ESTIMACIÓN DE PROBABILIDADES A POSTERIORI EN SVMS MULTICLASE PARA RECONOCIMIENTO DE HABLA CONTINUA

Rubén Solera Ureña, Fernando Pérez-Cruz, Fernando Díaz de María

Universidad Carlos III de Madrid
Departamento de Teoría de la Señal y Comunicaciones
Avda. de la Universidad, 30, 28911-Leganés (Madrid), ESPAÑA

RESUMEN

El presente artículo aborda el problema de la estimación de probabilidades en el marco del reconocimiento automático de habla mediante Máquinas de Vectores Soporte (SVMs). Aunque las SVMs están pensadas para tareas de clasificación, en ciertas aplicaciones, especialmente problemas multiclase, es conveniente conocer el nivel de incertidumbre asociado a la decisión del clasificador o la probabilidad a posteriori de cada una de las etiquetas posibles. Durante los últimos años se han desarrollado diversos métodos para estimar dichas probabilidades a partir de las salidas blandas de las SVMs; sin embargo, la base teórica subyacente no está suficientemente justificada. En este artículo se revisan las debilidades de algunos de estos métodos y se presentan diversas alternativas basadas en un uso más directo de las salidas de las SVMs.

1. INTRODUCCIÓN

Las Máquinas de Vectores Soporte [1] constituyen el estado del arte en tareas de clasificación no lineal. Sus ventajas frente a otros clasificadores tradicionales han llamado la atención de numerosos investigadores en diversas áreas, entre ellas la de reconocimiento automático de habla. Fundamentalmente, hay tres razones que han conducido a plantear el uso de las SVMs como posible alternativa a los Modelos Ocultos de Markov (HMMs) en reconocimiento de voz: a) el modelado acústico es fundamentalmente un problema de clasificación de patrones, mientras que los HMMs son modelos generativos y, por tanto, lo que buscan es estimar las funciones de densidad de probabilidad correspondientes a cada uno de los modelos; b) la función de coste definida por las SVMs establece un compromiso entre la minimización del riesgo empírico y del riesgo estructural, lo que produce máquinas con una mayor capacidad de generalización frente a otros clasificadores tradicionales; c) el entrenamiento de las SVMs busca la maximización del margen, definido como la distancia de las muestras de entrenamiento a la frontera de decisión, lo que a priori les hace más robustas frente a condiciones ruidosas, uno de los principales problemas en reconocimiento de habla.

No obstante, la aplicación de las SVMs en reconocimiento de habla no es inmediata, debido principalmente a que estas máquinas trabajan con vectores de entrada de dimensión fija. Por el contrario, los métodos más comunes de parametrización de la voz producen secuencias de parámetros de longitud variable, dependiendo de la duración de la locución. En los últimos años han aparecido diversas formas de abordar este problema. El sistema de reconocimiento de habla continua mediante SVMs que se presenta en [2] lo evita trabajando trama a trama. El reconocedor propuesto emplea una SVM multiclase para determinar a qué clase pertenece el segmento de voz considerado y, a continuación, se usa el algoritmo de Viterbi para obtener una secuencia de palabras a partir de las decisiones acústicas que proporciona la SVM.

Como es sabido, los HMMs contribuyen al proceso de reconocimiento de la secuencia de parámetros espectrales de entrada aportando la verosimilitud de que la muestra actual haya sido generada por cada uno de los modelos que constituyen el sistema de reconocimiento de voz (palabras, fonemas, trifonemas...). Las SVMs, en cambio, proporcionan la etiqueta de la clase asignada al vector de entrada. Para aquellas aplicaciones en las que el clasificador sólamente se encarga de una parte de la decisión global, se han propuesto diversas formas para estimar probabilidades *multiclase* a partir de las salidas blandas de las SVMs [3, 4]. Su planteamiento depende principalmente de la estrategia multiclase que se adopte en la SVM (*1-contra-1* ó *1-contra-el resto*), las cuales se detallarán en la siguiente sección. Una de las más usadas, empleada en [2], se basa en el cálculo de la probabilidad de Platt [5] para cada SVM binaria (en la aproximación *1-contra-1*) y el empleo de una variante del método de Refregier-Vallet para obtener las probabilidades *multiclase* [3].

No obstante, el fundamento teórico de los métodos citados no está suficientemente justificado, tal y como muestran los experimentos realizados. Su principal debilidad consiste en la hipótesis de gaussianidad en las funciones de densidad de probabilidad condicional de la salida de las SVMs binarias, dada una cierta clase ($p(f|y=+1)$ y $p(f|y=-1)$, donde f denota la salida blanda de la SVM binaria para la muestra actual y $+1$ y -1 son las etiquetas asociadas a las dos clases consideradas). En este artículo se exploran diversas alternativas basadas en el uso directo de las salidas blandas de las SVMs, evitando en lo posible

hipótesis como la señalada anteriormente.

El artículo está estructurado de la siguiente forma: la sección 2 presenta los conceptos básicos de las Máquinas de Vectores Soporte y repasa las formas más comunes de estimar probabilidades *multiclase* a partir de las salidas de las SVMs. En la sección 3 se presentan nuestras propuestas para estimar las probabilidades a posteriori a partir de la salida blanda de las SVMs. En la sección 4 se describe a grandes rasgos el sistema de reconocimiento de habla empleado en los experimentos ([2]) y se muestran los resultados obtenidos sobre una tarea de reconocimiento de dígitos conectados con el fin de evaluar las prestaciones de los métodos propuestos.

2. MÁQUINAS DE VECTORES SOPORTE

2.1. Fundamentos de las SVMs

Una SVM es un clasificador binario que asigna una etiqueta $y \in \{+1, -1\}$ al vector de entrada $\mathbf{x}$ conforme al signo de la siguiente expresión:

$$f(\mathbf{x}) = \mathbf{w}^T \cdot \phi(\mathbf{x}) + b, \tag{1}$$

donde $\phi(\mathbf{x}) : \Re^d \mapsto \Re^H$ ($d << H$) es una transformación del espacio de entrada a un espacio *de características* de mayor dimensión (incluso infinita), en el que se supone que las clases son linealmente separables. El vector $\mathbf{w}$ define el hiperplano de separación en dicho espacio y b representa el sesgo respecto al origen de coordenadas.

La razón que hace a las SVMs más robustas que otros clasificadores es su criterio de entrenamiento, consistente en un compromiso entre la minimización del riesgo empírico y del riesgo estructural. Éste último evita un posible sobreajuste de la máquina al conjunto de entrenamiento. La solución viene dada por el siguiente problema de minimización cuadrática:

$$\begin{aligned} \min_{\mathbf{w},b,\xi_i} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i, \\ \text{sujeto a} \quad & y_i(\mathbf{w}^T \cdot \phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0;\ i = 1, \cdots, n, \end{aligned} \tag{2}$$

donde $\mathbf{x}_i \in \Re^d$ ($i = 1, \ldots, n$) son las muestras de entrenamiento con etiquetas $y_i \in \{+1, -1\}$. Las variables ξ_i miden el error de entrenamiento de cada muestra y C es un factor de ponderación entre el riesgo empírico y el riesgo estructural. Para más detalles sobre la SVM se puede consultar [6, 1, 7].

El problema de programación cuadrática en (2) se suele resolver empleando el dual de Wolfe [8], en el que los multiplicadores de Lagrange α_i se calculan maximizando:

$$\begin{aligned} \max_{\alpha_i} \quad & \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j \phi^T(\mathbf{x}_i)\phi(\mathbf{x}_j) \\ \text{sujeto a} \quad & \sum_{i=1}^{n}\alpha_i y_i = 0,\ 0 \leq \alpha_i \leq C \end{aligned} \tag{3}$$

Este problema es cuadrático y convexo, por lo que la convergencia al mínimo global está asegurada. Una vez resuelto, el vector de pesos $\mathbf{w}$ se puede expresar de la siguiente forma:

$$\mathbf{w} = \sum_{i=1}^{n}\alpha_i y_i \phi(\mathbf{x}_i). \tag{4}$$

Sólo aquellas muestras cuyo multiplicador asociado α_i sea distinto de 0 contibuyen a la definición de la frontera de decisión, razón por la que reciben el nombre de *vectores soporte*.

Normalmente, la función $\phi(\mathbf{x})$ no se conoce de forma explícita o es imposible de evaluar. No obstante, esto no plantea ninguna dificultad, ya que si nos fijamos en las expresiones (1) y (3) veremos que únicamente se precisa calcular los productos escalares $\phi^T(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$, los cuales, empleando lo que se ha denominado *truco del kernel* [1], se pueden evaluar mediante la función de *kernel* $K(\mathbf{x}_i, \mathbf{x}_j)$. De esta forma, la salida blanda de la SVM adopta la siguiente expresión:

$$f(\mathbf{x}) = \sum_{i=1}^{n}\alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b. \tag{5}$$

Los *kernels* más comunes son el lineal (K_L):

$$K_L(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \cdot \mathbf{x}_j \tag{6}$$

y el gaussiano (K_{RBF}):

$$K_{RBF}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right). \tag{7}$$

Al comienzo de este apartado se indicó que una SVM es, en principio, un clasificador binario. Existen algunas aproximaciones al problema multiclase que reformulan las expresiones anteriores para considerar todas las clases a la vez [9, 10]. Su elevado coste computacional ha llevado, no obstante, a emplear combinaciones de SVMs binarias para abordar el caso multiclase. Existen básicamente dos versiones. La primera, denominada *1-contra-el resto*, consiste en comparar cada clase con todas las demás, mientras que en la segunda versión cada clase se compara con las restantes de forma separada (*1-contra-1*). Aunque en este último caso el número de SVMs binarias es mayor ($\frac{k(k-1)}{2}$ frente a k SVMs, siendo k el número de clases), en [2] se adopta esta solución dado que el menor número de vectores de entrenamiento en cada SVM conduce a un menor coste computacional con prestaciones similares [11]. A continuación se aplica un proceso de votación entre todas las SVMs binarias para decidir la etiqueta correspondiente a la muestra de entrada.

2.2. Estimación de probabilidades en SVMs

Dependiendo de la solución multiclase que se adopte, las SVMs binarias pueden comparar dos clases entre sí o

bien una de ellas con todas las demás. No obstante, por lo general el proceso de obtención de probabilidades a posteriori a partir de las salidas de las SVMs consta en ambos casos de dos pasos: 1) obtener la probabilidad de que la muestra pertenezca a cada clase en todas las SVMs binarias, y 2) transformar estas probabilidades *binarias* a probabilidades *multiclase*. En la aproximación *1-contra-el resto*, este último paso puede reducirse a una simple normalización para que la suma de las probabilidades a posteriori sea uno, ya que el número de SVMs binarias coincide con el número de clases.

La forma más comúnmente empleada para transformar la salida de una SVM en probabilidades *binarias* consiste en el uso de una función sigmoide. Su origen está en [12], donde se propone ajustar mediante sendas gaussianas las funciones de densidad de probabilidad condicional de la salida de la SVM ($p(f|y=+1)$ y $p(f|y=-1)$. Aplicando la regla de Bayes:

$$P\left(y=1|f\right)=\frac{p\left(f|y=1\right)P\left(y=1\right)}{\sum\limits_{i=\pm1}p\left(f|y=i\right)P\left(y=i\right)} \qquad (8)$$

y sustituyendo dichas expresiones para las funciones de densidad de probabilidad condicional se llega a:

$$P\left(y=1|f\right)=\frac{1}{1+exp\left(af^2+bf+c\right)}. \qquad (9)$$

Para simplificar esta función y evitar que no sea monótona, en [12] se asume que las gaussianas mencionadas están centradas en los márgenes (±1) y tienen la misma varianza, la cual debe estimarse. En este caso, la expresión de la probabilidad a posteriori se simplifica en una sigmoide, cuya pendiente en la zona lineal se calcula a partir de la varianza de las gaussianas. El sesgo se calcula de forma que $P\left(y=1|f\right)=0{,}5$ en $f=0$.

Basándose en este trabajo y asumiendo que en la zona comprendida entre los márgenes las funciones de densidad de probabilidad condicional son aproximadamente exponenciales, Platt propone un modelo paramétrico para la probabilidad *binaria* a posteriori [5]:

$$P\left(y=1|f\right)=\frac{1}{1+\exp\left(Af+B\right)}. \qquad (10)$$

La diferencia respecto al trabajo anterior consiste en que los parámetros A y B se estiman de manera discriminativa maximizando la verosimilitud.

Esta expresión proporciona directamente probabilidades *multiclase* en el caso *1-contra-el resto*. Si se tienen k clases distintas, la probabilidad a posteriori de la clase *i-ésima* se puede obtener como:

$$P\left(y=i|\mathbf{x}\right)=\frac{1}{1+exp\left(A_if_i\left(\mathbf{x}\right)+B_i\right)}, \qquad (11)$$

siendo $f_i\left(\mathbf{x}\right)$ la salida de la SVM binaria que clasifica la clase i contra el resto. En este caso no se garantiza que la suma de las probabilidades sea uno; una forma de obtener probabilidades a posteriori normalizadas consiste en usar la función *softmax*:

$$P\left(y=i|\mathbf{x}\right)=\frac{exp\left(\gamma f_i\left(\mathbf{x}\right)\right)}{\sum\limits_{j=1}^{k}exp\left(\gamma f_j\left(\mathbf{x}\right)\right)}, \qquad (12)$$

donde el parámetro γ se estima maximizando la verosimilitud.

En el caso *1-contra-1*, en primer lugar se debe calcular la probabilidad de Platt para la muestra de entrada en cada SVM binaria $\left(i,j\right),\forall i,j\in\{1,...,k\}$. La probabilidad de Platt de que **x** pertenezca a la clase i en la SVM binaria (i,j) se calcula como:

$$r_{ij}=P\left(y=i|\text{y=i ó j},\mathbf{x}\right)=\frac{1}{1+exp\left(A_{ij}f_{ij}(\mathbf{x})+B_{ij}\right)},$$

$$r_{ji}=P\left(y=j|\text{y=i ó j},\mathbf{x}\right)=1-r_{ij}(\mathbf{x}), \qquad (13)$$

siendo $f_{ij}\left(\mathbf{x}\right)$ la salida de la SVM binaria (i,j).

A continuación, se emplea una modificación del método de Refregier-Vallet para transformar estas probabilidades *binarias* $r_{ij}\ \forall\ i,j$ en probabilidades *multiclase* $P(y=i|\mathbf{x})\ \forall\ i$. En [13] se propone resolver un sistema lineal formado por $k-1$ ecuaciones del tipo:

$$r_{ji}P\left(y=i|\mathbf{x}\right)=r_{ij}P\left(y=j|\mathbf{x}\right), \qquad (14)$$

junto con otra que fuerce que la suma de las probabilidades sea uno. En [3] se señala que la solución que se obtiene depende en gran medida de las ecuaciones seleccionadas, por lo que se propone como alternativa el siguiente problema de minimización, que considera todas las ecuaciones posibles:

$$\min_{\mathbf{P}}\quad\frac{1}{2}\sum_{i=1}^{k}\sum_{j:j\neq i}^{k}\left(r_{ji}P\left(y=i|\mathbf{x}\right)-r_{ij}P\left(y=j|\mathbf{x}\right)\right)^2,$$

$$\text{sujeto a}\quad\sum_{i=1}^{k}P\left(y=i|\mathbf{x}\right)=1,$$

$$P\left(y=i|\mathbf{x}\right)\geq0\ \forall i, \qquad (15)$$

con $\mathbf{P}(\mathbf{x})=\left[P\left(y=1|\mathbf{x}\right),...,P\left(y=k|\mathbf{x}\right)\right]$.

Finalmente, en [14] se propone una forma alternativa para obtener las probabilidades a posteriori *multiclase* cuando la probabilidad a priori es la misma para todas las clases:

$$P(y=i|\mathbf{x})=\frac{\prod\limits_{j:j\neq i}r_{ij}}{\sum\limits_{m=1}^{k}\prod\limits_{j:j\neq m}r_{mj}} \qquad (16)$$

3. NUEVA ESTIMACIÓN DE PROBABILIDADES NO PARAMÉTRICAS

Tanto el trabajo presentado en [12] como el método de Platt asumen un comportamiento exponencial, ya

sea globalmente o de forma local entre ambos márgenes, en las funciones de densidad de probabilidad condicional de la salida de las SVMs binarias ($p(f|y=+1)$ y $p(f|y=-1)$). Sin embargo, normalmente no se cumple esta condición, por lo que la justificación teórica del uso de una sigmoide para obtener las probabilidades *binarias* no está suficientemente fundamentada. Otro problema, de carácter práctico, es que los parámetros de la sigmoide se estiman maximizando la verosimilitud de las muestras de entrenamiento. El procedimiento habitual para encontrar los mejores parámetros consiste en un proceso de validación cruzada en cada SVM binaria, por lo que el tiempo de entrenamiento de la SVM multiclase se multiplica aproximadamente por el número de grupos empleados en la validación. Véase que el gran tamaño de las bases de datos empleadas en habla y la elevada superposición de las clases consideradas (que produce máquinas con un gran número de vectores soporte) hacen del entrenamiento de la SVM multiclase un proceso costoso computacionalmente. El método de Platt produce un aumento considerable de este coste, a la vez que carece de una justificación teórica sólida. No obstante, hay que señalar que la función sigmoide resulta una manera intuitiva y directa de transformar la salida de la SVM, definida $\forall\ \Re$, en un valor asimilable a una probabilidad $\in [0, ..., 1]$.

Respecto al método de Refregier-Vallet modificado empleado en [3] para obtener las probabilidades *multiclase*, hay que señalar que su formulación produce que la probabilidad a posteriori de una clase dependa de la de todas las demás y, a través de ellas, de todas las probabilidades *binarias*. Es decir, en el cálculo de la probabilidad $P(y=i|\mathbf{x})$ intervienen las probabilidades $r_{jm}(\mathbf{x})$, $\forall\ j,m \in \{0, ..., k\}$. Esto significa que en el cálculo de la probabilidad a posteriori de una clase se están teniendo en cuenta decisiones de máquinas binarias en las que no participa, lo que a priori no resulta lógico.

Por todas estas razones, se han estudiado diversos métodos alternativos centrados en torno a tres objetivos principales: a) obviar en la medida de lo posible la hipótesis de gaussianidad en las funciones de densidad de probabilidad condicional de la salida de las SVMs; b) emplear únicamente la salida de las SVMs binarias en las que participa la clase considerada; c) reducir el coste computacional que introduce el proceso de estimación de probabilidades, tanto en el entrenamiento como en el test.

En reconocimiento de habla, el algoritmo de Viterbi suele trabajar con la suma de los logaritmos decimales de las probabilidades y verosimilitudes para evitar problemas numéricos. Así, la log-verosimilitud que proporcionan los Modelos Ocultos de Markov podría sustituirse por cualquier otra medida indicativa de la pertenencia de la muestra a cada uno de los modelos. Una medida de este tipo puede ser la distancia a la frontera de decisión (margen): a medida que crece podemos estar razonablemente más seguros de que la muestra pertenece a la clase considerada. Esta idea es la que se usa en las dos primeras alternativas que proponemos: emplear el margen en el Viterbi como medida directa de la pertenencia de la muestra a cada clase o modelo. El problema que aparece cuando usamos la versión multiclase *1-contra-1* es que no disponemos de un solo margen por clase, sino de los $k-1$ correspondientes a las SVMs binarias en las que está presenta la clase. La solución que proponemos es calcular un margen *multiclase* o *acumulado* como la suma de los márgenes en las SVMs binarias en las que participa la clase (método *Directo 1* $[P_{D1}]$).

$$P_{D1}(y=i|\mathbf{x}) = \sum_{j:j\neq i} f_{ij}(\mathbf{x}). \tag{17}$$

Cuando el número de clases es elevado, el valor absoluto de la suma de los márgenes puede dispararse, por lo que también se considera el uso del valor medio del margen (método *Directo 2* $[P_{D2}]$).

$$P_{D2}(y=i|\mathbf{x}) = \frac{P_{D1}(y=i|\mathbf{x})}{k}. \tag{18}$$

Como se muestra en el apartado 4.2, estos dos métodos no proporcionan buenos resultados, quizás debido a que el Viterbi empleado está optimizado para su uso habitual en reconocimiento de habla. Si la escala de los logaritmos de las verosimilitudes difiere en gran medida de la de los márgenes *acumulados*, podría suceder que se desajustase la ponderación entre el modelado acústico y el modelado del lenguaje, produciendo peores resultados. Para obtener valores asimilables a probabilidades en el rango $[0, ..., 1]$ se decidió usar la función *softmax* sobre las sumas de los márgenes (método *softmax* $[P_S]$):

$$P_S(y=i|\mathbf{x}) = \frac{exp(\gamma P_{D1}(y=i|\mathbf{x}))}{\sum_{j=1}^{k} exp(\gamma P_{D1}(y=j|\mathbf{x}))}, \tag{19}$$

siendo γ una constante normalizadora. Se puede observar que esta solución es la versión *1-contra-1* de la expresión (12).

La última de nuestras propuestas consiste en calcular las probabilidades de Platt para cada SVM y aplicar a continuación el método de Refregier-Vallet modificado, pero usando en este caso la misma función sigmoide en todas las SVMs binarias (método de *Platt modificado* $[r_{ij_M}]$):

$$r_{ij_M}(\mathbf{x}) = \frac{1}{1+exp(Af_{ij}(\mathbf{x}))}. \tag{20}$$

Esta expresión es similar a la (13), pero ahora el parámetro A es una constante negativa que no depende la SVM binaria y $B=0$. Este valor para el sesgo tiene sentido, ya que cualquier otro implica un desplazamiento de la frontera de decisión de la SVM binaria. Tanto en este método como en el anterior se emplea el logaritmo de la probabilidad estimada, mientras que en los métodos *directos* se usa la suma de los márgenes.

4. EXPERIMENTOS

4.1. Descripción del sistema

El sistema de reconocimiento de habla continua presentado en [2] se basa en el uso de una SVM multiclase que clasifica cada segmento de voz; como se ha explicado, se trata de un clasificador blando, ya que en realidad proporciona una estimación de la probabilidad a posteriori de cada clase. A continuación, un Viterbi se encarga de obtener una secuencia de palabras a partir de las decisiones acústicas de la SVM.

Por simplicidad, se aborda una tarea de reconocimiento de dígitos conectados en castellano; no obstante, el sistema es fácilmente extensible al reconocimiento de habla continua. Como unidades básicas se consideran tres clases por fonema, correspondientes a la parte inicial, la parte estable y la parte final del mismo. Se ha comprobado que esta forma de definir las clases (en vez de una sola por fonema) mejora las prestaciones del sistema, ya que se tiene en cuenta en mayor medida la variabilidad temporal del habla. Para los dígitos en castellano se definen 17 fonemas más el silencio, por lo que en total se tendrán $18 \cdot 3 = 54$ clases.

El software empleado para el entrenamiento de las SVMs, tanto en el sistema de referencia como en los experimentos que presentamos en el apartado 4.2, es *LIBSVM* [15]. Este programa implementa la solución multiclase *1-contra-1*, lo que en este caso implica entrenar y usar $\frac{54 \cdot 53}{2} = 1431$ SVMs binarias. Se usa un *kernel* gaussiano RBF, con parámetros C y σ obtenidos mediante validación cruzada. La estimación de las probabilidades a posteriori se basa en el cálculo de las probabilidades de Platt y el método de Refregier-Vallet modificado. Para la implementación del algortimo de Viterbi se ha empleado el software HTK [16].

La base de datos usada es SpeechDat [17]. Dispone de 4000 locutores, de los cuales se han empleado 3496 para entrenamiento (con un total de 71000 ficheros) y 350 para test. Para limitar la duración de los experimentos, en la fase de test se han utilizado únicamente 700 de los ficheros con dígitos conectados, disponiéndose de un total de 6546 palabras.

La parametrización de la voz consiste en 12 coeficientes MFCCs y la energía, y sus correspondientes diferencias primeras y segundas. En total, se obtienen 39 parámetros por muestra, con un periodo de muestreo de 10 ms. En experimentos anteriores se ha comprobado que la normalización de los datos es fundamental cuando se trabaja con Máquinas de Vectores Soporte, por lo que se ha realizado una normalización en media y varianza.

4.2. Resultados

La tabla 1 muestra los resultados obtenidos para cada uno de los modelos que se han estudiado: cálculo de las probabilidades de Platt seguido del método de Refregier-Vallet modificado, métodos directo 1 y 2, función *softmax* sobre los márgenes acumulados, y método de Platt modificado, para distintos valores de la pendiente de la sigmoide en la zona lineal. Se indican los casos en los que se utiliza el logaritmo decimal de la probabilidad estimada. Para el método *softmax* se han probado diversos valores de γ, obteniéndose el mejor resultado para un valor de $\gamma = 0{,}02$.

Método	Tasa de reconocimiento de palabras (%)
Platt+R-V (log)	$\mathbf{95{,}05}$ $[94{,}52 - 95{,}58]$
Directo 1	$90{,}65$ $[89{,}94 - 91{,}36]$
Directo 2	$91{,}60$ $[90{,}93 - 92{,}27]$
Softmax ($\gamma = 0{,}02$) (log)	$91{,}63$ $[90{,}96 - 92{,}30]$
Platt modificado+R-V ($A = -0{,}5$) (log)	$93{,}84$ $[93{,}25 - 94{,}43]$
Platt modificado+R-V ($A = -1{,}0$) (log)	$\mathbf{95{,}04}$ $[94{,}51 - 95{,}57]$
Platt modificado+R-V ($A = -1{,}5$) (log)	$94{,}88$ $[94{,}35 - 95{,}41]$
Platt modificado+R-V ($A = -2{,}0$) (log)	$94{,}74$ $[94{,}20 - 95{,}28]$

Tabla 1.
Tasas de reconocimiento de palabras (%) para los diversos métodos estudiados.

Las tasas de reconocimiento de palabras se presentan junto con sus correspondientes intervalos de confianza al $95\,\%$. Para calcular dichos intervalos se ha usado la siguiente expresión ([18], páginas 407-408):

$$\frac{\Delta}{2} = 1{,}96\sqrt{\frac{p\,(100-p)}{n}}. \qquad (21)$$

El parámetro p denota el porcentaje de palabras reconocidas correctamente en el experimento y n es el número de palabras presentes en el conjunto de test (6546). De esta forma, las tasas de reconocimiento que se presentan en la tabla se encuentran en el intervalo $\left[p - \frac{\Delta}{2}, p + \frac{\Delta}{2}\right]$ con una confianza del $95\,\%$.

Como se puede ver, todos los métodos que se basan en el cálculo del margen *acumulado* a partir de las salidas de las SVMs binarias obtienen peores resultados que aquellos que emplean la probabilidad de Platt. La razón más probable de este comportamiento se señaló en la sección 3: el cambio en la escala de los valores que aporta la etapa de modelado acústico en el Viterbi puede estar produciendo un desajuste en la ponderación del modelado del lenguaje que influya en los resultados finales.

Por otra parte, se observa que la modificación del método de Platt que se plantea en este artículo, consistente en usar una única sigmoide para transformar la salida de todas las SVMs binarias en las correspondientes probabilidades *binarias*, no produce diferencias significativas en la tasa de reconocimiento de palabras frente al método de Platt presentado en [5], para un rango de

$A \in [-2, ..., -1]$. Esto significa que no se obtiene ninguna ventaja real al estimar los parámetros A y B de forma individual en cada una de las 1431 SVMs binarias y, por tanto, este proceso puede eliminarse de la etapa de entrenamiento de la SVM multiclase, con lo que se consigue una importante reducción del tiempo de entrenamiento.

Finalmente, a tenor de los resultados presentados en la tabla 1, el método de Refregier-Vallet modificado [3] se presenta como una forma sensata de transformar las probabilidades *binarias* $r_{ij}(\mathbf{x})$ en probabilidades *multiclase* $P(y = i|\mathbf{x})$.

5. CONCLUSIONES

Ciertas aplicaciones requieren conocer el nivel de incertidumbre asociado a la decisión de un clasificador o la probabilidad a posteriori de cada una de las clases posibles. En el caso de las Máquinas de Vectores Soporte, la obtención de dichas probabilidades a partir de sus salidas no es obvia. La forma más habitual consiste en calcular la probabilidad de Platt y, en el caso multiclase, utilizar una modificación del método de Refregier-Vallet para obtener dichas probabilidades a posteriori. No obstante, ambos métodos presentan una serie de limitaciones tanto en su base teórica como en su aplicación práctica.

En este artículo se presentan diversas alternativas a estos dos métodos, en el marco del reconocimiento de habla continua mediante SVMs. Se ha comprobado experimentalmente que el método de Platt se puede simplificar, utilizando una única función sigmoide para obtener las probabilidades *binarias* a partir de las salidas de las SVMs binarias. Así mismo, se ha mostrado que los resultados obtenidos son poco sensibles a la pendiente de la sigmoide en la zona lineal, en un intervalo $[-2, -1]$. La consecuencia práctica es que el tiempo empleado en el entrenamiento de la máquina se reduce de una forma muy importante (entre 4 y 5 veces), factor de gran importancia dado el tamaño de las máquinas y de las bases de datos de voz consideradas.

6. BIBLIOGRAFÍA

[1] B. Schölkopf y A. Smola, *Learning with kernels*, MIT Press, Cambridge, MA, 2002.

[2] J. Padrell-Sendra, D. Martín-Iglesias, y F. Díaz de María, "Support Vector Machines for Continuous Speech Recognition," *Proc. EUSIPCO, Florence, Italy*, 2006.

[3] T. F. Wu, C. J. Lin, y R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *Proc. NIPS, Vancouver, Canada*, 2003.

[4] J. Milgram, M. Cheriet, y R. Sabourin, "Estimating accurate multi-class probabilities with support vector machines," *IEEE International Joint Conference on Neural Networks*, vol. 3, pp. 1906–1911, 2005.

[5] J. C. Platt, *Probabilities for SV Machines, in Advances in Large Margin Classifiers*, MIT Press, Cambridge, MA, 1999.

[6] V. N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, New York, 1998.

[7] F. Pérez-Cruz y O. Bousquet, "Kernel Methods and Their Potential Use in Signal Processing," *Signal Processing Magazine*, vol. 21, no. 3, pp. 57–65, 2004.

[8] J. Nocedal y S. J. Wright, *Numerical Optimization*, Springer, New York, USA, 1999.

[9] K. Crammer y Y. Singer, "On the algorithmic implementation of multiclass kernelbased vector machines," *Journal of Machine Learning Research*, vol. 2, no. 5, pp. 265–292, 2001.

[10] J. Weston y C. Watkins, "Multi-Class Support Vector Machines," Tech. Rep., Department of Computer Science, Royal Holloway, University of London, Egham, UK, 1998.

[11] C. W. Hsu y C. J. Lin, "A Comparison of Methods for Multi-class Support Vector Machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.

[12] T. Hastie y T. Tibshirani, "Classification by pairwise coupling," *Proc. NIPS, Denver, USA*, 1997.

[13] P. Refregier y F. Vallet, "Probabilistic approach for multiclass classification with neural netwoks," *Proc. International Conference on Artificial Netwoks*, pp. 1003–1007, 1991.

[14] T. Hamamura, H. Mizutani, y B. Irie, "A multiclass classification method based on multiple pairwise classifiers," *Proc. 7th International Conference on Document Analysis and Recognition, Edinburgh, UK*, vol. 2, pp. 809–813, 2003.

[15] C. C. Chang y C. J. Lin, "LIBSVM: A Library for Support Vector Machines," *Software disponible en http://www.csie.ntu.edu.tw/cjlin/∼libsvm*, 2001.

[16] S. Young et al., "HTK-Hidden Markov Model Toolkit (versión 3.2)," *Software disponible en http://htk.eng.cam.ac.uk/*, 2002.

[17] A. Moreno, "SpeechDat Spanish Database for Fixed Telephone Network," Tech. Rep., Technical University of Catalonia, 1997.

[18] N. A. Weiss y M. J. Hasset, *Introductory Statistics*, Addison-Wesley, Reading, MA, 3rd edition, 1991.