# An Empirical Comparison Of Feature Selection Methods In Problem Transformation Multi-label Classification

J. M. Rodriguez, D. Godoy and A. Zunino

*Abstract*— **Multi-label classification (MLC) is a supervised learning problem in which a particular example can be associated with a set of labels instead of a single one as in traditional classification. Many real-world applications, such as Web page classification or resource tagging on the Social Web, are challenging for existing MLC algorithms, because the label space grows exponentially as instance space increases. Under the problem transformation approach, the most common alternative for MLC, multi-label problems are transformed into several single label problems, whose outputs are then aggregated into a prediction to the whole classification problem. Feature selection techniques become crucial in large-scale MLC problems to help reducing dimensionality. However, the impact of feature selection in multi-label setting has not been as extensively studied as in the case of single-label data. In this paper, we present an empirical evaluation of feature selection techniques in the context of the three main problem transformation MLC methods: Binary Relevance, Pair-wise and Label power-set. Experimentation was performed across a number of benchmark datasets for multi-label classification exhibiting varied characteristics, which allows observing the behavior of techniques and assessing their impact according to multiple metrics.**

*Keywords*— **Multi-label Classification; Feature Selection; Problem Transformation Classification; Binary Relevance; Pair-Wise; HOMER**

## I. INTRODUCCIÓN

LOS PROBLEMAS de aprendizaje supervisado se ocupan comúnmente de asignar una única clase o etiqueta a un conjunto de ejemplos en base a las asociaciones previamente descubiertas entre las características que describen los ejemplos y las potenciales clases. Para descubrir estas asociaciones se utiliza un conjunto de ejemplos de entrenamiento para los cuales se conocen las clases asociadas. Sin embargo, en la actualidad cada vez hay más dominios en los que un ejemplo puede asociarse con más de una etiqueta, es decir las clases no son disjuntas. Este tipo de problemas se conocen como clasificación multi-etiqueta (MLC del inglés Multi-Label Classification). Un ejemplo de ellos son las páginas Web que suelen tener más de un tópico de interés o las imágenes que pueden contener más de un objeto visible.

J. M. Rodriguez, ISISTAN, Universidad Nacional del Centro de la Provincia de Buenos Aires (UNICEN) - Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET). Tandil, Buenos Aires, Argentina, juanmanuel.rodriguez@isistan.unicen.edu.ar

D. Godoy, ISISTAN, UNICEN-CONICET. Tandil, Buenos Aires, Argentina, daniela.godoy@isistan.unicen.edu.ar

A. Zunino, ISISTAN, UNICEN-CONICET. Tandil, Buenos Aires, Argentina, alejandro.zunino@isistan.unicen.edu.ar

Los problemas de clasificación multi-etiqueta no solo son cada vez más frecuentes en aplicaciones prácticas reales, sino también de mayor escala. La Web social, por ejemplo, ha traído consigo muchas aplicaciones donde los usuarios comparten recursos de distinto tipo (por ej. documentos, videos, imágenes) y los clasifican de acuerdo a sus diferentes intereses y puntos de vista. Por ejemplo, las folcsonomías son estructuras resultantes de los sistemas de anotación social o colaborativo que cada vez son más usados en la Web, como Delicious (http://delicious.com/) o Flickr (http://www.flickr.com/). En estos sistemas la recomendación de etiquetas [25] es un problema multi-etiqueta a gran escala, ya que millones de usuarios comparten recursos a los que etiquetan libremente, resultando en una gran base ejemplos, cada uno asociado a varias etiquetas.

En este contexto, cobran relevancia las técnicas dedicadas a la reducción de dimensionalidad para lograr un más eficiente y efectivo proceso de aprendizaje. La Selección de Características (FS del inglés Feature Selection) es un proceso mediante el cual se elige un subconjunto de características relevantes, del total de características o atributos que describen los ejemplos, para usarlos en la construcción del clasificador o modelo. Este proceso de selección por un lado reduce los tiempos de aprendizaje y, además, simplifica los modelos adquiridos tendiendo a generar mejores generalizaciones, reduciendo así el problema de sobreajuste.

Si bien el problema de selección de características ha sido ampliamente estudiado en el contexto de la clasificación tradicional, unos pocos estudios han evaluado su impacto en problemas multi-etiqueta [26, 4, 13]. Estos estudios, además, incluyeron un número limitado de algoritmos o para un único conjunto de datos. El presente trabajo busca evaluar empíricamente el efecto de la selección de características en una serie de algoritmos representativos de los tres enfoques usados para clasificación multi-etiqueta, utilizando múltiples conjuntos de datos.

El presente artículo está organizado como sigue. La Sección II introduce el problema de clasificación multi-etiqueta y los enfoques basados en transformación en los que se enfocó el presente estudio, mientras que la Sección III describe los métodos de selección de características en este contexto. La Sección IV detalla el estudio realizado y su diseño experimental. Los resultados obtenidos se presentan en la Sección V. La Sección VI discute los trabajos relacionados. Finalmente, en la Sección VII se presentan las conclusiones obtenidas como resultado de este estudio.

## II. APRENDIZAJE MULTI-ETIQUETA BASADO EN TRANSFORMACIÓN

La clasificación tradicional se ocupa del aprendizaje a partir de un conjunto de ejemplos que están asociados a una única etiqueta o clase $l$ de un conjunto disjunto de etiquetas $|L|$. En los problemas multi-etiqueta, en cambio, los ejemplos están asociados a un conjunto de etiquetas $Y \subseteq L$. Formalmente, sea $D$ un conjunto de datos compuesto de $N$ ejemplos $E_i = (x_i, Y_i)$ con $i = 1 \ldots N$. Cada ejemplo $E_i$ está asociado con un vector $\mathrm{x_i} = (\mathrm{x_{i1}}, \mathrm{x_{i2}}, \ldots, \mathrm{x_{iM}})$ descripto por $M$ características $x_t$, con $t = 1 \ldots M$, y un vector binario $Y_i = (y_{i1}, y_{i2}, \ldots, y_{iQ})$ donde $y_{ij} \in \{1, -1\}$ indicando si la $i$-ésima instancia está asociada a la $j$-ésima clase, siendo $L = \{l_1, l_2, \ldots, l_Q\}$ el conjunto de etiquetas. El objetivo de la clasificación multi-etiqueta es usar la información del conjunto de entrenamiento para obtener un clasificador $h: X \rightarrow Y$, que asigne correctamente un conjunto de etiquetas a una nueva instancia o ejemplo.

El enfoque más común para la clasificación multi-etiqueta es el que se conoce como *transformación del problema*, que tiene por objetivo convertir el conjunto de datos original, de naturaleza multi-etiqueta, en uno o más conjuntos de datos con una sola etiqueta que puedan ser tratados con modelos de clasificación tradicionales. En este caso, un clasificador realiza la clasificación respecto de una única clase y sus resultados se vuelven a transformar en representaciones multi-etiqueta. Para estos problemas más pequeños de etiqueta simple existen una variedad de algoritmos de aprendizaje disponibles. Las alternativas al enfoque de transformación son la adaptación de los algoritmos de etiqueta simple para manejar múltiples etiquetas, como por ejemplo Multi-label C4.5 [3] o ML-KNN [32]. La principal ventaja del enfoque de transformación con respecto a los métodos de adaptación del algoritmo radica en la fácil incorporación de métodos tradicionales de pre-procesamiento y clasificación. Además, la mayoría son fácilmente paralelizables, reduciendo los tiempos de aprendizaje en computadoras multi-núcleo [24].

Los métodos de transformación pueden agruparse en tres categorías: relevancia binaria, comparación por pares y conjunto potencia. A continuación se describen estos tres métodos y los algoritmos empleados en este estudio dentro de cada uno de ellos.

### *II-A. Relevancia Binaria*

Relevancia Binaria (RB) es una estrategia bien conocida que descompone un problema de aprendizaje multi-etiqueta en un clasificador para cada etiqueta. Cada clasificador binario se entrena utilizando todos los ejemplos de una etiqueta determinada como positivos y los ejemplos restantes como negativos. Para la predicción, cada clasificador decide si una etiqueta es relevante para el ejemplo, lo que resulta en un conjunto de etiquetas pertinentes, que pueden ser calificadas por probabilidad. Por lo tanto, los clasificadores RB entrenan $|L|$ clasificadores binarios $C_1, C_2, \ldots, C_{|L|}$, donde cada clasificador es responsable de la predicción de cada etiqueta correspondiente $l_j \in L$.

RB considera cada etiqueta como un problema binario independiente, y aunque la independencia de las etiquetas puede ser vista como un supuesto débil, debido a que RB ignora posibles correlaciones entre etiquetas, RB ha mostrado muy buen desempeño comparado con métodos más complejos [15]. Los Clasificadores en Cadena (CC) [23] involucran $|L|$ clasificadores binarios como RB, pero estos clasificadores se enlazan a lo largo de una cadena en la que cada clasificador trata un problema de relevancia binaria asociado con la etiqueta $l_j \in L$. El espacio de características de cada eslabón de la cadena se extiende con las asociaciones de la etiqueta de todos los enlaces anteriores. Por lo tanto, cada clasificador $C_j$ en la cadena es responsable del aprendizaje y la predicción de la asociación binaria de la etiqueta $l_j$ dado el espacio de características, aumentado por todas las predicciones de relevancia binarios anteriores en la cadena de $l_1, \ldots, l_{j-1}$. El proceso de clasificación se inicia en el primer clasificador y se propaga a lo largo de la cadena. Este método de encadenamiento pasa la información de las etiquetas entre los clasificadores, lo que permite tomar en cuenta las correlaciones de etiquetas y superar así el problema de independencia de los clasificadores RB manteniendo al mismo tiempo una complejidad computacional aceptable.

### *II-B. Comparación por Pares*

*II-B1. Ranking de Comparación por Pares:* La comparación por pares o round-robin [6, 30] trata el problema de multi-etiqueta aprendiendo un clasificador binario para cada par de etiquetas. Es decir, para las $Q$ etiquetas posibles $l_1, l_2, \ldots, l_Q$ , se generan un total de $Q(Q-1)/2$ clasificadores binarios, uno para cada par de etiquetas $(l_j, l_k)$ con $1 \leq j < k \leq Q$. Cada clasificador se entrena con los ejemplos positivos de la primera etiqueta y los ejemplos de la segunda como negativos. En general el método de comparación por pares adopta luego un algoritmo de votación por mayoría donde cada clasificador predice o vota por una etiqueta y las etiquetas se ordenan en un ranking de acuerdo a la suma de sus votos.

*II-B2. CLR:* CLR (Calibrated Label Ranking) [7] es una extensión de la calibración por pares que introduce una etiqueta adicional o etiqueta de calibración, que puede interpretarse como un punto de corte entre las etiquetas relevantes e irrelevantes. Loza Mencía et al. [18] propusieron una adaptación del algoritmo Quick Weighted Voting (QWeighted) [20] para aprendizaje multi-etiqueta denominado QWML (QWeighted Multi-label Learning).

### *II-C. Conjunto Potencia*

Este enfoque de aprendizaje multi-etiqueta transforma el problema en una tarea de aprendizaje multi-clase [1]. Para ello considera combinar conjuntos de etiquetas en una única etiqueta atómica, formando un problema de clasificación de etiqueta simple. En este último problema, el conjunto de

etiquetas posibles está dado por todos los subconjuntos de etiquetas distintos de la representación multi-etiqueta. En este caso, el enfoque toma en cuenta la correlación entre etiquetas. Sin embargo, el espacio posible de etiquetas puede ser extremadamente grande y para algunos valores de clase puede haber pocos ejemplos, convirtiéndose en un problema multi-clase desbalanceado.

Con este mismo enfoque, HOMER (Hierarchy Of Multi-label classifiERs) [27] construye una jerarquía de etiquetas en base a un algoritmo de agrupamiento, generando un árbol de clasificadores, cada uno de ellos para un subconjunto cada vez más reducido de etiquetas.

## III. SELECCIÓN DE CARACTERÍSTICAS PARA CLASIFICACIÓN MULTI-ETIQUETA

Las técnicas de selección de características se ocupan de la búsqueda de un subconjunto de características, entre todos los posibles conjuntos de características, que le permitan a un algoritmo de aprendizaje en particular inducir la mejor hipótesis de acuerdo a alguna medida de efectividad. Se pueden distinguir dos alternativas generales con respecto a la función de selección: selección por filtro o por wrapper [11]. Los métodos de selección de características por filtro son independientes del algoritmo de aprendizaje, mientras que los métodos de wrapper emplean el algoritmo de aprendizaje como parte de la evaluación del subconjunto de características.

Como la eficacia de clasificación depende fundamentalmente del desvío introducido por el algoritmo de aprendizaje, los enfoques de wrapper son generalmente preferidos a los enfoques de filtro, debido a su eficacia. Por el contrario, es computacionalmente caro aplicar el algoritmo de aprendizaje una vez o incluso más veces en los ejemplos de entrenamiento para cada subconjunto a ser considerado como requieren los métodos de wrapper. Esta desventaja puede ser particularmente seria cuando la cantidad de características es grande y, como en este caso, la escala se incrementa por ser problemas multi-etiqueta.

Los algoritmos de selección evalúan la bondad de una característica o bien en forma individual o bien como parte de un subconjunto. La evaluación individual es computacionalmente menos cara, ya que considera las características en forma aislada y les asigna un peso (ranking) de acuerdo con su poder de predicción de la clase. Con este fin, se han propuesto varias medidas de importancia de característica como $\chi^2$ (Chi-cuadrado), coeficiente de correlación, ganancia de información, información mutua y odds ratio [31], entre otras. Sin embargo, la evaluación individual no permite detectar características redundantes, ya que usualmente éstas obtienen puntaje similar usando estas medidas. El enfoque de la evaluación en subconjuntos en cambio considera tanto la relevancia de una característica como su redundancia.

Los métodos de subconjunto involucran la búsqueda en el espacio de características del subconjunto con mayor probabilidad de predecir la clase correctamente. Una medida numérica guía la búsqueda de dicho subconjunto. En este trabajo se usó la selección basada en correlación [8] (CFS del inglés Correlation-based Feature Subset) que determina la habilidad predictiva de una característica individualmente y la redundancia entre ellas, prefiriendo conjuntos de atributos que tengan alta correlación con la clase, pero baja intercorrelación.

Típicamente la búsqueda se realiza de forma greedy hacia adelante o hacia atrás, donde en cada paso el conjunto de características se modifica localmente agregando o quitando una característica. Las dos opciones usadas en este estudio fueron BestFirst y GreedyStepwise. La opción BestFirst realiza una búsqueda hill climbing con backtracking, que puede realizarse hacia adelante desde el conjunto vacío de atributos o hacia atrás desde el conjunto completo. La opción GreedyStepwise es una búsqueda greedy a través del espacio de atributos que, como BestFirst, puede ir hacia adelante o hacia atrás. A diferencia de la primera, no hace backtracking, pero termina tan pronto como la siguiente característica agregada o eliminada disminuye la métrica de evaluación.

La selección de características en el contexto de los problemas de transformación para clasificación multi-etiqueta consiste en primero transformar el conjunto de datos multi-etiqueta en múltiples conjuntos de una sola etiqueta, sobre los que se seleccionan las características. Por ejemplo, la Fig. 1 muestra la selección en el caso de usarse RB. La función $f^T$ es la encargada de descomponer el problema multi-etiqueta en $Q$ problemas binarios, mientras que la función $f^A$ denota la función usada para integrar la información de todos los clasificadores binarios. En la figura $y^k$ denota que el clasificador usa la $k$-esima columna para aprender una hipótesis $h_k$ para dicha etiqueta en base a un conjunto de entrenamiento $\mu_k$ . Una vez transformado el conjunto de datos original, el método de selección de características elegido, denotado FS, se aplica sobre cada conjunto transformado, de manera que cada clasificador utilice las características más relevantes para cada problema de clasificación individual (en RB distinguir una clase del resto).

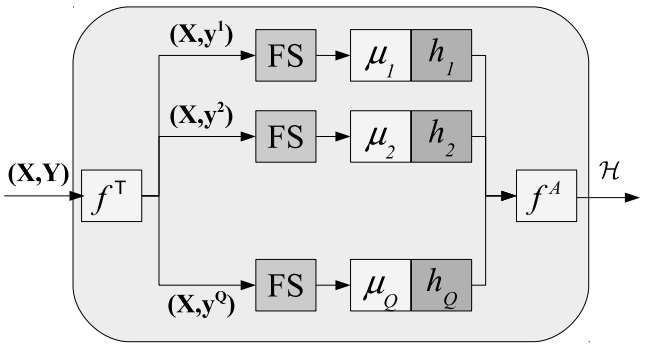


Figura 1. RB con selección de características.

## IV. DIMENSIONES DEL ESTUDIO

El presente estudio se enfocó en determinar el impacto de la selección de características en la clasificación multi-etiqueta

dentro del enfoque de transformación. Para ello se evaluaron algoritmos dentro de las tres ramas principales, relevancia binaria (RB y CC), comparación por pares (CLR y QWML) y conjunto potencia (HOMER). Utilizando diferentes conjuntos de datos se midió el efecto de la selección basada en evaluación por subconjuntos en términos de la efectividad del clasificador y el nivel de reducción de dimensionalidad. En esta sección se describen los conjuntos de datos, las métricas para evaluar la performance de los clasificadores y el diseño de los experimentos.

### *IV-A. Conjuntos de datos*

En este estudio se realizaron experimentos con 10 conjuntos de datos diferentes, con el fin de estudiar el impacto de la selección para distintas variaciones en los datos. Para caracterizar las propiedades de un conjunto de datos multi-etiqueta existen muchas posibilidades. La cardinalidad indica el numero promedio de etiquetas por instancia. La densidad normaliza la cardinalidad por el número de etiquetas posibles en el espacio de etiquetas. Otra medida habitual es la diversidad, el número de conjuntos de etiquetas distintos que aparecen en los datos, que se normaliza por el número de ejemplos para indicar la proporción de conjuntos de etiquetas distintos.

Los conjuntos de datos usados en este estudio se resumen en la Tabla I. Para cada uno de ellos se indica el número de ejemplos $N$ (con una d si son discretos o n si son numéricos), el número de características $M$, el número de etiquetas $|L|$, la cardinalidad $LC$, la densidad $LD$ y el número de combinaciones distintas de etiquetas $DC$.

TABLA I
CONJUNTOS DE DATOS USADOS EN EL ESTUDIO

| | N | M | \|L\| | LC | LD | DC |
|---|---|---|---|---|---|---|
| **bibtex** | 7395 | 1836 (d) | 159 | 2.40 | 0.02 | 2856 |
| **cal500** | 502 | 68 (n) | 174 | 26.04 | 0.15 | 502 |
| **corel16k** | 13766 | 500 (d) | 153 | 2.86 | 0.02 | 4803 |
| **corel5k** | 5000 | 499 (d) | 374 | 3.52 | 0.01 | 3175 |
| **emotions** | 593 | 72 (n) | 6 | 1.87 | 0.31 | 27 |
| **enron** | 1702 | 1001 (d) | 53 | 3.38 | 0.06 | 753 |
| **genbase** | 662 | 1186 (d) | 27 | 1.25 | 0.05 | 32 |
| **medical** | 978 | 1449 (d) | 45 | 1.25 | 0.03 | 94 |
| **scene** | 2407 | 294 (n) | 6 | 1.07 | 0.18 | 15 |
| **yeast** | 2417 | 103 (n) | 14 | 4.24 | 0.30 | 198 |

### *IV-B. Métricas*

Las métricas de evaluación en problemas multi-etiqueta pueden dividirse en las basadas en ejemplos o basadas en etiqueta. Las primeras evalúan el clasificador multi-etiqueta aprendido sobre cada ejemplo del conjunto de prueba de manera separada y devuelven una media de todos los ejemplos. En cambio, las métricas basadas en etiqueta evalúan el clasificador sobre cada etiqueta separadamente y luego devuelven una media calculada a nivel micro o macro con todas las etiquetas. En este trabajo se utilizó Hamming Loss o distancia de Hamming como métrica basada en ejemplos y Macro F-Measure o Valor F como métrica basada en etiqueta, tal como se definen en [33].

Hamming Loss evalúa la fracción de pares instancia-etiqueta mal clasificados, es decir, cuando una etiqueta relevante no se predice o se predice una irrelevante. De modo que, cuanto menor el valor de Hamming Loss, mejor es la performance del clasificador. Esta medida se define como:

$$Hamming - loss(H,D) = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i \Delta Z_i|}{|L|} \quad (1)$$

donde $Y_i$ y $Z_i$ son las etiquetas verdaderas y las que el clasificador predijo, respectivamente, y $\Delta$ representa la diferencia simétrica entre los dos conjuntos.

Macro F-Measure es una media pesada de la precisión y recall de un clasificador y se define como:

$$MacroF - Measure\ (H,D) = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{2|Y_i \cap Z_i|}{Z_i + Y_i} \quad (2)$$

### *IV-C. Diseño Experimental*

La comparación de los métodos multi-etiqueta fue realizada usando la librería MULAN [28], una librería Java que implementa varios algoritmos de aprendizaje multi-etiqueta tales como BR, HOMER y otros, que se encuentra a su vez basada en la librería de aprendizaje de máquina Weka [9]. Todos los resultados corresponden a la ejecución de los algoritmos con 10-fold cross validation. Como clasificadores de base se utilizó SMO [21], un algoritmo de optimización de secuencia mínima para entrenar máquinas de vectores de soporte (SVMs).

## V. RESULTADOS

Se evaluaron dos aspectos de los resultados de la selección de características. Primero, se midió el poder predictivo de los clasificadores aprendidos con y sin selección de características. Es decir, cuál es la capacidad predictiva de los clasificadores originales y cómo se vio afectada por la selección de características. Segundo, se midió cuál fue el grado de reducción de dimensionalidad en cada caso. Ambos aspectos se analizan en las secciones subsiguientes.

### *V-A. Performance de la Clasificación*

Las Fig. 2 y 3 muestran los valores de Macro F-Measure y Hamming Loss promediando todos los conjuntos de datos incluidos en el estudio y para los algoritmos de transformación usados. Como BASE se identifica el clasificador sin selección de características, es decir, usando la totalidad de las características disponibles. BestFirst (BF) y GreedyStepwise (GS) son las variantes de selección basada en subconjuntos descriptas anteriormente.

Como puede observarse los clasificadores BASE en todos los casos fueron superiores en performance en términos de Macro F-Measure. Si bien las ventajas obtenidas en promedio

fueron pequeñas, la distribución de los datos en cuartiles muestra que la mayoría de los datos se encuentran en el cuartil superior, no siendo este el caso en las variantes que usan selección de características. Los valores de Hamming Loss, por otro lado, muestran a algunas variantes obteniendo mejores niveles de performance en HOMER, CLR y BR. Nuevamente, las diferencias son de pequeña magnitud.

Otra observación que puede realizarse de estas figuras es que BestFirst y GreedyStepwise obtienen valores prácticamente equivalentes de performance en todos los casos. Por lo que la selección de uno u otro modo no afecta la capacidad predictiva de los clasificadores.

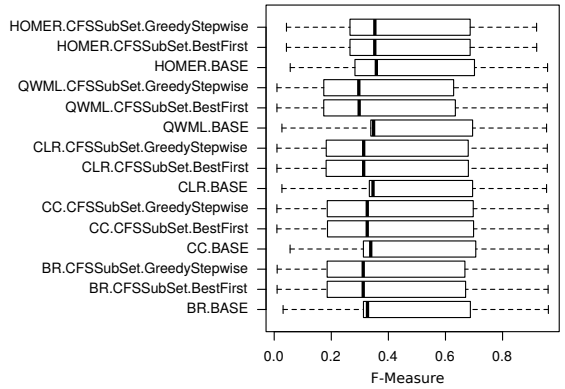


Figura 2. Resultados de la clasificación (Macro F-Measure).

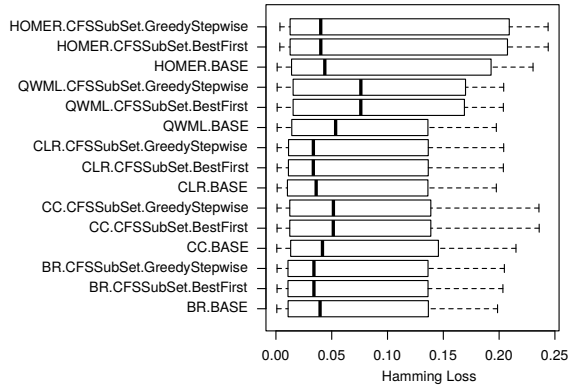


Figura 3. Resultados de la clasificación (Hamming Loss).

### *V-B. Reducción de Dimensionalidad*

Los valores de performance alcanzados para la clasificación multi-etiqueta no pueden ser analizados en forma aislada, sino junto con el grado de reducción de características en cada caso. Los clasificadores BASE consiguen mejor precisión a costa de procesar la totalidad de las características, lo cual puede requerir gran tiempo de procesamiento. En cambio, muchas veces puede ser preferible una pequeña pérdida en precisión si la selección reduce significativamente el costo computacional.

Para medir la reducción de dimensionalidad causada por la aplicación de un método de selección de características, se calculó el porcentaje usado de características por cada problema de clasificación individual con respecto a la totalidad de las características originales. En un problema de transformación, el número total de características usadas por el clasificador base es la misma para todos los clasificadores simples, lo que cambia son qué instancias se toman como positivas o negativas para una clase. Por ejemplo, los $Q$ clasificadores de relevancia binaria usan el número total de características, al igual que los $Q(Q-1)/2$ clasificadores del método de comparación por pares. Mientras que, aplicando selección de características cada clasificador de los $Q$ en RB, por ejemplo, van usar un número distinto de características que sean las relevantes para ese problema de clasificación de una única etiqueta.

La Tabla II muestra el porcentaje promedio de características usadas por los clasificadores estudiados, para las alternativas BestFirst (BF) y GreedyStepwise (GS), junto con su desvío estándar. Por ejemplo, para el conjunto *yeast*, cada clasificador usó en promedio 19.07% características, es decir alrededor de 20 características en cada clasificador binario contra las 103 características originales que se usarían sin un proceso de selección. Aunque en la última columna se muestra un promedio del porcentaje considerando todos los conjuntos (entre 91% y 92% de reducción en promedio), se puede ver que hay una gran variación en este sentido dependiendo del conjunto de datos. En conjuntos como *medical* y *genbase* se redujo el porcentaje de características usados a menos del 1%, mientras que en otros como *scene* y *emotions* la reducción fue menos drástica y se conservaron entre un 23% y 28% de las características originales.

Es natural pensar en este caso que el número de características relevantes para un clasificador guarda relación con la cantidad de características por clase o etiqueta. En la Fig. 4 se muestran los distintos conjuntos ordenados de menor a mayor de acuerdo a su densidad (LD), número de etiquetas por ejemplo divido por el número total de etiquetas, y el porcentaje promedio de características usadas en cada caso. Para todos los algoritmos de clasificación se muestra dicho porcentaje usando la alternativa GreedyStepwise, que es la que logró mejor reducción, aunque con diferencias no significativas respecto de BestFirst. Se puede observar que los conjuntos de mayor densidad son los que requieren mayor número de características, es decir que dejan menos margen de reducción de dimensiones.

La Fig. 5 ilustra la relación entre performance de los clasificadores y porcentaje de reducción, lo primero expresado en términos de Macro F-Measure y lo segundo en porcentaje de reducción realizado sobre la totalidad de las características. En todos los casos, se encuentran promediados los resultados para todos los conjuntos de datos con una selección BestFirst (BF). Los clasificadores BASE, que son los que lograron una mejor capacidad de predicción, se encuentra sobre el eje X ya que el aprendizaje se realizó sobre la totalidad de las características, sin reducción de características. En cambio, los

TABLA II
PORCENTAJE DE LAS CARACTERÍSTICAS ORIGINALES USADAS EN PROMEDIO PARA CLASIFICACIÓN

| | RB | | CC | | CLR | | QWML | | HOMER | |
|---|---|---|---|---|---|---|---|---|---|---|
| | BF | GS | BF | GS | BF | GS | BF | GS | BF | GS |
| **yeast** | 19.07±9.32 | 18.93±9.32 | 11.35±88.02 | 11.25±8.00 | 19.07±9.32 | 18.93±9.32 | 19.07±9.32 | 18.93±9.32 | 17.08±10.94 | 16.81±10.78 |
| **medical** | 0.94±0.66 | 0.93±0.66 | 0.94±80.68 | 0.94±0.68 | 0.94±0.66 | 0.93±0.66 | 0.94±0.66 | 0.93±0.66 | 1.32±1.34 | 1.31±1.34 |
| **scene** | 23.97±7.02 | 23.88±6.99 | 23.74±86.97 | 23.62±6.95 | 23.97±7.02 | 23.88±6.99 | 23.97±7.02 | 23.88±6.99 | 21.95±5.82 | 21.43±5.80 |
| **emotions** | 28.11±7.57 | 26.96±8.49 | 24.26±89.69 | 23.61±9.55 | 28.11±7.57 | 26.96±8.49 | 28.11±7.57 | 26.96±8.49 | 25.37±10.66 | 24.98±10.73 |
| **bibtex** | 1.63±0.57 | 1.61±0.56 | 1.60±80.49 | 1.58±0.48 | 1.63±0.57 | 1.61±0.56 | 1.63±0.57 | 1.61±0.56 | 2.25±0.73 | 2.20±0.72 |
| **cal500** | 4.07±3.08 | 4.05±3.06 | 15.88±87.63 | 15.50±7.41 | 4.07±3.08 | 4.05±3.06 | 4.07±3.08 | 4.05±3.06 | 3.81±2.12 | 3.81±2.11 |
| **corel5k** | 4.34±2.37 | 4.27±2.36 | 3.79±81.79 | 3.75±1.78 | 4.34±2.37 | 4.27±2.36 | 4.34±2.37 | 4.27±2.36 | 7.29±4.77 | 7.21±4.77 |
| **corel16k** | 6.52±4.05 | 6.39±4.09 | 4.53±82.71 | 4.44±2.69 | 6.52±4.05 | 6.39±4.09 | 6.52±4.05 | 6.39±4.09 | 11.98±2.84 | 11.92±2.89 |
| **genbase** | 0.55±0.20 | 0.55±0.20 | 0.55±80.20 | 0.55±0.20 | 0.55±0.20 | 0.55±0.20 | 0.55±0.20 | 0.55±0.20 | 0.61±0.20 | 0.60±0.20 |
| **enron** | 1.70±1.09 | 1.66±1.12 | 1.45±80.88 | 1.40±0.88 | 1.70±1.09 | 1.66±1.12 | 1.70±1.09 | 1.66±1.12 | 2.17±0.97 | 2.08±0.98 |
| **Promedio** | 9.09±10.47 | 8.92±10.23 | 8.81±9.42 | 8.66±9.25 | 9.09±10.47 | 8.92±10.23 | 9.09±10.47 | 8.92±10.23 | 9.38±9.19 | 9.23±9.02 |

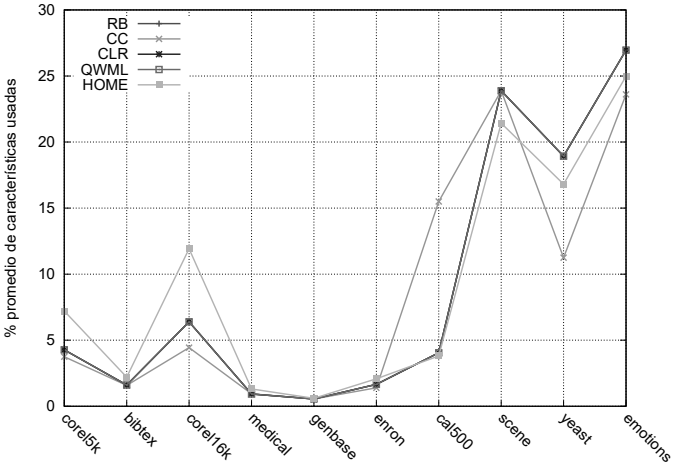

Figura 4. Reducción de características de acuerdo a densidad.

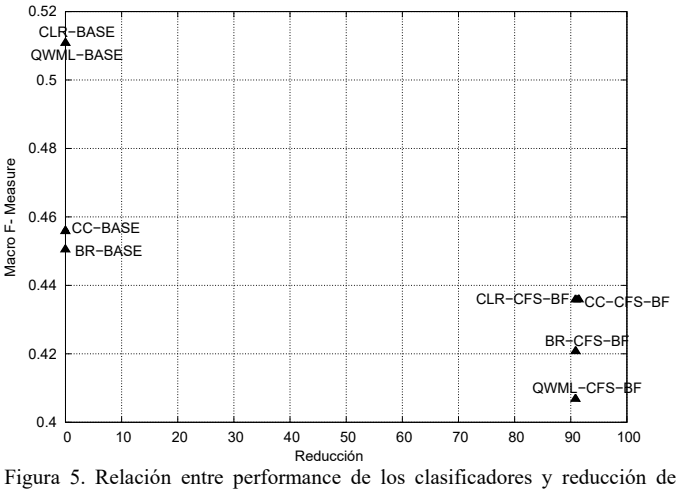

Figura 5. Relación entre performance de los clasificadores y reducción de características.

clasificadores con selección de características se encuentran sobre el otro extremo del eje X, con más del 90% de reducción. A pesar de haber usado un porcentaje mucho menor de características, la clasificación se redujo en menos de 0.1 (por ejemplo, en el caso de RB la performance cayó de 0.45 a 0.42). En conclusión, se puede observar que el compromiso entre performance y reducción es altamente beneficioso, ya que los clasificadores que se entrenan con un porcentaje muy reducido de características conservan en gran medida sus capacidades de predicción.

## VI. TRABAJOS RELACIONADOS

La clasificación multi-etiqueta es un problema que aparece con frecuencia en aplicaciones reales, por lo que está siendo cada vez más estudiado en trabajos que comparan los diferentes métodos de clasificación [16]. Por otro lado, las técnicas de selección de características también son objeto de estudio permanente dada la tendencia creciente a analizar datos de cada vez mayor dimensionalidad [19, 5], pero se analizan para casos de clasificación tradicional, donde una instancia se asocia con una única clase.

Pocos estudios han abordado el problema de selección de características en el contexto de la clasificación multi-etiqueta. En [26] dos métodos de selección como Ganancia de Información y ReliefF se estudian para dos clasificadores basados en transformación como relevancia binaria y conjunto potencia, sobre los mismos conjuntos de datos usados en este trabajo. Se concluye que el segundo genera conjuntos de características más pequeños sin degradar la performance de los clasificadores. Doquire et al. [4] usan Información Mutua para seleccionar características en un enfoque de transformación conocido como Pruned Problem Transformation (PPT) [22], una variante de conjunto potencia que realiza una poda de clases para reducir el gran número de combinaciones posibles de etiquetas. La selección basada en información mutua se realizó con un método greedy hacia adelante y se evaluó en tres conjuntos de datos, en los que se observó que la performance de los clasificadores mejoró respecto de los mismos sin selección. En un enfoque diferente, Lee et al. [13] propusieron un método de selección basado en información mutua multivariada que maximiza la dependencia entre las características y las etiquetas sin recurrir a la transformación del problema. Los experimentos en tres conjuntos de datos mostraron que este método es superior a la selección con métricas $\chi^2$ o Mutual Information en algoritmos de transformación como el mencionado PPT. Los estudios mencionados se han limitado a comparar el efecto de la selección en uno o dos métodos de transformación solamente, mientras que el presente trabajo buscó evaluar su

impacto en un espectro más amplio de problemas de transformación, considerando algoritmos clásicos en sus tres ramas principales, además de incluir una variedad de conjuntos de datos y variantes de selección de características basadas en la evaluación por subconjuntos.

## VII. CONCLUSIONES

La clasificación multi-etiqueta ha adquirido un gran atractivo en los últimos años debido a sus potenciales aplicaciones, que van desde la clasificación de texto y recursos Web de distinto tipo (como imágenes o vídeos), hasta aplicaciones en bioinformática. Este tipo de clasificación es una forma clásica de categorización de documentos, ya que estos usualmente pertenecen a múltiples tópicos [17]. En los últimos años se ha convertido en un método común para la recomendación de etiquetas en sistemas de anotación colaborativos [12, 25, 14]. En sitios como Delicious o Fickr los recursos, documentos e imágenes, tienen asociadas varias etiquetas o tags que son agregados de manera colectiva por una comunidad de usuarios. El aprendizaje de estas asociaciones se convierte entonces en un problema de clasificación multi-etiqueta a gran escala. En bioinformática se ha usado MLC en distintas tareas de predicción [29, 10, 2].

Dada la alta dimensionalidad que típicamente presentan los datos en esas aplicaciones, el uso de técnicas de selección de características se vuelve indispensable y, por lo tanto, es importante determinar el impacto de las mismas en la performance de los clasificadores. Este trabajo presentó un análisis experimental en este sentido, evaluando la performance de múltiples clasificadores dentro de los distintos enfoques de transformación con una variedad de conjuntos de datos. La selección se basó en las técnicas por evaluación de subconjunto que permiten eliminar no solo características irrelevantes sino también redundantes. Los resultados mostraron que es posible una gran reducción de la dimensionalidad, eliminando más de un 70% de las características en todos los conjuntos, sin degradar significativamente la performance de la clasificación.

## AGRADECIMIENTOS

Agradecemos el soporte financiero de ANPCyT brindado a través de los proyectos PICT-2012-0045 y PICT-2013-0464, el soporte financiero de CONICET brindado a través del proyecto PIP 11220120100185CO.

## REFERENCIAS

[1] M. Boutell, J. Luo and X. Shen and C. Brown, "Learning multi-label scene classification", Pattern Recognition (2004), 1757--1771.
[2] R. Cerri, R. R. O. da Silva and A. C. P. L. F. de Carvalho, "Comparing Methods for Multilabel Classification of Proteins Using Machine Learning Techniques, Lecture Notes in Computer Science (2009), vol. 5676, 109--120.
[3] A. Clare and R. D. King, "Knowledge Discovery in Multi-label Phenotype Data", Principles of Data Mining and Knowledge Discovery, Springer-Verlag (2001), 42--53.
[4] G. Doquire and M. Verleysen, "Feature Selection for Multi-label Classification Problems", Lecture Notes in Computer Science, vol. 6691, Springer (2011), 9--16.
[5] G. Forman, "An Extensive Empirical Study of Feature Selection Metrics for Text Classification", Journal of Machine Learning Research (2003), 1289--1305.
[6] J. Fürnkranz, "Round Robin Classification", Journal of Machine Learning Research (2002), 721--747.
[7] J. Fürnkranz, E. Hüllermeier, E. Loza Mencía and K. Brinker, "Multilabel classification via calibrated label ranking", Machine Learning (2008), 133--153.
[8] M. A. Hall, "Correlation-based Feature Subset Selection for Machine Learning" (1998).
[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, "The WEKA data mining software: An update", ACM SIGKDD Explorations Newsletter (2009), 10--18.
[10] D. Heider, R. Senge, W. Cheng and E. Hüllermeier, "Multilabel classification for exploiting cross-resistance information in HIV-1 drug resistance prediction", Bioinformatics (2013), 1946--1952.
[11] G. John, R. Kohavi and K. Pfleger, "Irrelevant Features and the Subset Selection Problem" Machine learning: proceedings of the eleventh international conference (1994), 121--129.
[12] I. Katakis, G. Tsoumakas and I. Vlahavas, "Multilabel text classification for automated tag suggestion" ECML PKDD discovery challenge 75 (2008).
[13] J. Lee and D. Kim, "Feature Selection for Multi-label Classification Using Multivariate Mutual Information", Pattern Recognition Letters (2013), 349--357.
[14] L. Li, Y. Yao, F. Xu and J. Lu, "MATAR: Keywords Enhanced Multi-label Learning for Tag Recommendation" Asia-Pacific Web Conference. Springer International Publishing (2015), 268--279.
[15] O. Luaces, J. Díez, J. Barranquero, J. J. del Coz and A. Bahamonde, "Binary relevance efficacy for multilabel classification", Progress in Artificial Intelligence (2012), 303--313.
[16] G. Madjarov, D. Kocev, D. Gjorgjevikj and S. Džeroski, "An extensive experimental comparison of methods for multi-label learning", Pattern Recognition (2012), 3084--3104.
[17] A. K. McCallum, "Multi-label text classification with a mixture model trained by EM" AAAI'99 workshop on text learning. (1999), 1--7.
[18] E. Loza Mencía, S. Park and J. Fürnkranz, "Efficient voting prediction for pairwise multilabel classification", Neurocomputing (2010), 1164-1176.
[19] L. C. Molina, L. Belanche and A. Nebot, "Feature selection algorithms: A survey and experimental evaluation" IEEE International Conference on Data Mining (2002), 306--313.
[20] S. Park, Johannes Fürnkranz, "Efficient Pairwise Classification" European Conference on Machine Learning (2007).
[21] J. Platt, "Fast Training of Support Vector Machines using Sequential Minimal Optimization", MIT Press (1998).
[22] J. Read, "A Pruned Problem Transformation Mathod for Multi-label Classification", New Zealand Computer Science Research Student Conference (2008), 143--150.
[23] J. Read, B. Pfahringer, G. Holmes and Eibe Frank, "Classifier Chains for Multi-label Classification" (2009).
[24] J. M. Rodriguez, D. Godoy, C. Mateos and A. Zunino, "A multi-core computing approach for large-scale multi-label classification", Intelligent Data Analysis (2016).
[25] C. Shen, J. Jiao, Y. Yang and B. Wang, "Multi-instance multi-label learning for automatic tag recommendation" IEEE International Conference on Systems, Man and Cybernetics (2009), 4910--4914.
[26] N. Spolaôr, E. Alvares Cherman, M. C. Monard and H. D. Lee, "A Comparison of Multi-label Feature Selection Methods using the Problem Transformation Approach", Electronic Notes in Theoretical Computer Science (2013), 135--151.
[27] G. Tsoumakas, I. Katakis and I. Vlahavas, "Effective and efficient multilabel classification in domains with large number of labels" Proc. ECML/PKDD 2008 Workshop on Mining Multidimensional Data (2008).
[28] G. Tsoumakas, E. Spyromitros-Xioufis, J. Vilcek and I. Vlahavas, "MULAN: A Java Library for Multi-Label Learning", Journal of Machine Learning Research (2011), 2411--2414.
[29] P. Vateekul, M. Kubat and K. Sarinnapakorn, "Hierarchical multi-label classification with SVMs: A case study in gene function prediction", Intelligent Data Analysis (2014).
[30] T. Wu and C. Lin and R. C. Weng, "Probability Estimates for Multi-class Classification by Pairwise Coupling", Journal of Machine Learning Research (2004), 975--1005.
[31] Y. Yang and J. O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization" (1997), 412--420.
[32] M. Zhang and Z. Zhou, "ML-KNN: A lazy learning approach to multi-label learning", Pattern Recognition (2007), 2038--2048.

[33] M. Zhang and Z. Zhou, "A Review on Multi-Label Learning Algorithms", IEEE Transactions on Knowledge and Data Engineering (2013), 1819--1837.

**Juan Manuel Rodriguez** holds a PhD degree in computer science from UNICEN. He is a teaching assistant at UNICEN. He also works a researcher at CONICET and is member of the Instituto Superior de Ingeniería de Software Tandil (ISISTAN). His research interests include Web services, mobile devices, grid computing, and service-oriented grids.

**Daniela Godoy** received his Ph.D. degree in computer science from UNICEN University in 2005. She is a full-time professor in the Computer Science Department at UNICEN, member of ISISTAN Research Institute and researcher at CONICET. Her research interests include intelligent agents, user profiling and text mining.

**Alejandro Zunino** has a PhD in computer science from UNICEN. He is an adjunct professor at UNICEN and member of ISISTAN and CONICET. His research areas include grid computing, service-oriented computing, Semantic Web services, and mobile computing.