REVIEW OF MINIMUM CLASSIFICATION ERROR TRAINING IN DIMENSIONALITY REDUCTION

Ramasubramanian Sundaram

Department of Electrical and Computer Engineering
Mississippi State University
Mississippi State, MS 39762 USA
email: sundaram@isip.mstate.edu

ABSTRACT

Several modelling techniques are used in speech recognition to model the short term variations in a speech signal. These techniques generally use a high dimensional feature vector which could be correlated. Several classical techniques of discriminant analysis are used for reducing the dimensionality of the input feature without affecting the overall performance. One such approach is the minimum classification error (MCE) in which the misclassification probability is minimized based on a given set of training samples and the dimensionality is reduced. This review will focus on analyzing the theory behind MCE and its performance when compared to other discriminant analysis techniques like LDA.

1. INTRODUCTION

Modelling techniques for speech recognition rely on high dimensional feature vector to summarize a speech signal. Typically, recognizers use a front-end involving feature vectors of more than thirty dimensions. Correlations between features can arise when the signal is non stationary or is corrupted by noise. In such cases, the dimensionality of the feature vector can be effectively reduced without losing recognition accuracy. The reduced features can be chosen in two ways namely feature selection methods and feature extraction methods. Minimum Classification Error (MCE) algorithm is one such approach for reducing the dimensionality of the input feature vector involving discriminative techniques.

The need for reducing the dimensionality arises out of several factors. A large feature set may help in gaining accuracy but the size of the training data may be a potential problem. By increasing the dimensionality, the number of parameters that need to be trained increases. There may not be enough training data to account for all the parameters. Also, more features directly implies that the computational complexity increases [1]. It has been observed in practice that having more features, beyond a certain point, leads to a degradation in performance. This could be attributed to the fact that more features tend to overlap leading to more confusing classes and increase in classification errors.

This paper will provide a critical review of "Using Minimum Classification Error Training in Dimensionality Reduction" authored by Xuechuan Wang and Kuldip K. Paliwal. The paper discusses the MCE algorithm and the authors propose a new misclassification measure which performs better than the conventional MCE algorithm. The experimental results are also compared with LDA. The new MCE algorithm does a better job than LDA which has been validated by proper experiments. The paper also discusses some performance issues with respect to MCE. The paper is definitely worth reading if one needs some insight on Minimum Classification Error techniques.

The critical review is organized as follows. In Section 2, we first compare the Bayesian approach and discriminative approach for classification. Section 3 discusses the theory behind MCE. The alternative approach to MCE is detailed in Section 4 which is followed by critical analysis in Section 5.

2. CLASSIFICATION TECHNIQUES

Consider a set of observations $\chi = \{x_1, x_2, \ldots, x_N\}$, where each x_i is a K-dimensional vector and is known to belong to one of the *M* classes C_i, i = 1, 2,..., M. A classifier normally consists of set of parameters Λ and a decision rule. The popular Bayes classification

technique is based on prior knowledge. If the *a posteriori* probability $P_\Lambda(C_i/x)$ for each class is known then the Bayesian decision rule is given by

$$C(x) = C_i, \text{ if } P_\Lambda(C_i/x) = max_j P_\Lambda(C_j/x) \quad (1)$$

where C(.) denotes the classification operation. The rule is also written in terms of a priori and conditional probabilities. Since the probability measure is seldom known exactly, the Bayesian approach boils down to estimating the a priori and conditional probabilities. If the form of the distribution function is not known or if the training set is rather limited then Bayesian approach suffers from several drawbacks [3].

Another alternative to the Bayesian approach is to use discriminant functions. This requires defining a set of M discriminant functions, $g_i(x;\Lambda)$, one for each class. Unlike the Bayesian approach, the problem of "optimal" classifier design becomes that of finding the right parameter set for the discriminant functions to minimize the misclassification error. Techniques like Linear Discriminant Analysis, Minimum Classification Error fall into this category. The discriminative procedure offers implementational simplicity and it is also possible to circumvent training sample issues by using better classifier structure.

3. MINIMUM CLASSIFICATION ERROR

A linear discriminant function of a K-dimensional feature vector x has the form $w^*x + w_0$ where * denotes the matrix transpose. The weight vector and threshold, w and w_0, respectively, are defined for each class, resulting in M discriminant functions and a parameter set $\Lambda = \{w_1, w_{01}, w_2, w_{02}, \ldots, w_M, w_{0M}\}$ which constitutes a classifier. Each discriminant function can be written as

$$g_i(x;\Lambda) = w_i^*x + w_{0i} = \lambda_i^*y \quad (2)$$

where $\lambda_i^* = [w_i^*, w_{0i}]$ and $y^* = [x^*, 1]$. Since the discriminant functions are linear, the decision boundaries are hyperplanes.

The classifier parameters are to be determined based on a given sample set of N observations. During training, the labels associated with each class in the training data is assumed to be known. If weight vectors and thresholds exists such that the classification based on the discriminant function in (2) produces no error at all, the sample set is called linearly separable. In other words, linear separability means that hyperplanes exists such that the data for each class are separated without any overlap. To find the parameters for the discriminant function a misclassification measure is introduced and the misclassification error is minimized using a cost function by iterating over several passes.

A traditional misclassification measure is to use the Bayes discriminant defined as

$$d(x) = P(C_2/x) - P(C_1/x) \quad (3)$$

where $P(C_i/x)$ are the a posteriori probabilities and are assumed to be known. Equation (3) is defined for a two category case. This enumerates how likely it is that a class 1 observation is misclassified as a class 2 observation. For multi category cases (M > 2), the decision boundary is not straight forward and hence the misclassification measure is defined in terms of linear discriminant functions as

$$d_k(x) = -g_k(x;\Lambda) + \left[\frac{1}{M-1}\sum_{j\neq k} g_j(x;\Lambda)^\eta\right]^\eta \quad (4)$$

where η is a positive number. By varying η one can take all the potential classes into consideration. As η approaches infinity (4) becomes

$$d_k(x) = -g_k(x;\Lambda) + g_i(x;\Lambda) \quad (5)$$

where C_i is the class with the largest discriminant value among those classes other than C_k. Hence, $d_k(x) > 0$ implies misclassification and $d_k(x) \leq 0$ means a correct decision. In this way, the decision rule becomes a judgement based on a scalar value.

The misclassification measure is used to formulate the objective criterion in the form of a cost function which is defined as

$$l_k(x;\Lambda) = l_k(d_k(x)) \quad (6)$$

which is expressed as a function of the misclassification measure. Typically, the cost functions exponential or sigmoid functions given by (7) and (8) respectively

$$l_k(d_k) = \begin{cases} d_k^{\zeta}, d_k > 0 \\ 0, d_k \le 0 \end{cases} \quad (7)$$

$$l_k(d_k) = \frac{1}{1 + \exp(-\xi(d_k + \alpha))} \quad (8)$$

Both the functions are smoothed zero-one cost functions suitable for gradient algorithms. Equations (7) and (8) also suggests that if the classification is correct ($d_k(x) \le 0$), no cost is incurred while a misclassification leads to some penalty. The empirical average cost is defined as

$$L_0(\Lambda) = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{M} l_k(x_i;\Lambda) 1(x_i \in C_k) \quad (9)$$

where N is the number of input observation, M is the number of distinct classes and 1(x) is an indicator function and is given by

$$1(x) = \begin{cases} 1, \text{ if x is true} \\ 0, \text{ otherwise} \end{cases} \quad (10)$$

The goal of discriminative training using MCE is to minimize the cost function and gradient descent algorithms are used for this purpose following the adaptation rule given by

$$\Lambda_{t+1} = \Lambda_t - \varepsilon \nabla L_0(\Lambda_t) \quad (11)$$

where Λ_t denotes the parameter set at the t^{th} iteration and ε is the adaptation constant. The parameters can be updated for each training sample or after looking at the complete training set. The initial transformation matrix is updated as follows

$$T_{ij}(k+1) = T_{ij}(k) - \varepsilon \frac{\partial L_0}{\partial T_{ij}} \quad (12)$$

where $T_{ij}(k)$ is the transformation matrix at the end of k^{th} iteration and $T_{ij}(k+1)$ is the updated transformation matrix. The

4. ALTERNATIVE MCE TRAINING

The MCE training algorithm is fairly robust and performs satisfactorily when there are a few classes to discriminate. But when the number of classes is more with many dimensions, the performance starts to degrade. The reason behind this is that the MCE classification criterion tries to minimize the misclassification measure $g_k(x;\Lambda) - g_i(x;\Lambda)$. For the gradient descent algorithm to minimize the error, $g_k(x;\Lambda)$ should decrease while $g_i(x;\Lambda)$ should increase. Depending upon the training sample, it might happen that both might decrease or increase leading to errors in training. Hence, the misclassification measure is modified as follows

$$d_k(x) = \frac{g_k(x;\Lambda)}{\left[\frac{1}{M-1} \sum_{j \ne k} g_j(x;\Lambda)\zeta\right]^{1/\zeta}} \quad (13)$$

When ζ approaches ∞, (12) becomes

$$d_k(x) = \frac{g_k(x;\Lambda)}{g_i(x;\Lambda)} \quad (14)$$

The misclassification criterion now becomes: $d_k(x) < 1$ means correct classification and $d_k(x) \ge 1$ means incorrect classification. The gradient descent procedure remains the same as in (11) and (12)

5. RESULTS AND ANALYSIS

Experiments were performed to verify the MCE algorithm and compare it with Linear Discriminant Analysis. The conventional MCE algorithm is also compared with the new MCE procedure proposed by the authors. The database chosen was a vowel set containing 11 classes and the each feature vector had 10 dimensions.

The plots in the paper show that the new MCE algorithm proposed by the authors performs well on the training part of the dataset when compared with LDA or conventional MCE procedure. It performs consistently better for any of the reduced dimensions.

The conventional MCE procedure performs worse than the LDA. However, the performance of all the three discriminative techniques are not satisfactory on the test set. But again, the new MCE procedure performs better than LDA and the conventional MCE procedure.

The paper and the algorithm is not without its drawbacks. The authors fail to explain why the new classification measure in (13) is better than the one in conventional MCE as in (5). They claim that in the conventional MCE there is no constriction on the joint behavior of the discriminant functions. It appears equation (13) also does not take into account the joint behavior of the discriminant functions. It could happen that the denominator in (13) could decrease more rapidly than the numerator leading to an error in classification. That the algorithm works on the vowel database does not prove that it works for all cases. It might fail for some other database where LDA might perform better.

The initial values for the weight vectors and the thresholds poses a great challenge to the performance of MCE. The experiments performed in the paper by choosing an unity matrix as the initial transformation matrix. A change in the initial transformation matrix seems to give better results on both training and test set. Now the initial transformation matrix is obtained from a LDA training process. In other words, by doing an incremental training, the MCE algorithms seems to yield better results. This variability with respect to initial transformation matrix does not augur well for the algorithm in comparison representative training using HMM's where the performance does not change significantly for different initial parameters. This also poses a question about the robustness of the algorithm.

The convergence criteria for obtaining the weight vectors and thresholds needs some discussion. With the initial transformation matrix playing a crucial role on the performance, an improper choice of the initial matrix could lead to a complete non-convergence. Also, the experiments performed uses a Gradient Descent procedure while Probabilistic Descent Algorithm guarantees local convergence [2]. This also accounts for different initial transformation matrices.

The MCE algorithm does not give any information on optimal number of dimensions for the feature vector. So the proper choice of dimensionality is left to the user itself. Hence, to choose the best set of features that give better discrimination, the user needs to run training and testing on every possible dimensionality and choose the one that gives better results. Principal Components Analysis, which is also a dimensionality reduction technique, makes the reduction easier by specifying the contribution of each feature to the overall information. In MCE, choosing the optimal number of features is a process of trial and error.

6. CONCLUSIONS

This paper discussed the MCE training algorithm and provided a critical review of the paper "Using Minimum Classification Error Training in Dimensionality Reduction" authored by Xuechuan Wang and Kuldip K. Paliwal. The MCE algorithm is straight forward and easy to implement. The algorithm works satisfactorily on a medium sized training set and feature vectors. The paper that was critiqued failed to substantiate the claim that the newly proposed MCE procedure will yield better for all databases. The algorithm was not proven mathematically but was only proved in the basis of the experiments performed. Finally, the demerits of the algorithm and suggestions to improve it were provided.

REFERENCES

[1] Richard O. Duda, Peter E. Hart, David G. Stork, *Pattern Classification and Scene Analysis*, Wiley Interscience, 2000.

[2] B. H. Juang and S. Katagiri, "Discriminative Learning for Minimum Error Classification", *IEEE Transactions On Signal Processing*, vol. 40, pp. 3043-3054, Dec. 1992

[3] Lawerence K. Saul and Mazin G. Rahim, "Maximum Likelihood and Minimum Classification Error Factor Analysis for Automatic Speech Recognition", IEEE Transactions On Speech And Audio Processing, vol. 8, pp. 115-125, March 2000.