ANALYSIS AND COMPARISON OF DISCRIMINATIVE TRAINING OBJECTIVES

Qi Li

Li Creative Technologies, Inc.
New Providence, NJ 07974, USA
qili@ieee.org; www.lilabs.com

ABSTRACT

In this paper, the *minimum classification error* (MCE) and *maximum mutual information* (MMI) objectives for discriminative training in automatic speech recognition and natural language processing are analyzed and compared theoretically. The results show that both objectives are related to posterior probability and error rates, and the MCE objective is more general and flexible than the MMI objective. The relations between the objectives and parameter optimization methods are also discussed. The results can help in understanding the discriminative objectives, in developing new objectives, and in discovering new training algorithms jointly with objectives.

1. INTRODUCTION

Several objectives and corresponding algorithms for discriminative training have been applied to speech recognition, speaker recognition, and natural language processing successfully. It has been reported that the discriminative training techniques provide significant improvements in recognition performance compared to the traditional maximum likelihood objective in training classifiers or recognizers, especially for robust speech and speaker recognition problems. Two of the most popular objectives are the *maximum mutual information* (MMI) [1] and *minimum classification error* (MCE) [2, 3] objectives. There are also other objectives, such as the H-criteria [4] and a recently proposed one for fast discriminative training [5, 6]. Since MMI and MCE objectives are popular and have been used for many years, we focus our discussions on them in this paper. The results and methods can also be extended to analyze other objectives.

Since both MMI and MCE objectives have shown good experimental results, there have been many discussions and comparisons between these two objectives through experiments or some degree of theoretical analysis (e.g. [7, 8]); however, the experimental comparisons are limited to particular tasks and the results are not general enough to help us understand the detailed mechanisms; on the other hand, the previous theoretical analyses are not conclusive or adequate enough to show the relations between these two objectives. In this paper, we intend to provide a new and more general analysis on the discriminative training objectives. To ensure the results in this paper are general and conclusive for any recognition or classification tasks, we focus our analysis and comparisons in theory instead of experiments since an experimental comparison on any specific recognition tasks may bias the comparison and lead to incomplete conclusions.

In the following, we will first establish the relations between each one of the discriminative objectives to the posterior probability, and then use the posterior probability to facilitate the comparisons among the discriminative objectives. The well-known posterior probability as an objective for classification was derived from the concept of minimizing the classification risk [9]. We review the theory as follows to facilitate further discussions:

In an M-class classification problem, we are asked to make a decision to identify a sequence of observations, $\mathbf{x}$, as member of a class, say, C_i. The true identity of $\mathbf{x}$, say C_j, is not known, except in the design or training phase in which observations of known identity are used as reference for parameter optimization. We denote event α_i as the action of identifying an observation as class C_i. The decision is correct if $i = j$; otherwise, it is incorrect. It is natural to seek a decision rule that minimizes the probability of error, or empirically, the error rate, which entails a zero-one loss function:

$$\mathcal{L}(\alpha_i|C_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j. \end{cases} \qquad i, j = 1, ..., M \tag{1}$$

It assigns no loss to a correct decision and assigns a unit loss to an error. The probabilistic risk of α_i corresponding

Qi (Peter) Li was with Bell Labs, Lucent Technologies, Murray Hill, NJ 07974.

to this loss function is

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{M} \mathcal{L}(\alpha_i|C_j)P(C_j|\mathbf{x}) = \sum_{j\neq i} P(C_j|\mathbf{x}) = 1 - P(C_i|\mathbf{x}) \quad (2)$$

where $P(C_i|\mathbf{x})$ is the posteriori probability that $\mathbf{x}$ belongs to C_i. Thus, the zero-one loss function links the error rates to the posterior probability. To minimize the probability of error, one should therefore maximize the posterior probability $P(C_i|\mathbf{x})$. This is the basis of Bayes' maximum *a posteriori* (MAP) decision theory and is also referred to as *minimum error rate* (MER) [9] in an ideal setup.

We note that the posterior probability $P(C_i|\mathbf{x})$ is often modeled as $P_{\lambda_i}(C_i|\mathbf{x})$, a function defined by a set of parameters λ_i. Since the parameter set λ_i has a one-to-one correspondence with C_i, we write $P_{\lambda_i}(C_i|\mathbf{x}) = P(\lambda_i|\mathbf{x})$ and other similar expressions without ambiguity.

If we consider all M classes and all data samples from all classes, an objective for MER or maximum posterior probability can be defined as:

$$\max J(\Lambda) = \frac{1}{N}\sum_{k=1}^{M}\sum_{i=1}^{N_k} P(C_k|\mathbf{x}_{k,i}) \quad (3)$$

$$= \frac{1}{N}\sum_{k=1}^{M}\sum_{i=1}^{N_k} P(\lambda_k|\mathbf{x}_{k,i}) \quad (4)$$

where N_k is the total number of training data of class k, $N = \sum_{k=1}^{M} N_k$, and $\mathbf{x}_{k,i}$ is the ith feature vector of class k. Λ is a set of parameters for models or classifiers, $\Lambda = \{\lambda_k\}_{k=1}^{M}$.

2. MINIMUM CLASSIFICATION ERROR VS. POSTERIOR PROBABILITY

The minimum classification error (MCE) objective was derived through a systematic analysis. It introduced a misclassification measure to embed the decision process in the overall minimum classification error formulation. During the derivation, it was also considered that the misclassification measure is continuous with respect to the classifier parameters. The empirical average cost as the typical objective in the MCE algorithm was defined as [2, 3]:

$$\min L(\Lambda) = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{M} \ell_k(d_k(\mathbf{x}_i);\Lambda)1(\mathbf{x}_i \in C_k). \quad (5)$$

where M and N are the total numbers of classes and training data, and $\Lambda = \{\lambda_k\}_{k=1}^{M}$. It can be rewritten as:

$$\min L(\Lambda) = \frac{1}{N}\sum_{k=1}^{M}\sum_{i=1}^{N_k} \ell_k(d_k(\mathbf{x}_{k,i});\Lambda) \quad (6)$$

where N_k is the total number of training data of class k, $N = \sum_{k=1}^{M} N_k$, and $\mathbf{x}_{k,i}$ is the ith feature vector of class k. ℓ_k is a loss function and a sigmoid function is often used for it:

$$\ell_k(d_k) = \frac{1}{1+e^{-\zeta d_k+\alpha}}, \qquad \zeta > 0 \quad (7)$$

where d_k is a class misclassification measure defined as [10]:

$$d_k(\mathbf{x}) = -g_k(\mathbf{x};\Lambda) + \log\left[\frac{1}{M-1}\sum_{j\neq k}\exp[\eta g_j(\mathbf{x};\Lambda)]\right]^{1/\eta} \quad (8)$$

where $\mathbf{x} = \mathbf{x}_{k,i}$. When function $g(.)$ in (8) is a logarithm of probability as used in many applications, the class misclassification measure in (8) can be rewritten as:

$$d_k(\mathbf{x}) = -\log p(\mathbf{x}|\lambda_k) + \log\left[\frac{1}{M-1}\sum_{j\neq k} p(\mathbf{x}|\lambda_j)^{\eta}\right]^{1/\eta}. \quad (9)$$

When $\eta = 1$, we have

$$d_k(\mathbf{x}) = -\log\frac{p(\mathbf{x}|\lambda_k)}{\sum_{j\neq k}\frac{1}{M-1}p(\mathbf{x}|\lambda_j)}. \quad (10)$$

It can be further presented as:

$$d_k(\mathbf{x}) = -\log\frac{p(\mathbf{x}|\lambda_k)P_k}{\sum_{j\neq k} p(\mathbf{x}|\lambda_j)P_j} \quad (11)$$

where $P_k = 1$ and $P_j = \frac{1}{M-1}$, and they are similar to the *a priori* probability if we conduct a normalization.

To facility our further comparison, we convert the minimization problem to a maximization problem. Let

$$\tilde{d}_k(\mathbf{x}) = -d_k(\mathbf{x}) = \log\frac{p(\mathbf{x}|\lambda_k)P_k}{\sum_{j,j\neq k} p(\mathbf{x}|\lambda_j)P_j}, \quad (12)$$

and take it into the sigmoid function in (7). Assuming $\zeta = 1$ and $\alpha = 0$, we have

$$\ell_k(\tilde{d}_k) = \frac{1}{1+e^{-\tilde{d}_k}} \quad (13)$$

$$= \frac{p(\mathbf{x}|\lambda_k)P_k}{p(\mathbf{x}|\lambda_k)P_k + \sum_{j\neq k} p(\mathbf{x}|\lambda_j)P_j} \quad (14)$$

$$= \frac{p(\mathbf{x}|\lambda_k)P_k}{\sum_{j=1}^{M} p(\mathbf{x}|\lambda_j)P_j}. \quad (15)$$

Thus, the objective in (6) is simplified to:

$$\max \tilde{L}(\Lambda) = \frac{1}{N}\sum_{k=1}^{M}\sum_{i=1}^{N_k}\frac{p(\mathbf{x}_{k,i}|\lambda_k)P_k}{\sum_{j=1}^{M}p(\mathbf{x}_{k,i}|\lambda_j)P_j} \quad (16)$$

$$= \frac{1}{N}\sum_{k=1}^{M}\sum_{i=1}^{N_k}P(\lambda_k|\mathbf{x}_{k,i}). \quad (17)$$

This demonstrates that the MCE objective can be equal to the maximum posterior probability if we make the following assumptions:

$$P_k = 1 \quad (18)$$

$$P_j = \frac{1}{M-1} \quad (19)$$

$$\eta = 1 \quad (20)$$

$$\zeta = 1 \quad (21)$$

$$\alpha = 0. \quad (22)$$

Among the parameters, $P_k = 1$ and $P_j \leq 1$ imply that the MCE objective weighs the true class higher or equal to the competing classes. The parameter η plays a role of Holder norm in (8). By changing η, the weights between the true class and competing classes can be adjusted. The rest of the parameters, ζ and α, are related to the sigmoid function. α represents the shift of the sigmoid function. Since other parameters can play the similar role, α is usually set to zero. ζ is related to the slope of the sigmoid function. For different tasks and data distributions, different values of ζ can be selected to achieve the best performance.

3. MAXIMUM MUTUAL INFORMATION VS. MINIMUM CLASSIFICATION ERROR

The objective of *maximum mutual information* (MMI) was defined in [1] as:

$$I(k) = \log\frac{p(\mathbf{x}_{k,i}|\lambda_k)P_k}{\sum_{j=1}^{M}p(\mathbf{x}_{k,i}|\lambda_j)P_j}. \quad (23)$$

If we consider all M models and all data as in the above discussions, the complete objective for MMI training is:

$$\max I(\Lambda) = \sum_{k=1}^{M}\sum_{i=1}^{N_k}\log\frac{p(\mathbf{x}_{k,i}|\lambda_k)P_k}{\sum_{j=1}^{M}p(\mathbf{x}_{k,i}|\lambda_j)P_j} \quad (24)$$

$$= \sum_{k=1}^{M}\sum_{i=1}^{N_k}\log P(\lambda_k|\mathbf{x}_{\mathbf{k},\mathbf{i}}). \quad (25)$$

By comparing (17) and (25), we can observe that the difference between the simplified version of the MCE objective and the MMI objective is only in the logarithm. Since the logarithm is a monotonically increasing function, a procedure to optimize (25) is equivalent to optimize (17); therefore, the MMI objective in (25) can be represented as:

$$\max \tilde{I}(\Lambda) = \frac{1}{N}\sum_{k=1}^{M}\sum_{i=1}^{N_k}P(\lambda_k|\mathbf{x}_{k,i}) \quad (26)$$

which is equivalent to the simplified version of MCE objective in (17).

4. DISCUSSIONS

It has been argued that it is not intuitive how the MMI objective relates to the error rate [11]. From the above discussions, the answer is straight forward because we have linked the MMI objective to posterior probability and linked the posterior probability to error rates. If we want to further investigate the differences between the original MCE in (6) and MMI objectives in (25), the differences are mainly in the parameter set listed from (18) to (22). In theory, those parameters provide the flexibility to adjust the MCE object for different recognition tasks and data distributions; therefore, MCE object is a more general objective compared to the MMI and MER objectives. In practice, from many reported experiments, we know that some of the parameters can play an important role in recognition or classification performances. For example, η and ζ can be adjusted to achieve better performances.

For pattern recognition or classification, the objectives and optimization methods for parameter estimation are related to each other, and they both play important roles in solving real-world problems, in terms of recognition accuracy and training speed. For optimization methods, in general, closed-form formulas, like the EM algorithm in maximum likelihood estimation, for parameter re-estimation are more efficient than a gradient-descent kind of approach. However, not every objective has the closed-form formulas. When an objective is complicated, such as the MCE objective, it has less of a chance to derive closed-form formulas. Thus, the algorithm has to rely on gradient-descent methods. For the MMI objective defined in (25), a closed-form parameter re-estimation algorithm was derived in [12] through an inequality; however, there is a constant D in the algorithm and the value of the constant needs to be pre-determined for parameter estimation. Like the learning rate in gradient-descent methods, it is difficult to determine the value of D as reported in [11]. It is our belief that for the best performances in terms of recognition accuracy and training speed, the objective and optimization method

should be developed jointly. A new algorithm developed under this consideration has been presented in [5, 6].

We have used the posterior probability to facilitate the above analysis. Actually, the posterior probability is also the root of the *sum-squared error* objective in training multiple-layer perception (MLP) neural networks. It has been proved mathematically that the outputs of the MLP network units represent the posterior probability also [9].

5. CONCLUSIONS

The theoretical analysis in this paper indicate that the discriminative objectives used in speech recognition are related to the posterior probability and error rates. While the MMI is directly from the posterior probability, the MCE can be equivalent to the posterior probability under some assumptions on its parameters. The results from this paper show that the MCE objective is more general and flexible than the MMI objective. The flexibility may benefit real applications for different recognition tasks or data distributions.

6. ACKNOWLEDGEMENT

The author would like to thank Biing-Hwang Juang for useful discussions on discriminative training.

7. REFERENCES

[1] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Proc. IEEE ICASSP*, pp. 49–52, 1986.

[2] S. Katagiri, C.-H. Lee, and B.-H. Juang, "New discriminative algorithm based on the generalized probabilistic descent method," in *Proceedings of IEEE Workshop on Neural Network for Signal Processing*, pp. 299–309, September 1991.

[3] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. on Signal Processing*, vol. 40, pp. 3043–3054, Dec. 1992.

[4] P. S. Gopalarishnan, D. Kanevsky, A. Nadas, D. Nahamoo, and M. A. Picheny, "Decoder selection based on cross-entropies," in *Proc. IEEE ICASSP*, pp. 20–23, 1988.

[5] Q. Li and B.-H. Juang, "A new algorithm for fast discriminative training," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, May 2002.

[6] Q. Li and B.-H. Juang, "Fast discriminative training for sequential observations with application to speaker identification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, April 2003.

[7] W. Reichl and G. Ruske, "Discriminant training for continuous speech recognition," in *Proceedings of Eurospeech*, 1995.

[8] R. Schluter and W. Macherey, "Comparison of discriminative training criteria," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 493–497, 1998.

[9] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification, Second Edition*. New York: John & Wiley, 2001.

[10] W. Chou, "Discriminant-function-based minimum recognition error rate pattern-recognition approach to speech recognition," *Proceedings of the IEEE*, vol. 88, pp. 1201–1222, August 2000.

[11] Y. Normandin, R. Cardin, and R. D. Mori, "High-performance connected digit recognition using maximum mutual information estimation," *IEEE Trans. on Speech and Audio Processing*, vol. 2, pp. 299–311, April 1994.

[12] P. S. Gopalakrishnan, D. Kanevsky, A. Nadas, and D. Nahamoo, "An inequality for rational functions with applications to some statistical estimation problems," *IEEE Trans. on Information theoty*, vol. 37, pp. 107–113, Jan. 1991.