Modificación del sesgo de una SVM entrenada sobre clases no balanceadas*

Haydemar Núñez[1], Cecilio Angulo[2], Luis Gonzalez-Abril[3]
[1] Laboratorio de Inteligencia Artificial, Universidad Central de Venezuela. 1020-A Caracas, Venezuela
[2] CETpD, Universitat Politècnica de Catalunya. 08800 Vilanova i la Geltrú, España
[3] Departamento de Economía Aplicada I, Universidad de Sevilla. 41018 Sevilla, España

Abstract

En el área de aprendizaje automático, uno de los problemas que se presenta es el relacionado con las clases no balanceadas. Esto ocurre cuando en el conjunto de datos se dispone de muchos ejemplos de una clase, pero muy pocos de otra. La principal contribución de este trabajo es la definición de un sesgo modificado, con la SVM entrenada original, de forma que se mejora la generalización sobre conjuntos no balanceados medida en forma de media geométrica. Una ventaja importante de nuestra propuesta es que el problema de optimización para hallar la SVM no cambia para el sesgo elegido y, por tanto, el coste computacional es casi nulo. Los resultados de experimentación confirman que la propuesta iguala en prestaciones aquellas con mayor rendimiento en la literatura, mientras que no añade coste computacional.

1. Introducción

Uno de los problemas que enfrentan los algoritmos de aprendizaje es el relacionado con las clases no balanceadas. Esto ocurre cuando en el conjunto de datos se dispone de muchos ejemplos de una clase, pero muy pocos de otra, lo cual dificulta la generación de un buen modelo de clasificación con técnicas tradicionales [He and Garcia, 2009]. Ejemplos de algunos dominios donde se presenta esta situación son el diagnóstico médico, la clasificación de textos, la detección de fraude en el uso de tarjetas de crédito, o la detección de intrusos en redes de comunicación, entre otros. En estos problemas resulta crítico el error de generalización sufrido en la clase minoritaria ya que, en general, resulta ser la clase de interés. En el caso de las Máquinas de Soporte Vectorial (SVM) [Vapnik, 1998], su mecanismo de aprendizaje las convierte en una opción interesante para tratar con conjuntos de datos no balanceados, ya que la SVM sólo toma en cuenta un subconjunto de las instancias de entrenamiento para construir un modelo de clasificación. Sin embargo, al igual que otros algoritmos de aprendizaje automático, para construir estos modelos la SVM busca minimizar el error total sobre el conjunto de datos, por lo que están inherentemente sesgadas hacia el concepto mayoritario; si el desequilibrio es severo, la SVM tenderá a clasificar todos los ejemplos presentados como pertenecientes a la clase mayoritaria. Este trabajo se centra en el estudio de las SVM en ambientes de aprendizaje con conjuntos no balanceados y, en particular, en las técnicas empleadas para mejorar su rendimiento mediante la determinación de un nuevo sesgo o umbral para el modelo de clasificación inducido. En la próxima sección se detalla una taxonomía de las estrategias que se han propuesto para encarar el desbalance en conjuntos de datos, así como las métricas que resultan más adecuadas para evaluar clasificadores en estos escenarios. Luego, en la Sección 2, se presenta el mecanismo de aprendizaje de las SVM y métodos que se han propuesto en la literatura para mejorar su rendimiento en este tipo de problemas. En la siguiente sección se describen en detalle las estrategias de post-procesamiento y se presenta una propuesta para la determinación de un nuevo sesgo para la SVM. Los experimentos y resultados se presentan en la Sección 5. Por último, se ofrecen conclusiones y futuros trabajos.

2. Aprendizaje con conjuntos de datos no balanceados

En problemas de aprendizaje binarios con conjunto de datos no balanceados, la clase con un menor número de ejemplos o instancias representativas se conoce como la clase minoritaria o positiva, mientras que la asociada al resto de los datos se refiere a la clase mayoritaria o negativa. En general, el desbalance entre las clases puede presentarse por varias razones relacionadas con:

- La naturaleza del problema, donde el desbalance es el resultado directo de las características de la población que genera los datos. Esta situación se presenta por ejemplo, en el diagnóstico de enfermedades raras, donde la clase minoritaria es muy limitada.
- Costo y/o dificultad en la obtención de datos de la clase de interés. Por ejemplo, en la clasificación de la morfología espermática, puede resultar que los casos clasifica-

*Haydemar Núñez agradece la financiación a través del proyecto PG-03-7678-2009/1, CDCH, UCV. Cecilio Angulo agradece la financiación a través de una beca I3 del Programa General de Intensificación de la Investigación, concedida por la Universitat Politècnica de Catalunya. Luis Gonzalez-Abril agradece la financiación del Gobierno de España a través del proyecto ARTEMISA, TIN2009-14378-C02-01.

dos como normales y disponible en un conjunto de datos sea relativamente pequeño. La obtención de más datos se ve limitado ya que, en general, provienen de sujetos que acuden a centros especializados por problemas de infertilidad, por lo que es probable la existencia de defectos de morfología.

- Utilización de clasificadores binarios, como la Máquina de Soporte Vectorial (SVM). En este caso el aprendizaje con datos no balanceados es inevitable cuando se intenta resolver problemas de clasificación multiclase, donde generalmente se adopta la estrategia de 1-versus-n y se entrena un clasificador por cada clase (datos positivos); el resto de los datos formarían el conjunto de ejemplos de la otra clase (datos negativos).

En estos escenarios, el aprendizaje con algoritmos tradicionales es muy limitado, debido a que estos métodos, en general, están diseñados para inducir un modelo de clasificación basado en el error que se comete sobre todo el conjunto de entrenamiento. Se busca entonces generalizar a partir de toda la muestra y producir la hipótesis más simple que mejor se ajuste a los datos, basado en la minimización de este error. Con conjuntos de datos no balanceados, la hipótesis más simple frecuentemente es la que clasifica casi todos los ejemplos como negativos. Para solventar este problema se han propuesto diferentes estrategias [He and Garcia, 2009; Sun *et al.*, 2009], las cuales pueden agruparse siguiendo la siguiente taxonomía de clasificación:

- Métodos de muestreo, que generan artificialmente datos para reequilibrar el conjunto de entrenamiento, ya sea mediante el sobre-muestreo (*over-sampling*) de la clase minoritaria, o con el sub-muestreo (*under-sampling*) de la clase mayoritaria.
- Aprendizaje sensitivo al costo, que considera los costos asociados a los ejemplos mal clasificados. Si el conjunto de datos no está balanceado, se asigna un mayor peso a los errores sobre la clase minoritaria y se desarrolla una hipótesis que minimice el costo total sobre el conjunto de entrenamiento.
- Técnicas que combinan múltiples clasificadores (*ensembles*), los cuales son entrenados a partir de conjuntos de aprendizaje que consideran diferentes distribuciones de datos.
- Métodos de post-procesamiento, que redefinen la función de clasificación aprendida, con el fin de mejorar el rendimiento sobre la clase minoritaria.
- Modificación de algoritmos tradicionales y propuestas de nuevos algoritmos.

El error de clasificación y la exactitud predictiva (*accuracy*) no son métricas de rendimiento apropiadas cuando las probabilidades a priori de las clases son muy diferentes, debido a que no consideran costos en las clasificaciones incorrectas y son muy sensitivas al sesgo entre las clases [He and Garcia, 2009]. Por ejemplo, en una situación donde la proporción de datos es 5/95, es decir 5 % de los datos pertenecen a la clase positiva y 95 % a la clase negativa, un clasificador que sólo predice correctamente la clase mayoritaria tendría un rendimiento de clasificación (exactitud) del 95 %. Sin embargo, el rendimiento sobre la clase de interés (la minoritaria) sería nulo. Esta descripción no refleja que se consigue un 0 % de identificación de los ejemplos de la clase positiva, por lo que no suministra una información adecuada para verificar el rendimiento de un modelo de clasificación con conjuntos no balanceados.

Debido a que estas métricas dependen de la distribución de los datos, en problemas de aprendizaje no balanceado se adoptan otras medidas de evaluación basadas en la información suministrada por la matriz de confusión (Cuadro 1).

Predicción/Real	Positiva	Negativa
Positiva	VP	FP
Negativa	FN	VN

Cuadro 1: Matriz de confusión

De esta tabla se deriva la definición de la *precisión* y el *recall*,

$$\text{Precision} = \frac{VP}{VP + FP}\,,\ \text{Recall} = \frac{VP}{VP + FN} \qquad (1)$$

La precisión es una medida de la sensibilidad del clasificador a la clase minoritaria, que determina, de los ejemplos clasificados como positivos, cuántos son clasificados correctamente. El recall es una medida de la completitud, que indica cuántos ejemplos de la clase positiva fueron clasificados correctamente. La precisión y el recall, a diferencia de la exactitud o accuray,

$$\text{Accuracy} = \frac{VP + VN}{VP + FP + VN + FN} \qquad (2)$$

no son sensitivas a los cambios en la distribución de los datos y pueden efectivamente evaluar el rendimiento de clasificación en escenarios de aprendizaje no balanceados.

A partir de estas dos métricas se define el *Valor-F*, que mide la efectividad de la clasificación en términos de la importancia ponderada sobre el recall y la precisión, determinada por un coeficiente definido por el usuario, β,

$$\text{Valor-F} = \frac{(1 + \beta) \cdot \text{Recall} \cdot \text{Precision}}{\beta^2 \cdot \text{Recall} + \text{Precision}} \qquad (3)$$

Otra medida que se utiliza comúnmente con conjuntos no balanceados es la media geométrica (*g-media*), que evalúa el rendimiento en términos de la exactitud positiva y la exactitud negativa.

$$\text{g-media} = \sqrt{\frac{VP}{VP + FN} \cdot \frac{VN}{VN + FP}} \qquad (4)$$

Finalmente, otras técnicas de evaluación que son utilizadas son el análisis de la curva ROC (por 'Receiver Operating Characteristics') y el análisis basado en las curvas de precisión y recall.

En el presente estudio se usarán la g-media y la precisión como medidas de evaluación de los clasificadores analizados.

3. SVM sobre conjuntos no balanceados

La Máquina de Soporte Vectorial (SVM) es una técnica de aprendizaje fundamentada en la teoría del aprendizaje estadístico [Vapnik, 1998; Hebrich, 2002], que ha sido aplicada con éxito en problemas de clasificación y regresión en diferentes dominios. El espacio de hipótesis de estas máquinas de aprendizaje son hiperplanos (superficies de decisión lineal) y durante el entrenamiento se busca aquel con un margen máximo de separación entre las clases. Para una tarea de clasificación binaria con un conjunto de datos de entrenamiento $(\mathbf{x}_i, y_i)_{i=1,\ldots,N}$, con $\mathbf{x}_i \in \mathcal{R}^m$, $y_i \in \{-1, +1\}$, y función de decisión del formato $f(\mathbf{x}) = \text{signo}(\mathbf{w}\cdot\mathbf{x}+b)$, este hiperplano óptimo se determina de la siguiente forma,

$$\text{mín}\ \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N}\xi_i \tag{5}$$

sujeto a

$$y_i \cdot (\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 - \xi_i\ ,\ \xi_i \geq 0\ \forall i \tag{6}$$

donde $\mathbf{w}$ es el vector perpendicular al hiperplano, que define su orientación, y b determina su posición. Las variables ficticias (ξ_i) miden el error sobre las instancias que violan la restricción $y_i \cdot (\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1$. El parámetro C, definido por el usuario, determina el balance o *tradeoff* entre maximizar el margen y minimizar el error; a mayor valor de C, la SVM se centra más en minimizar los errores; cuanto más pequeño, el objetivo principal será maximizar el margen.

En su forma dual, este problema de optimización puede resolverse como,

$$\text{máx}\sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i,k=1}^{N}\alpha_i y_i \alpha_k y_k \langle \mathbf{x}_i \cdot \mathbf{x}_k\rangle \tag{7}$$

sujeto a

$$0 \leq \alpha_i \leq C\ \forall i\ ,\ \sum_{i=1}^{N}\alpha_i y_i = 0 \tag{8}$$

dando lugar a la siguiente función de decisión,

$$f(\mathbf{x}) = \text{signo}\left(\sum_{i=1}^{sv}\alpha_i y_i \langle \mathbf{x}_i \cdot \mathbf{x}\rangle + b\right) \tag{9}$$

Para construir límites de decisión no lineales, se proyectan los vectores de entrada en un espacio de mayor dimensión dotado con producto interno, llamado espacio de características $\mathcal{F}$, utilizando un conjunto base de funciones no lineales $\phi(\cdot)$. En este nuevo espacio se determina el hiperplano óptimo, que corresponderá a una función de decisión no lineal cuya forma estará determinada por estas funciones. Mediante el uso de la teoría de núcleos (*kernels*) que cumplan con el teorema de Mercer, no es necesario conocer el nuevo espacio de características ya que todas las operaciones se pueden realizar directamente en el espacio de entrada utilizando $\phi(\mathbf{x}_i)\cdot\phi(\mathbf{x}_j) = K(\mathbf{x}_i, \mathbf{x}_j)$. La función de decisión se formula entonces en términos de estos núcleos,

$$f(\mathbf{x}) = \text{signo}\left(\sum_{i=1}^{sv}\alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b\right) \tag{10}$$

Entre todos los vectores de entrenamiento sólo unos pocos tienen asociado un peso $\alpha_i > 0$ en la función de decisión (9) o (10). Estos elementos yacen en el margen de decisión y son conocidos como *vectores de soporte* (SV). De esta forma, la hipótesis generada por la SVM se representa sobre la base de aquellos puntos, del espacio de entrada, más cercanos al límite de decisión. El valor de la función de decisión sin signo $f(\mathbf{x})$ es una medida de la distancia de un ejemplo $\mathbf{x}$ al hiperplano, mientras que el signo determina su etiqueta de clase (positiva o negativa).

Para conjuntos moderadamente desbalanceados, resultados empíricos muestran que, a diferencia de otras máquinas de aprendizaje, la SVM puede producir una buena hipótesis sin ninguna modificación [Akbani *et al.*, 2004; Imam *et al.*, 2006]. Sin embargo, no escapan al problema del desbalance entre las clases cuando el sesgo en la distribución de los datos es significativo.

También se ha mostrado empíricamente que el hiperplano aprendido por la SVM en presencia de conjuntos no balanceados tiene aproximadamente la misma orientación que el hiperplano ideal [Wu and Chang, 2005]. El problema de su pobre generalización sobre datos muy desbalanceados está pues relacionado con el sesgo b, ya que las instancias positivas yacen lejos de este límite ideal; es decir, la SVM aprende un límite que está muy cercano a la clase minoritaria. Como se indica en (5) la SVM trata de maximizar el margen entre ejemplos de clases opuestas con una penalización asociada a los errores. Para conjuntos muy desbalanceados, el penal para un pequeño número de ejemplos positivos es sobrepasado por el introducido por un gran número de ejemplos negativos. Entonces, el problema de minimización se orienta más en maximizar el margen a partir de los ejemplos de la clase mayoritaria, resultando en un hiperplano de separación más sesgado hacia la clase minoritaria. Este sesgo conduce a una solución con un rendimiento de generalización muy bajo o nulo para los ejemplos de esta clase.

Las estrategias que se han propuesto en la literatura, en el caso de las SVM, para tratar conjuntos de datos no balanceados pueden ser clasificadas de acuerdo al momento en que son aplicadas durante el proceso de aprendizaje,

- *Estrategias de pre-procesamiento*: Se aplican técnicas de sobre-muestreo [Bae *et al.*, 2010; Nguyen *et al.*, 2011], sub-muestreo [Li *et al.*, 2008c; Vivaracho, 2006; Yu *et al.*, 2006] o una combinación de éstas [Castro *et al.*, 2009; Vilarino *et al.*, 2005]. Algunos trabajos aplican métodos de muestreo con combinación de clasificadores o ensembles [Kang and Cho, 2006; Liu *et al.*, 2006; Waske *et al.*, 2009; Yu *et al.*, 2006]. En este grupo también se incluyen la aplicación de técnicas de selección de características y de ponderación de variables.

- *Estrategias de entrenamiento*: Se asignan diferentes costos a través del parámetro C [Cohen *et al.*, 2006; Yang *et al.*, 2008]. También se incluyen métodos para modificar la matriz kernel de acuerdo al desbalance observada en la distribución de los datos [Wu and Chang, 2005]. Otros esquemas combinan estrategias basadas en costo con técnicas de muestreo [Akbani *et al.*, 2004; Tang *et al.*, 2009] y con combinación de clasificadores

[Wang and Japkowicz, 2008].

- *Estrategias de post-procesamiento*: Se ajusta el límite de decisión aprendido por la SVM de tal forma que suministre un buen margen de separación para la clase positiva [Imam *et al.*, 2006; Li *et al.*, 2008b; Shanahan and Roma, 2003]. También se incluyen las técnicas que incorporan un módulo de post-procesamiento para darle una interpretación diferente a las salidas de la SVM, de tipo probabilístico [Wang and Zheng, 2008] o fuzzy [Li *et al.*, 2008a].

4. Estrategias de post-procesamiento

En general, las propuestas recogidas en la literatura pretenden la modificación del vector de pesos $\mathbf{w}$ en la función de decisión, o bien, la determinación de un nuevo sesgo o umbral. Otras investigaciones proponen incorporar un módulo de post-procesamiento que permita darle otra interpretación a las salidas de la SVM en escenarios no balanceados.

En el caso del ajuste del sesgo, conclusiones extraídas del análisis de trabajos como los presentados en [Sun *et al.*, 2009], en el domino de la clasificación de textos, sugieren investigar en estas estrategias, las cuales, además, no afectan directamente el entrenamiento de la SVM. Algunas de las propuestas con mayor rendimiento son:

- z-SVM [Imam *et al.*, 2006]: se resuelve un segundo problema de optimización para determinar un parámetro z, que pondera la contribución de los vectores de soporte de la clase minoritaria en la función de decisión obtenida luego del entrenamiento,

$$\begin{aligned} f(\mathbf{x}, z) &= z \sum_{\mathbf{vs}_p \in VS_p} \alpha_p y_p K(\mathbf{vs}_p, \mathbf{x}) \\ &+ \sum_{\mathbf{vs}_n \in VS_n} \alpha_n y_n K(\mathbf{vs}_n, \mathbf{x}) \end{aligned} \quad (11)$$

- WHM Offset [Li *et al.*, 2008b]: se calcula un *offset* a partir del promedio de los valores de decisión, S_i, generados por $f(\mathbf{x})$ sin signo, para los vectores de soporte y se construye una nueva función de clasificación de la siguiente forma, $f(\mathbf{x}) = \text{signo}(h(\mathbf{x}))$, con

$$h(\mathbf{x}) = \sum_{i=1}^{VS} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b - \frac{\sum_{i=1}^{n} S_i}{n} \quad (12)$$

- Aplicación del algoritmo Beta-Gamma [Shanahan and Roma, 2003] para el cálculo de un nuevo umbral (θ_{opt}) para la función de decisión $f(\mathbf{x}) = \text{signo}(h(\mathbf{x}))$, con

$$h(\mathbf{x}) = \sum_{i=1}^{VS} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b - \theta_{opt} \quad (13)$$

Nuestra propuesta está basada en los desarrollos presentados en [Gonzalez-Abril *et al.*, 2008] para el cálculo de un nuevo sesgo para la función de decisión de la SVM: Dado el conjunto de entrenamiento $\mathcal{Z} = \{(\mathbf{x}_i, y_i)\}_{i=1,\ldots,N}$, donde $\mathbf{x}_i \in \mathcal{X} \subset \mathcal{R}^m$, $y_i \in \mathcal{Y} = \{-1, +1\}$, sean $\mathcal{Z}_1$ y $\mathcal{Z}_2$ los conjuntos con los patrones pertenecientes a las clases positiva (+) y negativa (-), respectivamente. Entonces, se calculan dos valores α y β de la siguiente forma (Esquema_1),

$$\alpha = \max_{\mathbf{z}_i \in \mathcal{Z}_2} \langle \mathbf{x}_i \cdot \mathbf{w} \rangle \ , \ \beta = \min_{\mathbf{z}_i \in \mathcal{Z}_1} \langle \mathbf{x}_i \cdot \mathbf{w} \rangle \quad (14)$$

donde α es el valor máximo absoluto del hiperplano sin sesgo aplicado al conjunto $\mathcal{Z}_2$ y β es el valor mínimo absoluto del hiperplano sin sesgo aplicado al conjunto $\mathcal{Z}_1$. Luego, se proponen dos alternativas para el cálculo de un nuevo sesgo de las siguientes maneras,

1. Siguiendo la definición tradicional del sesgo,

$$b_s = \frac{\alpha + \beta}{2} \quad (15)$$

2. Si N_1 y N_2 representan el número de patrones de la clase (+) y (-), respectivamente[1], otra opción es

$$b_f = \frac{N_1 \alpha + N_2 \beta}{N_1 + N_2} \quad (16)$$

y se define una nueva función de decisión sustituyendo b, ya sea por b_s o b_f.

Como la hipótesis o función de decisión generada por la SVM se construye utilizando sólo los vectores de soporte (instancias más informativas para la tarea de clasificación), se propone como alternativa calcular los valores de α y β de la siguiente forma (Esquema_2),

$$\alpha = \max_{\mathbf{vs}_i \in VS_2} \langle \mathbf{vs}_i \cdot \mathbf{w} \rangle \ , \ \beta = \min_{\mathbf{vs}_i \in VS_1} \langle \mathbf{vs}_i \cdot \mathbf{w} \rangle \quad (17)$$

donde ahora α será el máximo valor absoluto del hiperplano sin sesgo aplicado al conjunto de vectores de soporte de la clase negativa (VS_2) y β el mínimo valor absoluto del hiperplano sin sesgo aplicado al conjunto de vectores de soporte de la clase positiva (VS_1). Luego, se determinan los nuevos sesgo b_s y b_f como se explicó en (15) y (16) y se construye una nueva función de decisión sustituyendo b por estos nuevos sesgos.

Una última posibilidad a considerar es calcular los valores de α y β como un promedio de los valores de hiperplano sin sesgo aplicado a los vectores de soporte de la clase negativa y positiva, respectivamente (Esquema_3). De esta forma se reduce la influencia de los vectores de soporte que supongan un valor *outlier*. Luego, se determinan los valores de b_s y b_f como se explicó anteriormente.

5. Experimentos

Para comprobar las prestaciones de los diferentes esquemas se utilizaron los conjuntos de datos que se describen en el Cuadro 2 del Repositorio UCI [Blake and Merz, 1998]. Como se hizo en otros artículos y para efectos de comparar el rendimiento de los clasificadores obtenidos, se asignó la clase (+) a la etiqueta que se muestra entre paréntesis y la clase (-) al resto de los datos; por ejemplo, para el conjunto de datos 'Ecoli', se seleccionó a la clase "5" como la clase positiva, para el resto de las clases, la negativa.

Para medir el rendimiento de los clasificadores se utilizó la media geométrica (g-media) como medida principal y la precisión como soporte de comparación con lo que sería el caso no desbalanceado. Como técnica de evaluación, se utilizó la validación cruzada de 10 particiones. Los Cuadros 3, 4 y 5

[1] en un problema con clases no balanceadas se suele presentar que $N_1 << N_2$.

Datos	Total	Positiva	Negativa	d
Abalone (19)	4177	32 (0.77 %)	4145 (99.23 %)	8
Yeast (5)	1484	51 (3.44 %)	1433 (96.56 %)	8
Car (3)	1728	69 (3.99 %)	1659 (96.01 %)	7
Ecoli (5)	336	20 (5.95 %)	316 (94.05 %)	8
Glass (7)	214	29 (13.55 %)	185 (86.45 %)	10
Segment (1)	2310	330 (14.29 %)	1980 (85.71 %)	19

Cuadro 2: Seis conjuntos de datos del repositorio UCI ordenadas en orden decreciente según nivel de desbalance

muestran los valores promedio de la media geométrica (GM) para la SVM original, así como la precisión (P), y para los clasificadores obtenidos al ajustar la posición del hiperplano definido por la SVM utilizando un nuevo sesgo, según los diferentes esquemas (Esquema_1, Esquema_2 y Esquema_3).

GM	b	b_s	b_f
Abalone	$0{,}0 \pm 0{,}0$	$0{,}0 \pm 0{,}0$	$0{,}0 \pm 0{,}0$
Yeast	$0{,}0 \pm 0{,}0$	$0{,}0 \pm 0{,}0$	$0{,}0 \pm 0{,}0$
Car	$93{,}6 \pm 7{,}4$	$68{,}8 \pm 28{,}4$	$\mathbf{96{,}8 \pm 3{,}6}$
Ecoli	$35{,}4 \pm 37{,}3$	$0{,}0 \pm 0{,}0$	$52{,}4 \pm 37{,}3$
Glass	$0{,}0 \pm 0{,}0$	$17{,}3 \pm 27{,}9$	$\mathbf{92{,}9 \pm 8{,}5}$
Segment	$\mathbf{99{,}0 \pm 1{,}5}$	$49{,}5 \pm 9{,}5$	$97{,}7 \pm 2{,}6$
P	b	b_s	b_f
Abalone	$99{,}2 \pm 0{,}1$	$40{,}1 \pm 50{,}8$	$10{,}6 \pm 31{,}3$
Yeast	$96{,}4 \pm 0{,}5$	$86{,}9 \pm 29{,}4$	$77{,}9 \pm 39{,}3$
Car	$99{,}1 \pm 0{,}7$	$98{,}1 \pm 1{,}1$	$98{,}0 \pm 3{,}5$
Ecoli	$95{,}5 \pm 1{,}6$	$94{,}0 \pm 0{,}1$	$96{,}4 \pm 1{,}9$
Glass	$85{,}9 \pm 0{,}5$	$87{,}4 \pm 2{,}4$	$95{,}8 \pm 3{,}5$
Segment	$99{,}7 \pm 0{,}4$	$89{,}7 \pm 2{,}7$	$99{,}4 \pm 0{,}7$

Cuadro 3: Porcentaje promedio de la media geométrica GM y de la precisión P para el clasificador SVM original (b) y los clasificadores según el Esquema_1 (b_s, b_f)

Analizando la g-media, se observa que en cuatro de las seis bases de datos analizadas (indicado en negrita en los Cuadros), el uso del sesgo b_f mejora las prestaciones tanto respecto a la SVM original, como respecto del sesgo convencional b_s definido según [Gonzalez-Abril *et al.*, 2008]. Además, su prestación es mayor cuanto mayor es el desbalance entre clases. De entre los esquemas propuestos, el Esquema_2 es el que indica un mejor comportamiento ante el desbalance elevado entre clases. La SVM original es la mejor opción, por muy poco, en el caso de la base de datos 'Segment', que es la que presenta menor desbalance entre las analizadas y sobre la que se produce un menor error de generalización en el sentido de la media geométrica.

Por otra parte, analizando la medida de precisión se observa el fenómeno que motiva nuestra propuesta de modificación de sesgo. La SVM original consigue una mayor valor de precisión en casi todos los casos que cualquiera de las alternativas propuestas. Ello es así porque el clasificador asociado ha tendido a etiquetar todas las instancias como pertenecientes a la clase negativa, que es la mayoritaria.

En vista de los resultados experimentales obtenidos y con objeto de no tomar ventaja en la comparativa con otros estudios proponiendo varios sesgos, se selecciona como alternativa de modificación del sesgo la que corresponde a b_f del Esquema_2.

GM	b	b_s	b_f
Abalone	$0{,}0 \pm 0{,}0$	$22{,}0 \pm 13{,}2$	$\mathbf{60{,}8 \pm 7{,}6}$
Yeast	$0{,}0 \pm 0{,}0$	$53{,}5 \pm 14{,}9$	$\mathbf{80{,}7 \pm 10{,}2}$
Car	$93{,}6 \pm 7{,}4$	$93{,}6 \pm 1{,}3$	$91{,}5 \pm 2{,}4$
Ecoli	$35{,}4 \pm 37{,}3$	$93{,}1 \pm 8{,}8$	$91{,}4 \pm 7{,}4$
Glass	$0{,}0 \pm 0{,}0$	$89{,}1 \pm 9{,}3$	$87{,}1 \pm 6{,}4$
Segment	$\mathbf{99{,}0 \pm 1{,}5}$	$98{,}6 \pm 1{,}5$	$95{,}7 \pm 2{,}4$
P	b	b_s	b_f
Abalone	$99{,}2 \pm 0{,}1$	$71{,}0 \pm 8{,}1$	$70{,}5 \pm 17{,}7$
Yeast	$96{,}4 \pm 0{,}5$	$37{,}5 \pm 25{,}2$	$87{,}7 \pm 8{,}7$
Car	$99{,}1 \pm 0{,}7$	$88{,}2 \pm 2{,}5$	$84{,}5 \pm 4{,}2$
Ecoli	$95{,}5 \pm 1{,}6$	$92{,}5 \pm 7{,}3$	$85{,}0 \pm 12{,}6$
Glass	$85{,}9 \pm 0{,}5$	$91{,}1 \pm 2{,}6$	$83{,}7 \pm 5{,}9$
Segment	$99{,}7 \pm 0{,}4$	$98{,}4 \pm 2{,}7$	$97{,}4 \pm 1{,}8$

Cuadro 4: Porcentaje promedio de la media geométrica GM y de la precisión P para el clasificador SVM original (b) y los clasificadores según el Esquema_2 (b_s, b_f)

GM	b	b_s	b_f
Abalone	$0{,}0 \pm 0{,}0$	$55{,}6 \pm 10{,}6$	$56{,}1 \pm 13{,}9$
Yeast	$0{,}0 \pm 0{,}0$	$69{,}7 \pm 7{,}1$	$70{,}1 \pm 7{,}9$
Car	$93{,}6 \pm 7{,}4$	$92{,}3 \pm 1{,}3$	$93{,}3 \pm 1{,}7$
Ecoli	$35{,}4 \pm 37{,}3$	$\mathbf{95{,}5 \pm 4{,}0}$	$92{,}9 \pm 8{,}6$
Glass	$0{,}0 \pm 0{,}0$	$86{,}5 \pm 9{,}2$	$88{,}2 \pm 9{,}0$
Segment	$99{,}0 \pm 1{,}5$	$90{,}5 \pm 2{,}3$	$89{,}9 \pm 3{,}8$
P	b	b_s	b_f
Abalone	$99{,}2 \pm 0{,}1$	$46{,}3 \pm 10{,}9$	$48{,}0 \pm 20{,}0$
Yeast	$96{,}4 \pm 0{,}5$	$55{,}4 \pm 11{,}3$	$56{,}6 \pm 14{,}6$
Car	$99{,}1 \pm 0{,}7$	$85{,}8 \pm 2{,}2$	$87{,}6 \pm 3{,}1$
Ecoli	$95{,}5 \pm 1{,}6$	$91{,}9 \pm 7{,}1$	$92{,}2 \pm 6{,}4$
Glass	$85{,}9 \pm 0{,}5$	$86{,}9 \pm 7{,}3$	$89{,}7 \pm 4{,}8$
Segment	$\mathbf{99{,}7 \pm 0{,}4}$	$84{,}6 \pm 3{,}5$	$83{,}6 \pm 5{,}8$

Cuadro 5: Porcentaje promedio de la media geométrica GM y de la precisión P para el clasificador SVM original (b) y los clasificadores según el Esquema_3 (b_s, b_f)

En el Cuadro 6 se muestran los resultados obtenidos para la media geométrica en las pruebas realizadas con este sesgo seleccionado, junto con los reportados en publicaciones relacionadas. Entre paréntesis se ha indicado el porcentaje de acierto en media geométrica de la SVM original reportado en cada trabajo. Se puede observar que, a nivel global, la aproximación KBA es la que obtiene un mejor rendimiento sobre las bases de datos analizadas hasta el momento. Sin embargo, su mejora es sólo significativa en el caso de la base de datos 'Car', respecto a nuestra propuesta. Por contra, en el caso de 'Abalone', la base de datos con mayor desbalanceo, KBA es la propuesta que peor se comporta, de forma significativa respecto a las otras dos.

Debe recordarse la manera como estas dos propuestas intentan tratar el problema de las clases no balanceadas: z-SVM [Imam *et al.*, 2006], resuelve un segundo problema de optimización para determinar un parámetro z, que pondera la contribución de los vectores de soporte de la clase minoritaria

Datos	b_f	z-SVM	KBA
Abalone	61 (0)	62 (0)	58 (0)
Yeast	81 (0)	72 (0)	82 (59)
Car	92 (93,6)	93 (0)	99 (99)
Segment	96 (99)	97 (92)	98 (98)

Cuadro 6: Resultados de porcentaje obtenidos para la media geométrica en las pruebas realizadas con el sesgo seleccionado (Esquema_3, b_f), junto con los reportados en publicaciones relacionadas. Entre paréntesis se ha indicado el porcentaje de acierto en media geométrica de la SVM original reportado en cada trabajo

en la función de decisión obtenida luego del entrenamiento. Por su parte, KBA [Wu and Chang, 2005], modifica la matriz kernel para que incluya información de la distribución de las clases. Por tanto, ambas aproximaciones añaden una complejidad computacional y un tiempo de cálculo importante al problema original, ya de por si caro en costo computacional. En cambio, nuestra propuesta de modificación del sesgo implica un añadido temporal y de complejidad casi nulos.

A continuación, se lista de forma abreviada algunos comentarios que pueden alimentar el análisis de resultados:

- Trabajando con datos no balanceados, parece claro que g-media y precisión son las medidas más acertadas.
- Para algunos conjuntos de datos, a pesar del desbalance, la SVM original logra obtener un modelo razonable (ejemplo: Ecoli); en otros no lo logra (ejemplo: Abalone). En ambos casos, utilizar el nuevo bias propuesto mejora considerablemente los resultados, sobre todo a nivel de Recall. Incluso con un desbalance moderado, se logra mejorar el rendimiento.
- La propuesta presentada iguala, pero no mejora, en prestacionas de generalización aquellas mejores en la literatura. Por contra, consigue igualarlas introduciendo un costo computacional casi nulo.

6. Conclusiones

La principal contribución de este trabajo es que la tasa de precisión sobre conjuntos no balanceados medida en forma de media geométrica y de precisión puede ser mejorada mediante el uso de un sesgo diferente con la SVM entrenada original. Así, una ventaja importante es que el problema de optimización para hallar la SVM no cambia para cada sesgo elegido y, por tanto, el coste computacional es casi nulo. Como futura investigación, se está desarrollando un marco teórico para el estudio de los movimientos del sesgo en el espacio de trabajo en función de su definición. Además, se está trabajando en un número mayor de conjuntos de datos para darle mayor consistencia a la demostración empírica.

Referencias

[Akbani *et al.*, 2004] R. Akbani, S. Kwek, and N. Japkowicz. Applying support vector machines to imbalanced datasets. In *Proc. 15th European Conference on Machine Learning*, pages 39–50, 2004.

[Bae *et al.*, 2010] Min Hyeok Bae, Teresa Wu, and Rong Pan. Mix-ratio sampling: Classifying multiclass imbalanced mouse brain images using support vector machine. *Expert Systems with Applications*, 37(7):4955 – 4965, 2010.

[Blake and Merz, 1998] C.L. Blake and C.J. Merz. UCI repository of machine learning databases, 1998.

[Castro *et al.*, 2009] Cristiano Leite Castro, Mateus Araujo Carvalho, and AntÃ´nio Padua Braga. An improved algorithm for svms classification of imbalanced data sets. In Dominic Palmer-Brown, Chrisina Draganova, Elias Pimenidis, and Haris Mouratidis, editors, *Engineering Applications of Neural Networks*, volume 43 of *Communications in Computer and Information Science*, pages 108–118. Springer Berlin Heidelberg, 2009.

[Cohen *et al.*, 2006] Gilles Cohen, Mélanie Hilario, Hugo Sax, Stéphane Hugonnet, and Antoine Geissbuhler. Learning from imbalanced data in surveillance of nosocomial infection. *Artificial Intelligence in Medicine*, 37:7–18, May 2006.

[Gonzalez-Abril *et al.*, 2008] L. Gonzalez-Abril, C. Angulo, F. Velasco, and J. A. Ortega. A note on the bias in SVMs for multiclassification. *IEEE Transactions on Neural Networks*, 19(4):723–725, 2008.

[He and Garcia, 2009] H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, September 2009.

[Hebrich, 2002] R. Hebrich. *Learning Kernel Classifiers. Theory and Algorithms*. The MIT Press, 2002.

[Imam *et al.*, 2006] T. Imam, K. Ting, and J. Kamruzzaman. z-SVM: An SVM for improved classification of imbalanced data. *AI 2006: Advances in Artificial Intelligence*, (4304):264–273, 2006.

[Kang and Cho, 2006] Pilsung Kang and Sungzoon Cho. EUS SVMs: Ensemble of under-sampled svms for data imbalance problems. In Irwin King, Jun Wang, Lai-Wan Chan, and DeLiang Wang, editors, *Neural Information Processing*, volume 4232 of *Lecture Notes in Computer Science*, pages 837–846. Springer Berlin / Heidelberg, 2006.

[Li *et al.*, 2008a] Boyang Li, Jinglu Hu, and Kotaro Hirasawa. An improved support vector machine with soft decision-making boundary. In *Proceedings of the 26th IASTED International Conference on Artificial Intelligence and Applications*, AIA '08, pages 40–45, Anaheim, CA, USA, 2008. ACTA Press.

[Li *et al.*, 2008b] Boyang Li, Jinglu Hu, and Kotaro Hirasawa. Support vector machine classifier with whm offset for unbalanced data. *Journal of Advanced Computational Intelligence and Intelligence Informatics*, 12(1):94–101, 2008.

[Li *et al.*, 2008c] Peng Li, Pei-Li Qiao, and Yuan-Chao Liu. A hybrid re-sampling method for svm learning from imbalanced data sets. In *Proceedings of the 2008 Fifth International Conference on Fuzzy Systems and Knowledge Dis-*

covery - Volume 02, pages 65–69, Washington, DC, USA, 2008. IEEE Computer Society.

[Liu *et al.*, 2006] Yang Liu, Aijun An, and Xiangji Huang. Boosting prediction accuracy on imbalanced datasets with svm ensembles. In Wee-Keong Ng, Masaru Kitsuregawa, Jianzhong Li, and Kuiyu Chang, editors, *Advances in Knowledge Discovery and Data Mining*, volume 3918 of *Lecture Notes in Computer Science*, pages 107–118. Springer Berlin / Heidelberg, 2006.

[Nguyen *et al.*, 2011] Hien M. Nguyen, Eric W. Cooper, and Katsuari Kamei. Borderline over-sampling for imbalanced data classification. *International Journal of Knowledge Engineering and Soft Data Paradigms*, 3:4–21, April 2011.

[Shanahan and Roma, 2003] James Shanahan and Norbert Roma. Improving svm text classification performance through threshold adjustment. In Nada Lavrac, Dragan Gamberger, Hendrik Blockeel, and Ljupco Todorovski, editors, *Machine Learning: ECML 2003*, volume 2837 of *Lecture Notes in Computer Science*, pages 361–372. Springer Berlin / Heidelberg, 2003.

[Sun *et al.*, 2009] A. X. Sun, E. P. Lim, and Y. Liu. On strategies for imbalanced text classification using SVM: A comparative study. *Decision Support Systems*, 48:191–201, 2009.

[Tang *et al.*, 2009] Yuchun Tang, Yan-Qing Zhang, Nitesh V Chawla, and Sven Krasser. Svms modeling for highly imbalanced classification. *IEEE transactions on systems man and cybernetics Part B Cybernetics a publication of the IEEE Systems Man and Cybernetics Society*, 39(1):281–288, 2009.

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

[Vilarino *et al.*, 2005] Fernando Vilarino, Panagiota Spyridonos, Jordi Vitrià, and Petia Radeva. Experiments with svm and stratified sampling with an imbalanced problem: Detection of intestinal contractions. In Sameer Singh, Maneesha Singh, Chid Apte, and Petra Perner, editors, *Pattern Recognition and Image Analysis*, volume 3687 of *Lecture Notes in Computer Science*, pages 783–791. Springer Berlin / Heidelberg, 2005.

[Vivaracho, 2006] Carlos Vivaracho. Improving svm training by means of ntil when the data sets are imbalanced. In Floriana Esposito, Zbigniew Ras, Donato Malerba, and Giovanni Semeraro, editors, *Foundations of Intelligent Systems*, volume 4203 of *Lecture Notes in Computer Science*, pages 111–120. Springer Berlin / Heidelberg, 2006.

[Wang and Japkowicz, 2008] Benjamin Wang and Nathalie Japkowicz. Boosting support vector machines for imbalanced data sets. In Aijun An, Stan Matwin, Zbigniew Ras, and Dominik Slezak, editors, *Foundations of Intelligent Systems*, volume 4994 of *Lecture Notes in Computer Science*, pages 38–47. Springer Berlin / Heidelberg, 2008.

[Wang and Zheng, 2008] Haiying Wang and Huiru Zheng. An improved support vector machine for the classification of imbalanced biological datasets. In De-Shuang Huang, Donald Wunsch, Daniel Levine, and Kang-Hyun Jo, editors, *Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues*, volume 5226 of *Lecture Notes in Computer Science*, pages 63–70. Springer Berlin / Heidelberg, 2008.

[Waske *et al.*, 2009] Björn Waske, Jon Benediktsson, and Johannes Sveinsson. Classifying remote sensing data with support vector machines and imbalanced training data. In Jón Benediktsson, Josef Kittler, and Fabio Roli, editors, *Multiple Classifier Systems*, volume 5519 of *Lecture Notes in Computer Science*, pages 375–384. Springer Berlin / Heidelberg, 2009.

[Wu and Chang, 2005] G. Wu and E. Y. Chang. KBA: Kernel Boundary Alignment Considering Imbalanced Data Distribution. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):786–795, 2005.

[Yang *et al.*, 2008] Chan-Yun Yang, Jianjun Wang, Jr-Syu Yang, and Guo-Ding Yu. Imbalanced svm learning with margin compensation. In Fuchun Sun, Jianwei Zhang, Ying Tan, Jinde Cao, and Wen Yu, editors, *Advances in Neural Networks - ISNN 2008*, volume 5263 of *Lecture Notes in Computer Science*, pages 636–644. Springer Berlin / Heidelberg, 2008.

[Yu *et al.*, 2006] Ting Yu, John Debenham, Tony Jan, and Simeon Simoff. Combine vector quantization and support vector machine for imbalanced datasets. In Max Bramer, editor, *Artificial Intelligence in Theory and Practice*, volume 217 of *IFIP International Federation for Information Processing*, pages 81–88. Springer Boston, 2006.