New Minimum Classification Error and Maximum Classification Accuracy Criteria for Bayesian Binary Hypothesis Testing Using Linear Matrix Transformations in GMM

Fernando D. Lorenzo-García¹, Juan L. Navarro-Mesa², Antonio Ravelo-García², Sofia Martín-Gonzalez²

¹Departamento de Ingeniería Telemática. ²Departamento de Señales y Comunicaciones.
Universidad de Las Palmas de Gran Canaria, Gran Canaria, Spain. fdlorenzo@dit.ulpgc.es

Abstract

We address the problem of improving the classification score in Bayesian binary hypothesis tests. We propose two methods to perform a dimensionality reduction by means of a linear matrix transformation from an original feature space to a new one. The criteria of the two methods is expressed in two different cost functions. The first function is based on the Bayes error and expresses the idea of minimizing the error and the second is based on the Chernoff distance and expresses the idea of maximizing the classification accuracy. In this study each class is modeled by means of a Gaussian mixture and the methods aim at finding a transformation matrix which is optimum depending on the selected criterion and its corresponding cost function. Once defined the cost functions we derive an iterative method which applies a steepest ascend strategy starting from an appropriate initialization. The experiments are carried out over a database of pathological speech and show that our approach represents an improvement in classification scores over other similar matrix transformations.

1. Introduction

The main problem in binary hypothesis testing is that classification scores degrade significantly when the classes are highly confusable. Various discriminative methods have been proposed in the literature to deal with this problem. Some methods use an optimization criteria based on mutual information [Bahl] or on minimum classification error [Chang] to estimate model parameters such that the classification scores were maximized. In [Loizou, 1999] an average divergence measure is used as criterion for finding a transformation matrix to map the original features into a more discriminative subspace to improve class separability performance. This and other approaches [Haeb-Umbach, Loizou 1996, Loizou 1999] use subspace projections by maximizing or minimizing an appropriately chosen cost function. Our work is developed in this context.

Our aims are either to minimize the classification error or to maximize the classification accuracy. These aims lead us to two different cost functions which through their maximization let to address a minimization/maximization of the classification error/accuracy. Thus, maximizing the cost functions is used here for estimating the matrix that maps the feature space into a new subspace. Once defined the cost functions, we apply the steepest ascent method for the estimation process. We use a Gaussian Mixture to model (GMM) each class as a probability density function (pdf). This is a classical choice in many applications of statistical pattern recognition. So, our methods could be adapted to these applications. Once estimated the transformation matrix, all vectors are mapped and we estimate the models of each class. It is in this space where the classification is performed.

We center our experiments in pathological speech classification. The vectors in the original space are formed by Mel-warped log-filterbank energies (MFE) features. A reference transformation matrix is the one based in the discrete cosine transform which is classic in the speech recognition literature. The experiments show that our approach represents an improvement in classification scores over similar matrix transformations.
2. Definition cost function based on Bayes error

The probability of error is a good measure of the performance of a decision rule [Fukunaga]. Let \( q_i(x) \) be the a posteriori probability of a class ‘i’ given a pattern ‘x’. Then, the conditional error given ‘x’, due to the Bayes decision rule is either \( q_i(x) \) or \( q_j(x) \) whichever smaller and the total error (Bayes error) can be estimated through the following expression \( E \{ r(x) \} \):

\[
\varepsilon = \min \{ p_i p_1(x), p_j p_2(x) \} dx = P \int p_i(x) dx + P \int p_j(x) dx
\]  

(2.1)

where \( p_i(x) = p(x/w_i) \) \( \{i=1,2\} \) and \( P_i = P(w_i) \) are the conditional probability density and the priori probability of the ‘i’ class, respectively. The integral regions \( L_1 \) and \( L_2 \) are the regions where ‘x’ is classified to \( \omega_1 \) and \( \omega_2 \) by this decision rule, and they are called the \( \omega_1 \) and \( \omega_2 \)-regions. In \( L_1 \), \( P_i p_1(x) > P_i p_2(x) \) and, therefore, \( r(x) = P_i p_1(x)/p(x) \). Likewise, \( r(x) = P_i p_1(x)/p(x) \) in \( L_2 \) because \( P_i p_1(x) < P_i p_2(x) \).

Now we define \( Y = \{(x',y'),..., (x',y')\} \) as a finite set of training instants, where each instant \( x' \) corresponds to the label \( y' = [1,2] \). Define \( A (k \times m) \) as a lineal matriz that maps the original observations \( x \) to a one transformed as \( y' = A^T x \), where \( x \) is a \( k \)-dimensional vector, \( y' \) is a \( m \)-dimensional vector and \( m \leq k \). We propose a cost function based on Bayes error (BEF, Bayes Error Cost Function) with the following expression:

\[
D = \sum_{i=1}^{N} \text{tanh} \left[ s \cdot \max \left\{ - \ln(p_i(x)), - \ln(p_j(x)) \right\} \right]
\]  

(2.2)

where the integral has been substituted by a summation because the training set is finite, the hyperbolic tangent has been introduced for convenience and \( s > 0 \) \( (=1/32) \) is a constant factor that controls the dynamic range of the argument. Notice that (2.2) use the maximum of \(-\ln()\), that is equivalent to search \( \min\{P_i p_1(x), P_i p_2(x)\} \) in (2.1). The resulting function is concave ascent and facilitates the search of a minimum error through a maximum of (2.2).

3. Definition cost function based on Chernoff distance

The Chernoff distance is a measure of similarity between two pdf’s. For example, each pdf may define the probability of pertaining to a given class, and therefore signifies how similar or different the two classes are. This distance can be defined as follows [5]:

\[
\bar{D} = \max_{\alpha \geq 1} \left\{ - \ln \left[ P_1^\alpha p_1^\alpha (x)p_2^{1-\alpha} p_2^{1-\alpha} (x) dx \right] \right\}
\]  

(3.1)

where \( p_i(X) = p(X/w_i) \) \( \{i=1,2\} \) and \( P_i = P(w_i) \) are the conditional probability density and the a priori probability of class ‘i’, respectively. Obviating the maximization in (3.1) with respect to \( \alpha \) and without loss of generality it is assumed to be constant hereafter and equal to \( \frac{1}{2} \) in the experiments. A maximum similarity between classes is seen as a tendency of \( \bar{D} \) to 0 while a minimum similarity is seen as tendency of \( \bar{D} \) to infinity. The important fact for our purposes is that the larger the distance \( \bar{D} \) is between the two distributions, the smaller the probability of misclassification between classes. In fact, the expression in (3.1) is often used to obtain an upper bound on the probability of misclassification such that the bigger the distance the smaller that probability. Thus, maximizing the Chernoff distance is used here as a key idea that will be further used for finding a transformation matrix from an original space to a transformed one where the classes are maximally separated.

The convex nature of (3.1) due to the application \(-\ln()\) over the integral is not appropriate to estimate the transformation matrix ‘A’ that maximizes the Chernoff distance. Instead, we propose a cost function as follows:

\[
D = \max_{\alpha, \beta} \text{tanh} \left[ s \cdot \ln \left( \sum_{t=1}^{N} P_1^\alpha p_1^\alpha (v') P_2^{1-\alpha} p_2^{1-\alpha} (v') \right) \right]
\]  

(3.2)

where again the integral has been substituted by a summation because the training set is finite, the hyperbolic tangent has been introduced for convenience and \( s > 0 \) \( (=1/32) \) is a constant factor that controls the dynamic range of the argument. Notice that now when the pdf’s of the two classes tend to overlap both \( \bar{D} \) and \( D \) tend to zero but when the overlap decreases then...
$D$ tends to one. The resulting function is concave increasing and facilitates the search for minimum overlapping (confusability) between classes by searching for the maximum of (3.2).

4. Cost function maximization

Let’s particularize for the case in which each class $j=\{1,2\}$ probability $p_j(X)$ is characterized by a mixture of $M$ Gaussian components with means $\mu^j$, covariance matrices $\Sigma^j$, weighting factors $\omega_j$ and $\{i=1,\ldots,M\}$. The objective is to obtain the matrix $A$ such that a given cost function is maximized. To do that we start by taking the partial derivatives of (2.2) with respect to $A$. Applying the chain rule the derivative first becomes

$$\frac{\partial D}{\partial A} = s \sum_{j=1}^{2} \tan^{-1} \left[ -L A \rho(v^j) \right] \left( \frac{\partial}{\partial A} \left[ -L A \rho(v^j) \right] \right)$$

(4.1)

where now $p(v^j)$ is either $p_j(v^j)$ or $p_{\bar{j}}(v^j)$ whichever smaller. For the cost function based in Chernoff distance we perform the same steps. Applying the chain rule to (3.2):

$$\frac{\partial D}{\partial A} = \left[ -L A \sum_{i=1}^{M} \rho_i(v) \rho_i^{-1}(v) \right] \left( \frac{\partial}{\partial A} \left[ -L A \sum_{i=1}^{M} \rho_i(v) \rho_i^{-1}(v) \right] \right)$$

(4.2)

Taking again the chain rule and making some derivations, the second factor in square brackets of the right hand side of equation (4.1) can be made equal to

$$\frac{\partial}{\partial A} \left[ -L A \rho(v) \right] = \frac{\partial}{\partial A} \left[ -L A \rho(v) \right] / p(v)$$

(4.3)

And the right hand of equation (4.2) can be made equal to

$$\frac{\partial}{\partial A} \left[ -L A \sum_{i=1}^{M} \rho_i(v) \rho_i^{-1}(v) \right] = \sum_{i=1}^{M} \frac{\partial \rho_i}{\partial A} \sum_{i=1}^{M} \rho_i(v) \rho_i^{-1}(v)$$

(4.4)

where the denominator in the right hand side of equation (4.3) y (4.4) acts as a normalization weight of the contribution to the derivative from each training vector $x^i$. The numerator has the partial derivatives with respect to matrix $'A'$. Since these derivatives include several terms and some of them are numerically negligible we will make the simplification that only the most important component from each mixture is taken into consideration. Thus, only single mixture components from each class are considered in (4.3) y (4.4). Substituting $v^j = A^T x^i$ in the expression of a normal pdf in the particular case of class $j=1$, the partial derivative with respect to the transformation matrix $'A'$ of a given component $'i'$ is:

$$\frac{\partial \rho_i^1}{\partial A^T} = \rho_i^1 \left[ A^T B_i^1 A - \alpha \Sigma_i A^T \Sigma_i A \right]^{-1} (A^T B_i^1 A + \Sigma_i A^T \Sigma_i A)^{-1}$$

(4.6)

where $\rho_i^1 = \rho_i^1 (x^i - \mu^1, x^i - \mu^1)^T$. The transformation matrix can be estimated by applying a steepest ascent method to estimate matrix $A$, i.e., given an initial matrix $A^{(0)}$ and update matrix $A^{(i)}$ in the following manner: $A^{(i+1)} = A^{(i)} + \gamma \partial D / \partial A$ for $i=0, \ldots, N_i$ where $i'$ is the iteration index, $N_i$ is the number of iterations, $\gamma = \gamma_0 (1 - I \gamma_0)$ is a step size that depends on the step and $\gamma_0 = (\gamma_0 - 1)$ is a small initialization constant.

Now we can make some considerations. The factor $(1 - \tanh^2[j])$ in (4.1) and (4.2) plays an interesting role because it is a reflection of the way in which the cost function progresses to its maximum when a new iteration is performed. Thus, this factor tends to zero making the derivative smaller step by step. The product between the smoothing factor 's' and the step size $\gamma$ act as a prevention from a fast progress to the maximum. The way in which the initialization matrix $A^0$ is estimated is open. In this paper expose the initialization of transformation matrix that obtain the best store [6]. The matrix $A$ which maximize the divergence among the classes can be build by selection of $m$ eigenvalue matrix of $\Sigma_i^{-1} \Sigma_i$ corresponding to de $m$ largest.

5. Experiments and results

For the classification experiments we used speech with (PS) and without pathology (NP). The database is composed of 54 speakers without pathology and 608 speakers with pathology from the Disordered Voice Database Model 4337. Recordings consisted in sustained vowel /ah/. A
half of the database was used for training and a half for testing. The original features were MFE obtained by applying \( m=20 \) triangular filters to the magnitude spectrum and the dimensionality in the transformed space is \( k=10 \). The speech waveforms are sampled at 25 KHz, and are blocked into 1500 samples from 30 msec. frames with 20 msec. of overlap between adjacent blocks. Each frame is passed through pre-emphasis filter and a Hamming window. Then, a 2048 points FFT is applied to the frame to produce a 1024-point power spectrum. The power spectra are combined using a weighted sum, shaped by the triangular filter, to obtain the filter output. Logarithms of the 20 outputs are then calculated arriving at 20 MFE for each frame. The whole set of training vectors is used to characterize each class by a mixture of Gaussian probability densities with \( M=7 \) components. For comparison purposes we have made experiments in the original space (MFE) and in the transformed space with MFCC, FEB y FEC. In tables 1 and 2 we show the classification scores. As we can see, both versions of method we propose outperform the MFE and MFCC.

<table>
<thead>
<tr>
<th>Method</th>
<th>Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFE</td>
<td>86.10%</td>
</tr>
<tr>
<td>MFCC</td>
<td>83.69%</td>
</tr>
<tr>
<td>BFC</td>
<td>93.33%</td>
</tr>
<tr>
<td>CFC</td>
<td>95.77%</td>
</tr>
</tbody>
</table>

TABLE 1. Classification score

<table>
<thead>
<tr>
<th>Method</th>
<th>Pathology</th>
<th>NP</th>
<th>PS</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFE</td>
<td>NS</td>
<td>74.07%</td>
<td>25.93%</td>
</tr>
<tr>
<td>MFE</td>
<td>PS</td>
<td>12.83%</td>
<td>87.17%</td>
</tr>
<tr>
<td>MFCC</td>
<td>NS</td>
<td>77.78%</td>
<td>22.22%</td>
</tr>
<tr>
<td>MFCC</td>
<td>PS</td>
<td>15.79%</td>
<td>84.21%</td>
</tr>
<tr>
<td>BFC</td>
<td>NS</td>
<td>62.96%</td>
<td>37.04%</td>
</tr>
<tr>
<td>BFC</td>
<td>PS</td>
<td>3.95%</td>
<td>96.05%</td>
</tr>
<tr>
<td>CFC</td>
<td>NS</td>
<td>77.78%</td>
<td>22.22%</td>
</tr>
<tr>
<td>CFC</td>
<td>PS</td>
<td>2.63%</td>
<td>97.37%</td>
</tr>
</tbody>
</table>

TABLE 2. Confusion matrices for different methods

6. Conclusions

We have formulated two new cost functions based in Bayes error and and we have given the formulae for an iterative maximization. The simplified version presented here is computationally less demanding than the original one while gives good classification scores. Obviating simplifications, from a theoretical point of view, in this paper we spread the scope of possibilities to obtain transformation matrices, but also all the parameters that parameterize the mixture components of each class. This and other related subjects will be a matter of future work in binary and M-ary hypothesis testing with a special emphasis in extending the formulation to hidden Markov models.

REFERENCES


