Universidad Central de Venezuela
Facultad de Ciencias
Escuela de Computación

Lecturas en Ciencias de la Computación
ISSN 1316-6239

Máquinas de Soporte Vectorial sobre Conjuntos de Datos no Balanceados: Propuesta de un Nuevo Sesgo

Haydemar Núñez, Luis Gonzalez-Abril, Cecilio Angulo

RT 2012-06

Centro de Ingeniería de Software y Sistemas
ISYS-UCV
Caracas, Septiembre 2012.

Máquinas de Soporte Vectorial sobre conjuntos de datos no balanceados: propuesta de un nuevo sesgo

Haydemar Núñez[1], Luis Gonzalez-Abril[2], Cecilio Angulo[3]

[1]Laboratorio de Inteligencia Artificial, Centro de Ingeniería de Software y Sistemas.
Facultad de Ciencias. Universidad Central de Venezuela
[2]Applied Economics I Dept, Universidad de Sevilla. 41018 Sevilla, España
[3]CETpD, Universitat Politècnica de Catalunya. 08800 Vilanova i la Geltrú, España

Resumen.- En el aprendizaje con conjuntos de datos no balanceados, la máquina de soporte vectorial (SVM) puede exhibir un bajo rendimiento sobre la clase minoritaria ya que, como otras máquinas de aprendizaje, están diseñadas para inducir un modelo de clasificación basado en un error global. Con el fin de mejorar su desempeño en este tipo de problemas, en este trabajo se propone una estrategia de post-procesamiento basada en el cálculo de un nuevo sesgo o umbral que toma en cuenta la proporción de las clases en el conjunto de datos y que permite ajustar la función aprendida por la SVM para mejorar su desempeño sobre la clase minoritaria. Esta solución no supone la entonación de nuevos parámetros ni la modificación del problema de optimización estándar para entrenar la SVM. Los resultados obtenidos de la experimentación sobre 23 conjuntos de datos con diferentes grados de desbalance, muestran que efectivamente se logra mejorar las clasificaciones sobre la clase minoritaria, medidas en función de g-media y la sensibilidad.

1. Introducción

Un problema al que se enfrentan los algoritmos de aprendizaje en entornos de clasificación es el relacionado con el desbalance en el conjunto de datos [1]-[4]. Esto ocurre cuando se dispone de muchos ejemplos de una clase, pero muy pocos de otra. Ejemplos de algunos dominios donde se presenta esta situación son: el diagnóstico médico, la clasificación de textos, detección de fraude en el uso de tarjetas de crédito, detección de intrusos en redes de comunicación, entre otros.

En estos escenarios es muy importante obtener modelos que exhiban un alto rendimiento de predicción sobre la clase minoritaria ya que ésta, por lo general, representa el objetivo o *target* de la tarea de clasificación. Sin embargo, los algoritmos de aprendizaje tradicionales tenderán a producir una hipótesis que sólo tendrá un buen desempeño sobre la clase mayoritaria. Esto es debido a que están diseñados para inducir un modelo de clasificación basado en el error que se comete sobre todo el conjunto de entrenamiento, sin tomar en cuenta la representatividad o balance de las clases

En el caso de la Máquina de Soporte Vectorial (SVM) [5], [6], su mecanismo de aprendizaje las convierte en una opción interesante para tratar con conjuntos de datos no balanceados, debido a que la SVM sólo toma en cuenta un subconjunto de las instancias de entrenamiento para la construcción de un modelo de clasificación. Pero, al igual que otras máquinas de aprendizaje, para generar estos modelos la SVM busca minimizar el error total sobre el conjunto de datos, por lo que está inherentemente sesgada hacia el concepto mayoritario cuando el desbalance es severo.

Para mejorar el rendimiento de la SVM sobre problemas con clases no balanceadas se han propuesto diversas soluciones, algunas de aplicación general, como las técnicas de muestreo para re-balancear el conjunto de datos en una etapa de pre-procesamiento; otras, más específicas, que toman en cuenta sus características particulares y que se basan, por ejemplo, en el aprendizaje sensitivo al costo [1]. Algunas investigaciones también sugieren utilizar una etapa de post-procesamiento con el fin de reducir el sesgo, hacia la clase mayoritaria, de la función de clasificación aprendida por la SVM [7].

Siguiendo esta última línea de investigación, en este trabajo se propone una estrategia para Maquinas de Soporte Vectorial basada en el cálculo de un nuevo sesgo o umbral (*bias*), que considera la proporción de las clases en el conjunto de datos y que permite ajustar la función aprendida por la SVM para mejorar su desempeño sobre la clase minoritaria. La solución que se propone no supone la entonación de nuevos parámetros (como ocurre en otros esquemas), ni la modificación del problema de optimización estándar para su entrenamiento.

Este artículo se encuentra estructurado de la siguiente manera: en la próxima Sección, se describe el problema de aprendizaje con conjuntos de datos no balanceados y se presentan las métricas que resultan más adecuadas para evaluar clasificadores en estos escenarios. En la Sección 3 se presenta el mecanismo de aprendizaje de la SVM y se da una revisión de algunas estrategias dirigidas a mejorar su rendimiento en este tipo de problemas. La Sección 4 presenta en detalle la propuesta de post-procesamiento basado en la determinación de un nuevo sesgo. Luego, en la Sección 5 se describen los experimentos realizados para verificar su aplicabilidad, con un análisis de los resultados. Por último, se dan las conclusiones y trabajos futuros.

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 (minoritaria) se conoce como la clase positiva, mientras que la asociada al resto de los datos (mayoritaria) se refiere a la clase negativa. En general, el desbalance entre las clases en un conjunto de datos puede presentarse por diversas razones, algunas de ellas 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 positiva (datos de pacientes con este tipo de enfermedad) es muy limitada.
- El costo y/o dificultad en la obtención de datos de la clase de interés. Por ejemplo, en la clasificación de la morfología espermática, puede presentarse que el número de casos clasificados como normales (clase positiva) en un conjunto de datos sea relativamente pequeño. La obtención de más datos de esta clase se verá limitado si, en general, éstos provienen de sujetos que acuden a centros especializados por problemas de infertilidad, por lo que es más probable la existencia de defectos de morfología (clase negativa).

En estos escenarios, el aprendizaje con algoritmos tradicionales es muy limitado debido a que en general, están diseñados para inducir un modelo de clasificación basado en el error que se comete sobre todo el conjunto de entrenamiento, sin tomar en cuenta la representatividad o balance de las clases. 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 todas las instancias como negativas. Por ejemplo, en una situación donde la proporción de datos es de 2/98, es decir 2% de los datos pertenecen a la clase positiva y 98% a la clase negativa, el modelo más simple es el que asigna cualquier instancia a la clase mayoritaria, con un rendimiento de clasificación del 98% de exactitud. Sin embargo, el rendimiento sobre la clase de interés (la minoritaria) será nulo.

Para solventar este problema es pertinente entonces disponer de mecanismos que, acoplados o integrados a estos algoritmos, permitan construir modelos que exhiban un alto grado de exactitud en las clasificaciones de los datos de la clase minoritaria. En este sentido, se han propuesto diversas estrategias como el re-balanceo del conjunto de datos con técnicas de muestreo, la construcción de clasificadores que tomen en cuenta el costo de los errores sobre las

diferentes clases, combinación (*ensemble*) de los resultados de varios clasificadores entrenados con diferentes distribuciones de datos, entre otros [1], [3], [4].

2.1. Métricas de evaluación

En el aprendizaje supervisado, las métricas que se utilizan comúnmente para evaluar el rendimiento de generalización de los modelos inducidos, son el error de clasificación y la exactitud predictiva. En base a la matriz de confusión (Figura 1), estas medidas se definen de la siguiente manera:

	Clase real	
Clase predicha	**Positiva**	**Negativa**
Positiva	**VP** ***(Verdadero positivos)***	**FP** ***(Falsos positivos)***
Negativa	**FN** ***(Falsos negativos)***	**VN** ***(Verdaderos negativos)***

Figura 1. Matriz de confusión para un problema de clasificación binario

$$Exactitud = \frac{VP + VN}{VP + FP + VN + FN} \qquad Error = \frac{FP + FN}{VP + FP + VN + FN} \tag{1}$$

Sin embargo, estas métricas no son apropiadas cuando las probabilidades a priori de las clases son muy diferentes, ya que no consideran los costos en las clasificaciones incorrectas y son muy sensitivas al sesgo entre las clases [1], [2].

Debido a que estas medidas dependen de la distribución de los datos, en problemas de aprendizaje no balanceado se adoptan otras métricas de evaluación que permitan medir el rendimiento sobre cada una de las clases de manera independiente. Algunas de éstas son la precisión y la sensibilidad (o *recall*), las cuales se definen a partir de la matriz de confusión de la siguiente manera:

$$Precisión = \frac{VP}{VP + FP} \qquad Sensibilidad = \frac{VP}{VP + FN} \tag{2}$$

La precisión (o pureza) es una medida de la exactitud que determina, de los ejemplos clasificados como positivos, cuántos son clasificados correctamente. La sensibilidad o *recall*, es una medida de la completitud o exactitud positiva, que indica cuántos ejemplos de esta clase fueron clasificados correctamente. A partir de estas dos métricas se definen otras medidas de evaluación, como el valor F:

$$Valor - F = \frac{(1+\beta) \times Sensibilidad \times Precisión}{\beta^2 \times Sensibilidad + Precisión} \tag{3}$$

Donde β es definido por el usuario y generalmente tiene un valor de 1. Otra medida que se utiliza en escenarios no balanceados es la media geométrica (g-media), la cual evalúa el rendimiento en términos de la sensibilidad y la especificidad (exactitud negativa) de la siguiente forma:

$$g-media=\sqrt{\frac{VP}{VP+FN}\times\frac{VN}{VN+FP}}=\sqrt{Sensibilidad\times Especificidad} \quad (4)$$

Por su parte, el análisis de la curva ROC (*Receiver Operating Characteristics*) es una técnica de evaluación que es utilizada comúnmente y que además, constituye una herramienta visual de comparación entre diferentes clasificadores. La curva ROC muestra gráficamente las relaciones entre la sensibilidad (*eje Y*) y la proporción de FP (*eje X*); ésta última calculada como el número de falsos positivos entre el número total de instancias negativas (*VN+FP*). Así, es posible representar el rendimiento global de un clasificador por un punto en esta gráfica. Por ejemplo, el punto (0,0) representaría a un modelo que clasifica a todos las instancias como negativas, y el punto (0,1) uno que clasifica bien a todos los datos.

Además, para evaluar el rendimiento de diferentes clasificadores se puede utilizar el área total bajo la curva ROC (AUC), la cual puede ser estimada de la siguiente manera [8]:

$$AUC=\frac{1+\left(\frac{VP}{VP+FN}\right)-\left(\frac{FP}{VN+FP}\right)}{2} \quad (5)$$

Esta medida varía entre 0 y 1, donde a mayor valor mejor rendimiento de clasificación.

3. Máquinas de soporte vectorial sobre conjuntos de datos no balanceados

La Máquina de Soporte Vectorial está fundamentada en la teoría del aprendizaje estadístico y ha sido aplicada con éxito en problemas de clasificación y regresión en diferentes dominios [5], [6]. 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 $(\mathbf{x}_i,y_i)_{i=1..N}$, con $\mathbf{x}_i \in \Re^m$, $y_i=\{-1,+1\}$, y la función de decisión del tipo $f(\boldsymbol{x})=signo(\boldsymbol{w}\cdot\boldsymbol{x}+b)$, este hiperplano óptimo se determina de la siguiente forma:

$$\text{Minimizar} \quad \frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{N}\xi_i \quad (6)$$

$$\text{Sujeto a} \quad y_i\left(\boldsymbol{w}\cdot\boldsymbol{x}_i+b\right)\geq 1-\xi_i \qquad \forall i,$$

$$\xi_i\geq 0 \quad \forall i$$

donde **w** es el vector perpendicular al hiperplano, el cual define su orientación, y *b* determina su posición. Las variables ficticias (ξ_i) miden el error sobre las instancias que violan la restricción $y_i(\mathbf{w}\,\mathbf{x}_i + b) \geq 1$. El parámetro *C* (definido por el usuario) determina el balance o *tradeoff* entre maximizar el margen ($1/\|\boldsymbol{w}\|$) y minimizar el error; es decir, mientras más alto el valor de *C*, la

SVM se enfoca más en minimizar los errores; mientas más pequeño, el objetivo principal será maximizar el margen.

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

$$\text{Maximizar} \qquad W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,k=1}^{N} \alpha_i y_i \alpha_k y_k \langle \boldsymbol{x}_i \cdot \boldsymbol{x}_k \rangle \tag{7}$$

$$\text{Sujeto a} \qquad 0 \le \alpha_i \le C \;\; \forall i, \;\; \sum_{i=1}^{N} \alpha_i y_i = 0$$

Dando lugar a la siguiente función de decisión

$$f(\boldsymbol{x}) = signo\left(\sum_{i=1}^{N} \alpha_i y_i \langle \boldsymbol{x}_i \cdot \boldsymbol{x} \rangle + b \right) \tag{8}$$

Para construir límites de decisión no lineales, se proyectan los vectores de entrada en un espacio de producto interno de más alta dimensión, llamado espacio de características *F*, utilizando un conjunto base de funciones no lineales ϕ. En este nuevo espacio se determina el hiperplano óptimo, el cual corresponderá (en el espacio original) 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 $\phi(\boldsymbol{x}_i) \cdot \phi(\boldsymbol{x}_j) = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$. La función de decisión se formula entonces en términos de estos núcleos:

$$f(\boldsymbol{x}) = signo\left(\sum_{i=1}^{N} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b \right) \tag{9}$$

Entre todos los vectores de entrenamiento sólo unos pocos tienen asociado un peso α_i mayor que cero en las ecuaciones (8) y (9). Estos elementos yacen en el margen de decisión y son conocidos como *vectores de soporte* (SV). El valor de $f(\boldsymbol{x})$ sin signo, es una medida de la distancia de un ejemplo **x** al hiperplano, mientras que el signo determina su etiqueta de clase (positiva o negativa).

Para conjuntos de datos moderadamente desbalanceados resultados empíricos muestran que, a diferencia de otras máquinas de aprendizaje, la SVM puede producir una buena hipótesis sin ninguna modificación [9], [10]. Una explicación a este fenómeno es que la SVM sólo utiliza a los vectores de soporte para construir un modelo de clasificación, por lo cual instancias negativas lejos del límite de decisión no serán tomadas en cuenta y la SVM no se verá afectada por ellas, aunque sean numerosas.

Sin embargo, la SVM no escapa al problema del desbalance entre las clases cuando el sesgo en la distribución de los datos es significativo. En estos casos, se ha observado que el hiperplano de separación aprendido por la SVM está muy cercano a la clase minoritaria [11]-[13], lo que

resulta en un rendimiento de generalización muy bajo o nulo para los ejemplos de esta clase, en comparación con aquellos de la clase mayoritaria.

3.1. Estrategias para SVM sobre conjuntos de datos no balanceados

Se han propuesto diversas estrategias para mejorar el rendimiento de la SVM sobre conjuntos de datos no balanceados. Algunas se describen a continuación, presentándolas de acuerdo al momento en pueden ser aplicadas durante el proceso de aprendizaje:

a) Estrategias de pre-procesamiento

Se basan principalmente en técnicas de muestreo para re-balancear el conjunto de datos, con el fin de construir un nuevo conjunto de aprendizaje donde la clase minoritaria esté mejor representada. Una forma de hacerlo es mediante el sobre-muestreo (*over-sampling*) de los datos de esta clase, donde se crean nuevas instancias con el fin de aumentar su proporción en el conjunto de datos. En contraposición, el sub-muestreo (*under-sampling*) busca disminuir el tamaño de la clase mayoritaria mediante la eliminación de un subconjunto de estos datos.

Como el objetivo de estos algoritmos es balancear el conjunto de datos antes del entrenamiento, en general no están dirigidos a una máquina de aprendizaje en particular. Uno muy utilizado es SMOTE [14], el cual se basa en la técnica de K-vecinos más cercanos para sobre-muestrear la clase minoritaria; un ejemplo de su aplicación se muestra en [15]. También es posible utilizar otras estrategias para balancear el conjunto de datos, por ejemplo, mediante la aplicación de algoritmos de agrupación para sub-muestrear la clase mayoritaria [16], [17].

Sin embargo, tomando en cuenta que la función aprendida por la SVM se construye a partir de los vectores de soporte, algunas propuestas buscan aumentar la representatividad de la clase minoritaria sólo en las áreas bordes entre las dos clases. La información del límite puede ser obtenida aplicando heurísticas basadas en la técnica K-vecinos, como se propone en [18]. Otros trabajos se basan en la utilización de una SVM para obtener los vectores de soporte positivos y sobre-muestrear a partir de estos datos [12], [19]. Esta característica también ha sido explotada para construir algoritmos de sub-muestreo, por ejemplo, como el que se presenta en [20], donde se utiliza una SVM para construir un nuevo conjunto de datos constituido sólo por vectores de soporte negativos y los datos positivos.

Asimismo, se han propuesto soluciones que utilizan métodos de muestreo con combinación de clasificadores o *ensembles*. Los *ensembles* se basan en la construcción de un conjunto de modelos cuyas decisiones son combinadas para producir un resultado final. En este caso, se entrenan varias SVMs con diferentes conjuntos de datos que han sido balanceados, producto de aplicar sub-muestreo y/o sobre-muestreo [21]-[24].

b) Estrategias de entrenamiento:

En este grupo se incluyen aquellas propuestas que modifican el problema de optimización estándar para el entrenamiento de la SVM, con el fin de incorporar la información relacionada con las proporciones de las clases en el conjunto de datos. Una manera de hacerlo es mediante el aprendizaje sensitivo al costo el cual, para determinar el modelo de clasificación, incorpora en el problema de aprendizaje la información relacionada al costo o *penal* asociado a las predicciones incorrectas para cada clase. En escenarios no balanceados esta estrategia se utiliza de manera similar, pero se asigna un mayor peso a los errores sobre la clase minoritaria.

Para el caso de las SVM, la información de los costos en los dos tipos de errores puede ser introducida en la formulación del problema de aprendizaje en (6), utilizando dos parámetros de regularización de la siguiente forma [25], [26]:

$$\text{Minimizar} \qquad \frac{1}{2}\|\boldsymbol{w}\|^2 + C^+ \sum_{i/y_i=+1}^{N} \xi_i + C^- \sum_{i/y_i=-1}^{N} \xi_i \tag{10}$$

$$\text{Sujeto a} \qquad y_i\left(\boldsymbol{w}\cdot\boldsymbol{x}_i + b\right) \geq 1 - \xi_i \qquad \forall i,$$

$$\xi_i \geq 0 \quad \forall i$$

donde C^+ *y* C^- son los costos asociados a los errores sobre la clase positiva y negativa, respectivamente. Algunos trabajos también añaden nuevas restricciones sobre las variables ficticias ξ, con el fin de tratar de controlar el margen de separación entre las dos clases [11], [27]. Un enfoque diferente se presentan en [28], donde se utiliza un solo parámetro de regularización *C*, pero la información del costo de los errores se incorpora asignando un peso diferente a cada variable ξ; el resultado es que a cada dato se le asocia un costo de acuerdo a su importancia en la tarea de clasificación.

Otros soluciones proponen combinar el aprendizaje sensitivo al costo con otras técnicas. Por ejemplo en [9], se utiliza el algoritmo SMOTE para sobre-muestrear la clase minoritaria antes de entrenar una SVM con diferentes costos. También se ha utilizado esta estrategia en la construcción de *ensembles* de clasificadores [29]. Por último, se incluyen en este renglón propuestas como la que se describe en [13], donde se presenta el algoritmo KBA, que modifica la matriz kernel de acuerdo al desbalance observado en la distribución de los datos.

c) Estrategias de post-procesamiento:

En general, los trabajos están dirigidos hacia la modificación del vector de pesos **w** de la función de decisión o en la determinación de un nuevo sesgo o umbral, con el fin de ajustar el límite de decisión aprendido por la SVM de tal forma que suministre un buen margen de separación para la clase positiva.

Por ejemplo, en [10] se propone el método z-SVM, el cual determina el valor de un nuevo parámetro z (resolviendo un problema de optimización), que pondera la contribución de los vectores de soporte de la clase minoritaria en el vector **w** de la función de decisión obtenida luego del entrenamiento, de la siguiente forma:

$$f(x,z) = z \sum_{i/y_i=+1}^{N} \alpha_i y_i K(\boldsymbol{x_i}, \boldsymbol{x}) + \sum_{i/y_i=-1}^{N} \alpha_i y_i K(\boldsymbol{x_i}, \boldsymbol{x}) \tag{11}$$

Por su parte, en [30] se buscan modificar el umbral de la función de decisión calculando un *Offset*, a partir del promedio de los valores generados por *f(**x**)* sin signo para los vectores de soporte (S_i), y construyen una nueva función de clasificación como:

$$f(\boldsymbol{x}) = signo\left(\sum_{i=1}^{N} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b - \frac{\sum_{i=1}^{n} S_i}{n}\right) \tag{12}$$

Además, con el fin de disminuir la contribución de los vectores de soporte afectados por ruido, en este trabajo se propone calcular otro *Offset* basado en la media harmónica ponderada de los valores S_i.

Una estrategia similar es utilizada por [31], pero para el cálculo del *Offset* o nuevo umbral (θ_{opt}), aplican el algoritmo *Beta-Gamma* y generan una nueva función de decisión ***f(x)*** de la siguiente manera:

$$f(\boldsymbol{x}) = signo\left(\sum_{i=1}^{N} \alpha_i y_i \langle \boldsymbol{x}_i \cdot \boldsymbol{x} \rangle + b - \theta_{opt}\right) \tag{13}$$

Otras investigaciones proponen sustituir el esquema de clasificación estándar (basada en la función signo), por otro que permita darle otra interpretación a las salidas de la SVM en escenarios no balanceados. Por ejemplo, en [32], se utiliza una función de decisión difusa cuyos parámetros son estimados a partir de la distribución observada en el conjunto de datos. En [33], se incorpora al proceso de decisión un módulo de post-procesamiento, cuya construcción se basa en métodos de la teoría de la información para construir un nuevo umbral de clasificación.

4. Propuesta de post-procesamiento basada en un nuevo bias

Las estrategias que se han propuesto para mejorar el rendimiento de la SVM sobre conjuntos de datos no balanceados en general requieren de la entonación de nuevos parámetros, como la tasa de muestreo, el número k de vecinos, etc. Otros pueden ser computacionalmente costosos al considerar la construcción de varios clasificadores (como los métodos basados en *ensembles*), o basarse en algoritmos iterativos, como los que modifican la matriz kernel (KBA) y algunas técnicas de muestreo, que requieren de varios pasos de re-entrenamiento. En los enfoques sensitivos al costo es necesario modificar el problema de optimización estándar de la SVM y conocer los costos de los errores sobre las clases. Además, pueden producir modelos sobre-ajustados.

Por otra parte, 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 [11], [13], [22]. La baja generalización sobre la clase minoritaria está asociada al sesgo *b*, ya que la SVM aprende un límite que está muy cercano a esta clase. Otros trabajos, como los presentados en [7], en el domino de la clasificación de textos, sugieren investigar en las estrategias dirigidas a determinar un nuevo umbral para la función de decisión de la SVM basado en la distribución de las clases en el conjunto de datos, las cuales, además, no afectarían directamente el entrenamiento de la SVM.

Siguiendo esta última línea de investigación, en este trabajo se propone una estrategia de post-procesamiento basada en el cálculo de un nuevo sesgo o umbral que considera la proporción de las clases en el conjunto de datos y que permite ajustar la función aprendida por la SVM para mejorar su desempeño sobre la clase minoritaria. La solución que se propone no supone la entonación de nuevos parámetros. Tampoco requiere modificar el problema de optimización estándar para entrenar la SVM ni pasos adicionales de re-entrenamiento.

Nuestra propuesta está basada en los desarrollos presentados en [34] y con ella se busca modificar, después del entrenamiento, el margen de separación del hiperplano hacia la clase mayoritaria, con el fin de conseguir un mejor rendimiento de generalización sobre los datos de la clase minoritaria. Para ello se propone calcular un nuevo sesgo de la siguiente forma:

Dado el conjunto de aprendizaje $Z = \{(\mathbf{x}_i, y_i)\}_{i=1..N}$, donde $\mathbf{x}_i \in \Re^m$, $y_i \in Y = \{+1,-1\}$. Sean Z_1 y Z_2 los conjuntos con los datos pertenecientes a las clases positiva (+) y negativa (-), respectivamente. La formulación estándar del sesgo para el caso linealmente separable indica que *b*, en la ecuación (8), podría obtenerse como:

$$b_s = -\frac{\alpha + \beta}{2} \tag{14}$$

donde α es el valor máximo del hiperplano sin sesgo aplicado al conjunto Z_2 y β es el valor mínimo del hiperplano sin sesgo aplicado al conjunto Z_1; es decir,

$$\alpha = \max_{x_i \in Z_2} \langle \boldsymbol{x}_i . \boldsymbol{w} \rangle \qquad\qquad \beta = \min_{x_i \in Z_1} \langle \boldsymbol{x}_i . \boldsymbol{w} \rangle \tag{15}$$

A partir de la expresión para b_s, se ha considerado utilizar la misma para el caso no separable, pero tomando en cuenta la proporción de las clases en el conjunto de datos. Entonces, sea N_1 y N_2 el número de patrones de las clases (+) y (-) respectivamente, un nuevo sesgo b_f puede definirse de la siguiente forma,

$$b_f = -\frac{N_1 \alpha + N_2 \beta}{N_1 + N_2} \tag{16}$$

En este caso, si $N_1 << N_2$, lo cual es lo común en problemas no balanceados, este nuevo sesgo acercará el límite de decisión hacia la clase negativa, aumentando de esta forma el margen de separación para la clase positiva. Además, como los valores máximos y mínimos del hiperplano sin sesgo se alcanzan en los vectores de soporte, sólo es necesario considerar estos puntos para el cálculo de α y β. La nueva función de decisión quedaría entonces expresada de la siguiente manera,

$$f_{new}(\boldsymbol{x}) = signo\left(\sum_{i=1}^{N} \alpha_i y_i \langle \boldsymbol{x}_i \cdot \boldsymbol{x} \rangle + b_f \right) \tag{17}$$

Por otra parte, como la hipótesis o función de decisión generada por la SVM se construye sólo a partir de los vectores de soporte (instancias más informativas para la tarea de clasificación), una modificación adicional que puede realizarse a esta propuesta considera, en vez de los valores de N_1 y N_2, el número de vectores de soporte de las clases positiva y negativa (N_{sv1} y N_{sv2}), de la función de decisión obtenida por la SVM y se construye un nuevo *bias* b_{f2} como,

$$b_{f2} = -\frac{N_{SV1} \alpha + N_{SV2} \beta}{N_{SV1} + N_{SV2}} \tag{18}$$

donde ahora la nueva función de clasificación sería,

$$f_{new2}(\boldsymbol{x}) = signo\left(\sum_{i=1}^{N} \alpha_i y_i \langle \boldsymbol{x}_i \cdot \boldsymbol{x} \rangle + b_{f2}\right) \quad (19)$$

Con estos nuevos sesgos, el hiperplano aprendido por la SVM es desplazado para conseguir un mejor rendimiento de clasificación para la clase positiva, tomando en cuenta las proporciones de las clases: a mayor desbalance, mayor margen de separación para la clase minoritaria. Esto puede observarse en la Figura 2, donde se muestra el desplazamiento del límite de decisión para los casos separable y no separable. Sin embargo, este resultado se obtiene a expensas de una disminución del margen para la clase negativa. Estas implicaciones serán discutidas posteriormente.

5. Experimentos y análisis de resultados

Para comprobar las prestaciones de esta nueva estrategia de post-procesamiento se utilizaron 23 conjuntos de datos del Repositorio UCI [35]. Las características de estos conjuntos se muestran en la Tabla 1. En ellos, se asignó la clase (+) a la etiqueta que se muestra entre paréntesis y la clase (-) al resto de los datos. Para medir el rendimiento de los clasificadores obtenidos al utilizar los nuevos sesgos, se utilizó la media geométrica (g-media) y la sensibilidad, ésta última para verificar el rendimiento sobre la clase minoritaria. Como técnica de evaluación, se aplicó la validación cruzada de 10 particiones. Para el entrenamiento de las SVM se utilizó un kernel RBF y el Toolbox de Bioinformática de Matlab.

Tabla 1. Descripción de los conjuntos de datos utilizados en este estudio

CONJUNTO DE DATOS	NO. TOTAL DE INSTANCIAS	NO. DE INSTANCIAS POSITIVAS	PORCENTAJE DE INSTANCIAS POSITIVAS
Abalone (19)	4177	32	0.76
Winequality_red (8)	1593	18	1.11
Page-Blocks (5)	5473	115	2.05
Tiroides (1)	3772	93	2.47
Nursey (3)	12960	328	2.53
Yeast (5)	1483	51	3.44
Letter (a)	20000	789	3.95
Car (3)	1728	69	3.99
Ecoli (5)	336	20	5.95
Balance(2)	625	49	7.24
Satimage (4)	6435	626	9.70
Euthyroid	2000	238	11.90
Glass (7)	214	29	13.55
Segment (1)	2310	330	14.29
Hepatitis	129	24	18.60
Cmc (2)	1473	333	22.61
Vehicle (1)	846	199	23.52
Transfusion	748	178	23.80
Haberman	306	81	26.50
German	1000	300	30.00
Waveform (0)	5000	1657	33.00
Pima Diabetes	768	268	34.00
TicTac (2)	958	332	34.66

La Tabla 2 muestra los valores promedio de g-media y sensibilidad para cada conjunto de datos aplicando 10 validaciones cruzadas de 10 particiones (en total 100 experimentos). A partir de esta tabla se puede concluir:

Tabla 2. Valores promedio de la media geométrica y la sensibilidad para 10 validaciones cruzadas de 10 particiones.

CONJUNTO DE DATOS	G-MEDIA			SENSIBILIDAD		
	SVM	b_f	b_{f2}	SVM	b_f	b_{f2}
Abalone (19)	0.000	0.625	0.552	0.000	**0.828**	0.592
Winequality_red (8)	0.000	**0.603**	0.296	0.000	**0.835**	0.280
Page-Blocks (5)	0.535	0.843	**0.894**	0.298	**0.982**	0.936
Tiroides (1)	0.851	**0.950**	0.886	0.733	**0.976**	0.794
Nursey (3)	0.878	**0.983**	0.602	0.774	**0.986**	0.370
Yeast (5)	0.000	0.595	**0.728**	0.000	**0.911**	0.601
Letter (a)	0.975	0.995	**0.994**	0.951	**0.996**	0.991
Car (3)	0.000	**0.941**	0.595	0.000	**0.911**	0.390
Ecoli (5)	0.823	**0.935**	0.891	0.740	**0.925**	0.850
Balance(2)	0.749	0.830	**0.851**	0.623	**0.909**	0.867
Satimage (4)	0.811	**0.889**	0.830	0.678	**0.865**	0.716
Euthyroid	0.772	0.748	**0.855**	0.615	**0.956**	0.796
Glass (7)	0.862	0.839	**0.923**	0.782	**0.915**	0.892
Segment (1)	0.988	**0.994**	0.991	0.977	**0.995**	0.984
Hepatitis	0.643	0.732	**0.744**	0.562	**0.855**	0.840
Cmc(2)	0.545	**0.597**	0.584	0.370	**0.609**	0.492
Vehicle (1)	0.982	**0.985**	0.985	0.975	**0.988**	0.986
Transfusion	0.540	**0.614**	0.557	0.327	**0.462**	0.350
Haberman	0.464	**0.617**	0.609	0.273	**0.606**	0.512
German	0.667	0.660	**0.695**	0.517	**0.804**	0.680
Waveform (0)	0.877	**0.884**	0.863	0.824	**0.921**	0.783
Pima Diabetes	0.672	**0.725**	0.676	0.518	**0.687**	0.530
TicTac (2)	0.974	**0.996**	0.997	0.950	0.995	**0.998**

- 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).
- Para todos los conjuntos de datos, utilizar un nuevo sesgo (b_f o b_{f2}) mejora considerablemente los resultados, sobre todo a nivel de sensibilidad, con respecto a la SVM original. Además, el sesgo b_f es el que ofrece mejores prestaciones; esto se refleja particularmente en la mejora en la precisión sobre la clase positiva.
- Al trabajar con conjuntos de datos no balanceados las métricas de evaluación g-media y sensibilidad logran medir el desempeño de los clasificadores independiente de la distribución, por lo que su selección resulta acertada para este tipo de problemas.

6. Conclusiones y trabajos futuros.

A partir de los resultados experimentales sobre conjuntos de datos con diferentes grados de desbalance, se puede concluir que el desempeño de la SVM logra mejorarse notablemente

utilizando un nuevo sesgo que toma en cuenta la proporción de las clases. Una ventaja importante es que el problema de optimización para hallar la SVM no es modificado ni es necesaria la entonación de nuevos parámetros, por lo cual el coste computacional es casi nulo.

Por otra parte, se pudo determinar que la mejora en la sensibilidad es a costa de una desmejora en la precisión de la clase negativa. Sin embargo, a pesar de un incremento de los falsos positivos, se logra el objetivo de mejorar en gran medida la precisión sobre la clase positiva, lo cual es crucial en las aplicaciones a las cuales va dirigido este estudio. Por otro lado, hay que resaltar que en los trabajos relacionados que fueron revisados, son pocos los que ofrecen información de cómo afecta la mejora en las clasificaciones de la clase positiva en el rendimiento sobre la clase negativa.

Como trabajo futuro se está considerando la aplicación de la propuesta a otros conjuntos de datos no balanceados, así como su comparación con otros métodos reportados en la literatura.

Referencias

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

[2] Fernández, A., García, S., Herrera, F. (2011). *Addressing the Classification with Imbalanced Data: Open Problems and New Challenges on Class Distribution*. E. Corchado, M. Kurzyński, M. Woźniak (Eds.): HAIS 2011, Part I, LNAI 6678, pp. 1–10, 2011. Springer-Verlag.

[3] Garcia, V., Sánchez, J.S., Mollineda, R.A., Alejo, R., Sotoca, J.M. (2007). *The class imbalanced problem in pattern classification and learning*. II Congreso Español de Informática. 283-291. Thomson.

[4] Guo, X., Yin, Y., Dong, C., Yang, G., Zhou, G. (2008). *On the class imbalanced problem*. Proceedings 4th International Conference on Neural Computation. 192-201.

[5] Cristianini, N., Shave-Taylor, J. (2000). *An Introduction to Supoprt Vector Machine and other kernel-based learning methods*. Cambridge University Press.

[6] Vapnik, V. (1999). *The nature of Statistical Learning Theory*. Second edition. Springer.

[7] Sun, A., Lim, E., Liu, Y. (2009). *On strategies for imbalanced text classification using SVM: A comparative study*. Decision Support Systems. Elsiever. 48:191-201.

[8] López, V., Fernández, A., Moreno-Torres, J., Herrera, F. (2012). *Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics*. Expert Systems with Applications 39(7):6585-6608.

[9] Akbani, R., Kwek, S. Japkowicz, N. (2004). *Applying Support Vector Machines to imbalanced datasets*. J.-F. Boulicaut et al. (Eds.): ECML 2004, LNAI. 3201:39–50. Springer-Verlag.

[10] Iman, T. Ming, K., Kamruzzaman, J. (2006). *z-SVM: An SVM for improved classificaction of imbalanced data*. A. Sattar and B.H. Kang (Eds.): AI 2006, LNAI. 4304:264–273. Springer-Verlag.

[11] He, H., Ghodsi, A. (2010). *Rare Class Classification by Support Vector Machine*. Proceedings 20th International Conference on Pattern Recognition. 548-551. Istanbul, Turkey. IEEE, ISBN: 978-0-7695-4109-9.

[12] Nguyen, H., Cooper, E., Kamei, K. (2009). *Bordeline Over-sampling for Imbalanced Data Classification*. Proceedings Fifth International Workshop on Computational Intelligence & Applications. IEEE. 24-29.

[13] Wu, G., Chang, E. (2005). *KBA: Kernel Boundary Alignment considering imbalanced data distribution*. IEEE Transactions on Knowledge and Data Engineering. 17(6):786-795.

[14] Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W. (2002). *SMOTE: Synthetic Minority Over-sampling TEchnique*. Journal of Artificial Intelligence Research. 16:341-378.

[15] Vilariño, F. Spyridonos, P., Vitríà, J., Radeva, P. (2005). *Experiments with SVM and Stratified Sampling with an imbalanced problem: detection of intestinal contractions*. S. Singh et al. (Eds.): ICAPR 2005, LNCS. 3687:783–791. Springer-Verlag.

[16] Li, P., Qiao, P., Liu, Y. (2008). *A Hybrid Re-sampling method for SVM learning form imbalanced data sets*. Proceedings of Fifth International Conference on Fuzzy Systems and Knowledge Discovery. IEEE. 65-69.

[17] Yu, T., Debenham, T., Simoff, S. (2006). *Combine Vector Quantization and Support Vector Machines for imbalanced datasets*. Artificial Intelligence in Theory and Practice. IFIP International Federation for Information Processing. 217:81-88.

[18] Castro, C., Carvalho, M., Braga, A. (2009). *An Improved Algorithm for SVMs Classification of Imbalanced Data Sets*. D. Palmer-Brown et al. (Eds.): EANN 2009, CCIS. 43:108–118. Springer-Verlag.

[19] Wang, H. (2008). *Combination approach of SMOTE and biased-SVM for imbalanced datasets*. Proc. IEEE International Joint Conference on Neural Networks. 228-231.

[20] Tang, Y. Zhang, Y., Chawla, N., Krasser, S. (2009). *SVMs modeling for high imbalanced classificaction*. IEEE Transactions on Systems, Man and Cybernetics – Part B. 39(1):281-288.

[21] Kang, P., Cho, S. (2006). *EUS SVMs: Ensemble of Under-Sampled SVMs for data imbalanced problems*. I. King et al. (Eds.): ICONIP 2006, Part I, LNCS. 4232:837–846. Springer-Verlag.

[22] Liu, Y. An, A., Huang, X. (2006). *Boosting prediction accuracy on imbalanced datasets with SVM Ensembles*. W.K. Ng, M. Kitsuregawa, and J. Li (Eds.): PAKDD 2006, LNAI 3918:107–118. Springer-Verlag.

[23] Waske, B., Benediktsson, J., Sveinsson, J. (2009). *Classifying Remote Sensing Data with Support Vector Machines and Imbalanced Training Data*. J.A. Benediktsson, J. Kittler, and F. Roli (Eds.): MCS 2009, LNCS. 5519:375–384. Springer-Verlag.

[24] Yang, P., Zhang, Z., Zhou, B., Zomaya, A. (2011). *Sample subset optimization for classifying imbalanced biological data*. Advances in Knowledge Discovery and Data Mining. Lecture Notes in Computer Science. 6635:333-344. Springer-Verlag.

[25] Cohen, G. Hilario, M. Sax, H., Hugonnet, S., Geissbuhler, A. (2006). *Learning from imbalanced data in surveillance of nocosomial infection*. Artificial Intelligence in Medicine. 37:7-18. Elsiever.

[26] Veropoulos, K., Campbell, C., Cristianini, N. (1999). *Controlling the sensitivity of support vector machines*. Proc. Int. Joint Conf. on Artificial Intelligence. 55–60.

[27] Yang, C., Wang, J., Yang, J., Yu, G. (2008). *Imbalanced SVM learning with margin compensation*. F. Sun et al. (Eds.): ISNN 2008, Part I, LNCS. 5263:636–644. Springer-Verlag.

[28] Batuwita, R., Palade, V. (2010). *FSVM-CIL: Fuzzy Support Vector Machines for Class Imbalance Learning*. IEEE Transactions on Fuzzy Systems, 18(3):558-571.

[29] Wang, B. Japkowicz, N. (2008). *Boosting Support Vector Machines for Imbalanced Data Sets*. A. An et al. (Eds.): ISMIS 2008, LNAI. 4994:38-47. Springer-Verlag.

[30] Li, B., Hu, J., Hirasawa, K. (2008a). *Support Vector Machines with WHM offset for unbalanced data*. Journal of Advanced Computational Intelligence and Intelligence Informatics. 12(1):94-101.

[31] Shanathan, J., Roma, N. (2003). *Improving SVM text classification performance through threshold adjustment*. N. Lavrač et al. (Eds.): ECML'2003. LNAI 2837:361–372. Springer-Verlag.

[32] Li, B., Hu, J., Hirasawa, K. (2008b). *An improved Support Vector Machine with soft decision-making boundary*. Proceedings International Conference on Artificial Intelligence and Applications. Innsbruck, Austria. 40-45.

[33] Wang, H., Zheng, H. (2008). *An improved Support Vector Machine for the classification of imbalanced biological datasets*. D.-S. Huang et al. (Eds.): ICIC'2008. LNCS 5226:63-70. Springer-Verlag.

[34] Gonzalez-Abril, L., Angulo, C. Velasco, F., Ortega, J.A. (2008). *A Note on the Bias in SVMs for classification*. IEEE Transactions on Neural Networks. 19(4): 723-725.

[35] Frank, A., Asuncion, A. (2010). *UCI Machine Learning Repository*. Irvine, CA: University of California, School of Information and Computer Science. [http://archive.ics.uci.edu/ml].