)}} \right)^{2\beta}} \\ &= \sum_k \int dx \cdot p(x, k) \cdot \frac{1}{1 + \left(\frac{p_\theta(k|x)}{r \sqrt{\sum_{k' \neq k} p_\theta^r(k'|x)}} \right)^{2\beta}} \\ &= \int dx \cdot p(x) \cdot \underbrace{\sum_k \frac{p(k|x)}{1 + \left(\frac{p_\theta(k|x)}{r \sqrt{\sum_{k' \neq k} p_\theta^r(k'|x)}} \right)^{2\beta}}}_{=: f_{\text{MCE}}(e|x)} \end{aligned}$$

where $f_{\text{MCE}}(e|x)$ denotes the local MCE error. To obtain a link between the local MCE criterion and the Bayes error, we need the following inequality:

$$\begin{aligned} &\frac{\sum_{k' \neq k} p_\theta^{2\beta}(k'|x)}{\left[\sum_{k'' \neq k} p_\theta^r(k''|x) \right]^{2\beta/r}} \\ &= \sum_{k' \neq k} \left[\frac{p_\theta^r(k'|x)}{\sum_{k'' \neq k} p_\theta^r(k''|x)} \right]^{2\beta/r} \leq 1 \quad \text{for } \frac{2\beta}{r} \geq 1. \end{aligned} \quad (2)$$

By using the fact that for $2\beta/r = 1$, the equality is true, it is easy to prove the inequality. Using this inequality, we obtain

$$\begin{aligned} f_{\text{MCE}}(e|x) &= \sum_k \frac{p(k|x)}{1 + \frac{p_\theta^{2\beta}(k|x)}{\left[\sum_{k' \neq k} p_\theta^r(k'|x) \right]^{2\beta/r}}} \\ &\geq \sum_k \frac{p(k|x)}{1 + \frac{p_\theta^{2\beta}(k|x)}{\sum_{k' \neq k} p_\theta^{2\beta}(k'|x)}} \\ &= \sum_k p(k|x) \cdot \frac{\sum_{k' \neq k} p_\theta^{2\beta}(k'|x)}{\sum_{k'} p_\theta^{2\beta}(k'|x)} \\ &= \sum_k p(k|x) \cdot \left[1 - \frac{p_\theta^{2\beta}(k|x)}{\sum_{k'} p_\theta^{2\beta}(k'|x)} \right] \\ &= 1 - \underbrace{\sum_k p(k|x) \cdot \frac{p_\theta^{2\beta}(k|x)}{\sum_{k'} p_\theta^{2\beta}(k'|x)}}_{=: f_{\text{GG}}(e|x)} \\ &\geq 1 - \max_c p(c|x) \cdot \sum_k \frac{p_\theta^{2\beta}(k|x)}{\sum_{k'} p_\theta^{2\beta}(k'|x)} \\ &= 1 - \max_k p(k|x) \\ &= p_{\text{Bayes}}(e|x) \end{aligned} \quad (3)$$

where the ‘‘intermediate’’ local error $f_{\text{GG}}(e|x)$ will be used to define a new discriminative criterion, which will be discussed in Section III-B. The aforementioned result shows that the MCE criterion gives an upper bound to the true Bayes error rate independent of the discrimination function model used in the MCE criterion. This result supplements the fact that the MCE criterion gives an approximation to the empirical error rate on the training data produced by the corresponding discrimination function model.

III. MODEL-FREE OPTIMIZATION

A. ML and MMI Criterion

For the ML and the MMI criterion, model-free optimization in the asymptotic case of infinite training data leads to closed form solutions. Under consideration of the normalization constraints for probability models, the model-free optimization of the ML and the MMI criterion leads to the true distributions

$$\begin{aligned} p_\theta(x|k) &= \frac{p(x, k)}{p(k)} = p(x|k) \quad (\text{ML criterion}), \\ p_\theta(k|x) &= \frac{p(x, k)}{p(x)} = p(k|x) \quad (\text{MMI criterion}). \end{aligned}$$

Therefore, under the assumption that the true *a priori* probability $p(k)$ is known, both the ML and the MMI criterion lead to Bayes' decision rule.

B. MCE and Related Criteria

For the asymptotic case of infinite training data, in this section, we will derive a closed-form solution for the model-free

optimization of both the original MCE criterion as well as for the GG criterion. The GG criterion is derived from the intermediate local error $f_{GG}(e|x)$ and gives even a closer upper bound to the true Bayes' error than the MCE criterion itself, cf. (3). Because of its similarity to the Gini criterion ([1], pp. 38), the new criterion will be called the GG criterion

$$F_{GG} = \int dx \cdot p(x) \cdot \sum_k p(k|x) \cdot \left[1 - \frac{p_\theta^{2\beta}(k|x)}{\sum_{k'} p_\theta^{2\beta}(k'|x)} \right].$$

In (3), a system of inequalities between the local MCE error $f_{MCE}(e|x)$, the local GG error $f_{GG}(e|x)$, and the true local Bayes' error rate $p_{\text{Bayes}}(e|x)$ is given. The equality, i.e., optimality for both the MCE and the GG criterion is obtained if the model distribution $p_\theta(k|x)$ is set to

$$\hat{p}_\theta(k|x) = \begin{cases} 1, & \text{if } k = \operatorname{argmax}_{k'} p(k'|x) \\ 0, & \text{otherwise.} \end{cases}$$

When we replace the true distribution $p(k|x)$ in Bayes' decision rule with this model distribution $\hat{p}_\theta(k|x)$, we obtain the important result that the decision about the unknown class remains the same

$$\operatorname{argmax}_k \hat{p}_\theta(k|x) = \operatorname{argmax}_k p(k|x).$$

Therefore, both the MCE criterion and the GG criterion lead to optimal error rate in the asymptotic case of infinite training data, although the structure of the corresponding model distributions differs considerably from the structure of the true distributions.

IV. DISCUSSION

Despite the fact that the MCE criterion only *approximates* the empirical error rate on the training data, in Section II-B we show that the MCE criterion gives an upper bound to the true Bayes' error rate. This result is independent of the discrimination function model used in the MCE criterion.

An optimization of the MCE criterion aims at improving the empirical error rate. Moreover, in the limit of $2\beta \rightarrow \infty$, the MCE criterion *equals* the empirical error rate on the training

data, and any discrimination function minimizing the empirical error rate gives a possible solution in this case. Nevertheless, for finite β , the *optimal* outcome of an MCE training might very well depend on the model choice. In this letter, for the case of infinite training data, a model-free optimization is performed for the MCE criterion, which leads to a closed form solution, which leads to the true Bayes' error rate. We found the same result for the GG criterion defined in this letter, which is shown to be similar to the MCE criterion, and which even gives a closer upper bound to the true Bayes' error rate than the MCE criterion itself.

Although being a function of the true distributions, the model distribution resulting from the model-free optimization of both the MCE and the GG criterion differs considerably from the corresponding true distribution and from the structure of the model distributions usually used for statistical pattern recognition. This result for the MCE criterion suggests that the structure of the models usually used for ML and MMI training might not be optimal for MCE training.

REFERENCES

- [1] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1984.
- [2] M. Ben-Bassat, "Use of distance measures, information measures and error bounds in feature evaluation," in *Handbook of Statistics*, P. R. Krishnaiah and L. N. Kanal, Eds. Amsterdam, The Netherlands: North Holland, 1982, vol. 2, pp. 773–791.
- [3] W. Chou, C.-H. Lee, and B.-H. Juang, "Minimum error rate training based on N -best string models," in *Proc. 1993 Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 2, Minneapolis, MN, Apr. 1993, pp. 652–655.
- [4] P. A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*. London, U.K.: Prentice-Hall, 1982.
- [5] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. Signal Processing*, vol. 40, pp. 3043–3054, Dec. 1992.
- [6] Y. Normandin, "Maximum mutual information estimation of hidden Markov models," in *Automatic Speech and Speaker Recognition*, C.-H. Lee, F. K. Soong, and K. K. Paliwal, Eds. Norwell, MA: Kluwer, 1996, pp. 57–81.
- [7] R. Schlüter, W. Macherey, B. Müller, and H. Ney, "Maximum mutual information and maximum likelihood approach for mixture density splitting," in *Proc. 1999 Eur. Conf. Speech Communication and Technology*, vol. 4, Budapest, Hungary, Sept. 1999, pp. 1715–1718.
- [8] V. Valtchev, J. J. Odell, P. C. Woodland, and S. J. Young, "MMIE training of large vocabulary recognition systems," *Speech Commun.*, vol. 22, pp. 303–314, Sept. 1997.