ESTIMACIÓN DE MATRICES DE TRANSFORMACIÓN MEDIANTE FUNCIONES DE COSTE BASADAS EN EL MÍNIMO ERROR DE BAYES Y SU APLICACIÓN A GMM

J. L. Navarro-Mesa[†], F. D. Lorenzo-García[*], A.G. Ravelo-García[†], S. I. Martín-González[†], P. J. Quintana-Morales[†], E. Hernández-Pérez[†]

Departamento de Señales y Comunicaciones[†]. Departamento de Ingeniería Telemática[*]

*Universidad de Las Palmas de Gran Canaria. Spain. †jnavarro@dsc.ulpgc.es, *fdlorenzo@dit.ulpgc.es*

ABSTRACT

Presentamos dos nuevos métodos para mejorar los resultados de clasificación en problemas de hipótesis binarias donde las clases son modeladas mediante mezclas de Gausianas. Partimos de una función de coste basada en el error de Bayes y a partir de la cota del error dada por Chernoff proponemos otra basada en la distancia del mismo nombre. A partir de ambas funciones se estima una matriz de transformación a un nuevo espacio de características que minimice el error de clasificación entre las clases. La estimación de las matrices se hace por el método de paso ascendente y se da la formulación resultante. Para ello, obtenemos las derivadas de las funciones y las aplicamos a una versión simplificada donde se selecciona una componente de cada mezcla participar en la estimación. Se estudia la inicialización del método y se propone el método de inicialización que mejores resultados nos ha proporcionado en los experimentos. Este método de inicialización está basado en una maximización de la divergencia entre las clases. Los experimentos se han realizado sobre una base de datos de voz con y sin patología y muestra que nuestra aproximación representa una mejora en los resultados de clasificación sobre otros métodos clásicos de diseño de matrices de transformación.

1. INTRODUCCIÓN

El problema de la discriminación entre clases es bien conocido y ha sido estudiado en múltiples trabajos. El principal problema es que los resultados de clasificación se degradan significativamente cuando las clases son altamente confundibles y, en consecuencia, el error es alto. Nuestro principal objetivo es abordar este problema en función de conseguir una baja tasa de error en un test de hipótesis binario. En la literatura se han propuesto varios métodos discriminantes para dar solución al problema de la confusión entre clases. Algunos métodos usan criterios optimizados basados en información mutua [1] o en error mínimo de clasificación [2] para calcular el modelo de parámetros tal que la separación entre las clases sea maximizada. En [3] se usa una medida ponderada de la divergencia como criterio para encontrar una matriz de transformación que mapee las características originales a un subespacio más discriminante. Esta y otras aproximaciones [4,3,6] usan proyecciones lineales que mapean las características del espacio original a un subespacio transformado maximizando o minimizando un criterio convenientemente formulado, p.e., maximizar la separación entre clases.

En este artículo partimos de la idea de que las propiedades estadísticas de las clases están recogidas mediante un modelo de mezcla de gaussianas. Estos modelos describirán las clases en un espacio de características original y mediante una matriz de transformación pasaremos a un nuevo espacio en el cual se cumplan unos objetivos expresados en unas funciones de coste. Una vez obtenida la matriz se transforman todos los vectores y se calculan los modelos de cada clase. Es en este espacio donde se realizará finalmente la clasificación.

Nuestro objetivo fundamental es minimizar el error de clasificación bayesiano. Para ello, partimos de una formulación clásica del mismo dando una primera función de coste y mostramos los elementos matemáticos principales. Dada la complejidad de las expresiones que se obtienen elegimos el método de paso descendente para estimar una matriz de transformación óptima. Una vez hecho esto aprovecharemos la desigualdad que relaciona el error de clasificación con la cota de Chernoff y, posteriormente, de ésta pasar a la distancia del mismo nombre [7]. Tomando como nuevo punto de referencia dicha distancia se llega a una nueva función de coste cuyo objetivo principal es el mismo, minimizar el error de clasificación, si bien en este caso se hará una maximización de dicha función. Las expresiones a las que se llegan también son complejas y utilizamos el método de paso ascendente para estimar una matriz de transformación óptima. Una idea clave de nuestro trabajo es que si bien con la primera función de coste pretendemos minimizar el error de clasificación a partir de su definición con la segunda minimizamos una cota del mismo.

Una vez definidas las funciones de coste y el método de estimación de las matrices centramos nuestro trabajo en los experimentos, particularmente, clasificación de habla con y sin patología. Los vectores en el espacio original representan características de tipo energía logarítmica obtenida a la salida de un banco de filtros (Mel-warped log-

Filterbank Energies, MFE). Entonces, el objetivo de los experimentos es comparar los métodos de estimación de las matrices de transformación a través de la tasa de error. A efectos de realizar comparaciones, como matriz de transformación de referencia elegimos el basado en la transformada de coseno discreta (MFCC) por ser un clásico en la literatura sobre reconocimiento del habla dados los buenos resultados que se obtienen.

2. DEFINICIÓN DE UNA FUNCIÓN DE COSTE BASADA EN EL ERROR DE BAYES

En general, cualquier regla de decisión que utilicemos para obtener una clasificación no va a dar una clasificación perfecta. Para evaluar el rendimiento de una regla de decisión debemos calcular la probabilidad del error, es decir, la probabilidad de que una muestra sea asignada a la clase equivocada. El error condicional dado una muestra 'x' es aquel en que la probabilidad a posteriori de la clase *i-ésima* dada 'x', $q_1(x)$ o $q_2(x)$, sea el menor. Es decir,

$$r(x) = \min[q_1(x), q_2(x)] \tag{2.1}$$

El error total, llamado error de Bayes, se obtiene calculando la siguiente expresión $E\{r(x)\}$.

$$\begin{aligned} \varepsilon &= E\{r(x)\} = \int r(x)\,p(x)dx \\ &= \int \min[P_1 p_1(x), P_2 p_2(x)]dx \\ &= P_1 \int_{L_2} p_1(x)dx + P_2 \int_{L_1} p_2(x)dx \end{aligned} \tag{2.2}$$

donde $p_i(x)= p(x/w_i)$ $\{i=1,2\}$ y $P_i=P(w_i)$ son la densidad de probabilidad condicional y la probabilidad a priori de la clase *'i'*, respectivamente. La ecuación (2.2) muestra varias formas de expresar el error de Bayes, ε. La primera línea es la definición del ε. La segunda línea se obtiene insertando la ecuación (2.1) en la primera línea y aplicando el teorema de Bayes. Las regiones integrales L_1 y L_2 de la tercera línea son las regiones donde 'x' se clasifica como ω_1 y ω_2 por esta regla de decisión, y se llaman las regiones ω_1 y regiones ω_2. En L_1 $P_1p_1(x) > P_2p_2(x)$ y, por tanto, $r(x) = P_2p_2(x) / p(x)$. Asimismo, $r(x) = P_1p_1(x) / p(x)$ en L_2 porque $P_1p_1(x) < P_2p_2(x)$. Desde un punto de teórico práctico nos centraremos en la segunda línea mientras que desde un punto de vista práctico los haremos sobre la tres pues nos dice que hemos de evaluar el error ε sobre los ejemplos mal clasificados.

Ahora definimos $Y=\{(x^1,y^1),\ldots,(x^N,y^N)\}$ como un conjunto finito de instantes de entrenamiento, donde a cada instante x^t corresponde a la etiqueta $y^t=\{1,2\}$. Definimos A $(k \times m)$ como una matriz lineal que mapea las observaciones originales x^t a uno transformado como $v^t = A^T x^t$, donde x^t es un vector k-dimensional, v^t es un vector m-dimensional y $m \leq k$. La naturaleza de (2.2, línea 2) no es apropiada para obtener una matriz de transformación *'A'* que maximice la separación entre las clases. Como alternativa proponemos una función de coste basada en el error de Bayes (FEB, Función de Error de Bayes) como se muestra a continuación:

$$D = \sum_{t=1}^{N} \tanh\left\{ s \cdot \max_{\alpha, A}\left[-Ln(P_1 p_1(v^t)), -Ln(P_2 p_2(v^t)) \right] \right\} \tag{2.3}$$

donde la integral ha sido sustituida por un sumatorio porque el conjunto de entrenamiento es finito, se ha introducido la tangente hiperbólica por conveniencia y $s>0$ es un factor constante que controla el rango dinámico del argumento. Observar que en (2.3) utilizamos el máximo de $-Ln(\cdot)$, que es equivalente a buscar $min[P_1p_1(x),P_2p_2(x)]$ en (2.2). La función resultante es cóncava creciente y facilita la búsqueda de un mínimo error entre clases mediante búsqueda de un máximo de (2.3). Esta búsqueda se hará con un método de paso ascendente que será presentado en la sección 4.

3. FUNCIÓN DE COSTE BASADA EN LA DISTANCIA DE CHERNOFF

La distancia de Chernoff es una medida de similitud entre dos funciones de densidad de probabilidad (pdf). Por ejemplo, cada pdf podría definir la probabilidad de pertenecer a una clase dada, y por tanto significa cuan similar o diferente son las dos clases. La distancia se puede definir como sigue [5]:

$$\overline{D} = \max_{0 \leq \alpha \leq 1}\left\{ -Ln\left(\int P_1^{\alpha} p_1^{\alpha}(x) P_2^{1-\alpha} p_2^{1-\alpha}(x)dx \right) \right\} \tag{3.1}$$

Obviando la maximización en (3.1) con respecto a α, y sin pérdida de generalidad, se asume que es una constante de ahora en adelante y es igual a ½ en los experimentos. Para nuestro propósito, el aspecto importante es que cuanto mayor sea la distancia $\overline{D}$ entre las dos distribuciones será la probabilidad de error de clasificación entre las clases. De hecho, a menudo la expresión (3.1) se usa para obtener una frontera superior en la probabilidad de error de clasificación tal que a mayor distancia, menor probabilidad.

La naturaleza convexa de (3.1) debido a la aplicación del $-Ln[.]$ sobre la integral no es apropiada para obtener la matriz de transformación *'A'*, por lo que usaremos la siguiente función de coste [7] (FEC, Función de Error de Chernoff):

$$D = \max_{\alpha, A} \tanh\left\{ -s \cdot Ln\left[\sum_{t=1}^{N} P_1^{\alpha} p_1^{\alpha}(v^t) P_2^{1-\alpha} p_2^{1-\alpha}(v^t) \right] \right\} \tag{3.2}$$

donde el sumatorio, la tangente hiperbólica y $s>0$ tienen el mismo significado de la sección anterior. Téngase en cuenta que ahora cuando las pdf de las dos clases tiendan a solaparse, $\overline{D}$ y D tienden a cero pero cuando el solapamiento disminuye entonces D tiende a uno. La

función resultante es cóncava creciente y facilita la búsqueda de mínimo solapamiento (confusión) entre clases mediante búsqueda de un máximo. Esta búsqueda se hará con un método de paso ascendente que será presentado en la siguiente sección.

4. MAXIMIZACIÓN DE LAS FUNCIONES DE COSTE

Vamos a particularizar para el caso en el que cada probabilidad $p_j(x)$ de la clase $j=\{1,2\}$ se representa por una mezcla de M Gausianas con medias μ_j^i, matrices de covarianzas Σ_j^i, factores de peso ω_j^i y $\{i=1,\ldots,M\}$. El objetivo es obtener la matriz de transformación A tal que la función de coste se minimiza o maximiza según proceda ya que está asociada a un error de clasificación mínimo. Empezamos por realizar la derivada parcial de (2.3) con respecto de A. Aplicando la regla de la cadena obtenemos

$$\frac{\partial D}{\partial A} = s\sum_{t=1}^{N} 1-\tanh^2\left[-s\cdot Ln\left(p(v^t)\right)\right]\left[\frac{\partial}{\partial A}\left(-Ln\left(p(v^t)\right)\right)\right] \quad (4.1)$$

donde $p(v^t)$ es la función de densidad de probabilidad mayor obtenida de la ecuación (2.3) dado v^t.

Para la función de coste basada en la distancia de Chernoff realizamos los mismos pasos. Aplicando la regla de la cadena a (3.2) con respecto de A obtenemos

$$\frac{\partial D}{\partial A} = s\left[1-\tanh^2\left(-s\cdot Ln\left(\sum_{t=1}^{N} p_1^{\alpha}(v^t)p_2^{1-\alpha}(v^t)\right)\right)\right]\left[\frac{\partial}{\partial A}\left(-Ln\left(\sum_{t=1}^{N} p_1^{\alpha}(v^t)p_2^{1-\alpha}(v^t)\right)\right)\right] \quad (4.2)$$

Ahora vamos a desarrollar en más detalle la derivada parcial en el lado derecho de las ecuaciones (4.1) y (4.2). Tomando de nuevo la regla de la cadena y derivando el logaritmo de esta forma $\partial Ln[g(x)]/\partial x = g'(x)\,\partial Ln[g(x)]/\partial[g(x)]$, el segundo factor entre corchetes del lado derecho de la ecuación (4.1) se puede igualar a

$$\frac{\partial}{\partial A}\left(-Ln\left(p(v^t)\right)\right) = \frac{\partial/\partial A(p(v^t))}{p(v^t)} \quad (4.3)$$

Y el lado derecho de la ecuación (4.2) se puede igual a

$$\frac{\partial}{\partial A}\left(-Ln\left(\sum_{t=1}^{N} p_1^{\alpha}(v^t)p_2^{1-\alpha}(v^t)\right)\right) = \sum_{t=1}^{N}\left[\frac{\partial/\partial A(p_1^{\alpha}(v^t)p_2^{1-\alpha}(v^t))}{\sum_{t=1}^{N} p_1^{\alpha}(v^t)p_2^{1-\alpha}(v^t)}\right] \quad (4.4)$$

donde el denominador en parte derecha de las ecuaciones (4.3) y (4.4) actúa como peso normalizador de la contribución de la derivada para cada vector de entrenamiento x^t. Debido a que estas derivadas tienen muchos términos y algunos de ellos son numéricamente insignificantes tomaremos la simplificación de tomar en cuenta solo los componentes más importantes de cada mezcla. Así, sólo se considera una mezcla por clase en (4.3) y (4.4).

La expresión de la componente *'i'* de la clase *'j'* en el espacio transformado se puede expresar como:

$$p_j^i(x^t, A) = \frac{1}{(2\pi)^{n/2}\left|A^T\Sigma_j^i A\right|^{1/2}}\exp\left\{-\frac{1}{2}\left(x^t-\mu_j^i\right)^T A\left(A^T\Sigma_j^i A\right)^{-1}A^T\left(x^t-\mu_j^i\right)\right\} \quad (4.5)$$

Por ejemplo, en el caso particular de la clase $j=1$, la derivada parcial con respecto a la matriz de transformación 'A' de un componente dado *'i'* es:

$$\frac{\partial p_1^{i\,\alpha}(A)}{\partial A} = p_1^{i\,\alpha}(A)\left[\alpha\cdot B_1^i A-\alpha\cdot\Sigma_1^i A\left(A^T\Sigma_1^i A\right)^{-1}\left(A^T B_1^i A\right)+\Sigma_1^i A\left(A^T\Sigma_1^i A\right)^{-1}\right] \quad (4.6)$$

donde $B_j^i = A^T\left(x^t-\mu_j^i\right)\left(x^t-\mu_j^i\right)^T A$. Note que (4.6) se calcula para cada vector de entrenamiento x^t y, por tanto, cada uno de ellos influirá directamente en el cálculo de A.

La matriz de transformación se podría obtener resolviendo la ecuación de igualar a cero la ecuación (4.1) o (4.2), lo cual es difícil. En cambio, aplicaremos un método de paso ascendente para encontrar la matriz A, p.e., dada una matriz inicial A^0 y una matriz actualizada $A^{(l)}$ de la siguiente manera:

$$A^{(l+1)} = A^{(l)} + \gamma_l \left.\frac{\partial D}{\partial A}\right|_{A=A^{(l)}} \quad (4.7)$$

para $\{l=1,2,\ldots, N_l\}$ donde *'l'* es el índice de iteración, N_l es el número de iteraciones, $\gamma_l = \gamma_0\,(1-l/N_l)$ es el tamaño del paso que depende del paso y γ_0 es la inicialización de una constante que ha sido puesta a 0'1 en los experimentos.

Ahora podemos tomar algunas consideraciones. El factor $(1-\tanh^2[.])$ en (4.1) y (4.2) juega un papel interesante porque es un reflejo de la forma en que la función de coste progresa hacia un máximo cuando se lleva a cabo una nueva iteración. Así, este factor tiende a cero conforme se avanza hacia la convergencia. El producto entre el factor de suavizado *'s'* y el tamaño del paso γ_l actúa como prevención de un progreso rápido hacia un máximo.

La forma en que se inicializa la matriz A^0 está abierta. En los métodos iterativos la inicialización es muy importante porque afecta a la evolución hacia un máximo global o local. En este artículo exponemos aquella inicialización de la matriz de transformación que mejores resultados nos ha dado [6]. La matriz A que maximiza la divergencia entre las clases se puede construir seleccionando los m autovectores de la matriz $\Sigma_2^{-1}\Sigma_1$ correspondiente a los m mayores autovalores, donde Σ_i es la matriz de covarianza de las Gausianas más cercanas entre sí.

4. EXPERIMENTOS Y RESULTADOS

Para los experimentos de clasificación utilizamos habla con y sin patología. Usamos la base de datos Disordered Voice Database Model 4337 compuesta de 54 locutores sin patología y 608 locutores con patología. Las grabaciones consisten en la vocal sostenida /ah/. La mitad de la base de datos se usó para el entrenamiento y la otra mitad para el test. Las características originales MFE fueron obtenidas aplicando m=20 filtros triangulares dando lugar a una dimensión de 20 en el espacio original. La dimensionalidad del espacio transformado es k=10. La forma de onda está muestreada a 25 KHz, se toman tramas de 30 msec. Con un solape de 20 msec. entre bloques adyacentes. Cada trama se pasa a través de un filtro pre-énfasis con un factor de 0'95 y se aplica una ventana Hamming. Entonces, se hace una FFT de 2048 puntos para producir 1024 puntos del espectro de energía. A este se le aplican los *'m'* filtros triangulares para obtener la energía logarítmica. La totalidad del conjunto de vectores de entrenamiento se usa para representar cada clase en el espacio original mediante una mezcla de Gausianas con M=7 componentes. A efectos de comparación hemos realizado experimentos en el espacio original (MFE) y en el espacio transformado con MFCC, FEB y FEC con la inicialización introducidas en la sección 4. Hemos encontrado que s=1/32 es una buena elección como factor de control en (2.3). En la tabla 1 mostramos los resultados de clasificación para los cuatro métodos mencionados en la sección anterior. Como puede verse, ambas versiones del método que proponemos mejora a MFE y MFCC.

Method	Scores
MFE	*86.10%*
MFCC	*83.69%*
FEB	*93.35%*
FEC	*95.77%*

Tabla 1. Resultados de clasificación para diferentes métodos

Es interesante observar que la transformación MFCC no mejora los resultados de clasificación en comparación con MFE como se podría esperar de la literatura. Los resultados obtenidos a partir de nuestras transformaciones son realmente mejores.

Metodo	Patología	NP	HP
MFE	NP	*74.07%*	*25.93%*
MFE	*HP*	*12.83%*	*87.17%*
MFCC	NP	*77.78%*	*22.22%*
MFCC	*HP*	*15.79%*	*84.21%*
FEB	NP	*62.96%*	*37.04%*
FEB	*HP*	*3.95%*	*96.05%*
FEC	NP	*77.78%*	*22.22%*
FEC	*HP*	*2.63%*	*97.37%*

Tabla 2. Matrices confusión para diferentes métodos

Ya que la cantidad de ejemplos para el habla no patológica es pequeña, en la tabla 2 vemos la matriz de confusión de los diferentes métodos entre habla patológica (HP) y no patológica (NP).

Podemos ver que el método de Chernoff ofrece los menores resultados de confusión incluso para la clase NS que tiene menos cantidad de ejemplos comparados con la clase PS.

5. CONCLUSIONES

Hemos formulado dos funciones de coste basada en el error de Bayes y hemos dado la formula para una maximización iterativa. La versión simplificada presentada aquí tiene menos demanda computacional que el original mientras que da unos resultados de clasificación buenos. Obviando simplificaciones, desde el punto de vista teórico, en este artículo abrimos el ámbito de las posibilidades de obtener matrices de transformación, pero además todos los parámetros que caracterizan las mezclas de componente de cada clase. Este y otros temas relacionados se tratarán en el trabajo futuro en test de hipótesis binaria y M-aria con un énfasis especial en extender la formulación a modelos ocultos de Markov.

6. REFERENCIAS

[1] L.R. Bahl, P.F. Brown, P.V. Souza, & R.L. Mercer, "Maximum mutual information estimation of hidden Markov models for speech recognition". Proc ICASSP'86, pp 49-52, 1986.

[2] P.C. Chang & B-H. Juang, "Discriminative training of dynamic programming based speech recognizers". IEEE Trans. Speech Audio Processing, Volume 1, number 2, pp 135-143, 1993.

[3] P.C. Loizou, & A.S. Spanias, "Improved speech recognition using a subspace projection approach". IEEE Transactions on Speech and Audio Processing, Volume: 7, Issue: 3, pp 343–345, 1999.

[4] R. Haeb-Umbach & H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition". Proc ICASSP'92, pp 13-16, 1992.

[5] Fukunaga, K, "*Introduction to Statistical Pattern Recognition*". Academic Press, Inc. 2nd Edition, 1990.

[6] P.C. Loizou, & A.S. Spanias, "High-performance alphabet recognition". IEEE Transactions on Speech and Audio Processing, Volume: 4, pp 430-445, 1996.

[7] F.D. Lorenzo, A.G. Ravelo, J.L. Navarro, S.I. Martín, P.J. Quintana, E. Hernández. "A Chernoff-based Approach to the Estimation of Transformation Matrices for Binary Hypothesis Testing". IEEE-ICASSP, pp V-753 – V-756