09-12 Nov

CAEPIA'15

Albacete

Libro de Actas

Albacete, 9-12 Noviembre 2015

1. Indice

CAEPIA - Sesión General

mldr: Paquete R para Exploración de Datos Multietiqueta

David Charte and Francisco Charte

Universidad de Granada, Granada, España
fdavidcl@correo.ugr.es, fcharte@ugr.es

Resumen La clasificación tradicional de datos trabaja con varios atributos de entrada y uno solo de salida. La clasificación multietiqueta, por el contrario, implica predecir simultáneamente más de un atributo de salida, por lo que los algoritmos de clasificación han de afrontar una tarea considerablemente más compleja. La clasificación de datos multietiqueta es una labor de creciente importancia, presente en ámbitos como la categorización de textos, el etiquetado de imágenes o la identificación de sonidos en un medio.
mldr es un paquete R que incorpora una serie de herramientas para cargar y crear datasets multietiqueta, calcular diversas medidas de caracterización de los datos, generar varios gráficos que muestran su distribución, y manipular los datasets para tratar de facilitar la tarea de los algoritmos de preprocesamiento y de clasificación. Además, la interfaz gráfica de usuario incluida permite que usuarios sin experiencia con R accedan a las mismas funcionalidades.

Keywords: Clasificación multietiqueta · R · Análisis exploratorio · Software

1. Introducción

La clasificación de datos [1] es una tarea ampliamente requerida en la actualidad, ya que representa problemas obtenidos de muy diversas fuentes: detección de spam en mensajes de correo, diagnóstico de enfermedades y concesión de créditos, entre otros ámbitos. Los casos más estudiados se plasman en conjuntos de datos de tipo binario o multiclase, es decir, para cada instancia se predice un único valor de salida: uno de entre dos en el caso binario, o uno de entre más de dos en el caso multiclase.

En los últimos años se ha producido un crecimiento de un tipo de información que, por su naturaleza, no es posible tratar con clasificadores tradicionales. Se trata de la información multietiqueta [2], representada por fotografías de cámaras y teléfonos, publicaciones en blogs y redes sociales, contenido audiovisual en plataformas de *streaming*, etc. Todos estos son ejemplos de datos que se pueden clasificar a la vez en varias categorías no excluyentes: por ejemplo, una fotografía puede ser del mar, la playa y una puesta de sol; una publicación puede ser una pieza de opinión sobre política y economía, etc.

En su representación habitual, todos estos conjuntos de datos se almacenan mediante estructuras similares a tablas, donde cada columna es un atributo y cada fila corresponde a una instancia. En el caso de los datasets binarios y multiclase, únicamente es necesario reservar una columna para contener la información de clase, y generalmente se asumirá que esta es la última de la tabla. Esto no ocurre así para los datasets multietiqueta (MLDs), ya que se requiere de varias de las columnas, cada una representando una etiqueta, para abarcar toda la información a predecir.

R [3] es una conocida herramienta dirigida a tareas estadísticas que provee una estructura de datos capaz de contener tablas de estos tipos: el `data.frame`. Esta estructura basta para representar problemas de tipo binario y multiclase, pero para datos multietiqueta es necesaria otra adicional que identifique las columnas que corresponden a etiquetas. Las funcionalidades disponibles para R se pueden ampliar mediante la instalación de paquetes, que están disponibles para su descarga e instalación desde el propio entorno de trabajo de dicha herramienta. El paquete mldr introduce la funcionalidad básica para la exploración y el análisis de los MLDs en R, y proporciona una base para la programación de algoritmos de clasificación. Se trata además del primer paquete que permite trabajar de forma nativa y genérica con MLDs en R.

Hasta ahora, la mayoría del software existente para trabajar con MLDs está escrito en Java, y las dos herramientas más conocidas para esta tarea son MULAN y [4] MEKA [5], que dependen de WEKA [6]. Pese a que el paquete RWeka [7] proporciona una conexión con WEKA desde R, no facilita el manejo de los MLDs. Además, a diferencia de MULAN y MEKA, el hecho de que mldr esté implementado en R facilita el tratamiento estadístico de los datos y la generación de gráficos a partir de ellos, aportando asimismo una interfaz de usuario web que permite el acceso a las funcionalidades del paquete.

La siguiente sección introduce las características principales de los MLDs. A continuación, la Sección 3 presenta el paquete mldr y su funcionalidad. Finalmente, la Sección 4 expone las conclusiones.

2. Clasificación Multietiqueta

El problema de clasificación pertenece al aprendizaje supervisado, y se presenta mediante un conjunto de instancias o dataset, pertenecientes a un espacio de características que pueden ser de distinto tipo. Estas instancias vienen asociadas a su información de clase, y un algoritmo de clasificación pretende ser capaz de predecir dicha información para nuevas instancias sin clasificar.

Frente a la clasificación tradicional, ya sea binaria o multiclase, la clasificación multietiqueta (MLC) [8] presenta una diferencia fundamental en cuanto a la información a predecir: si L es el conjunto de clases o etiquetas disponibles, entonces cada instancia en un problema binario o multiclase estará asociado a un único $y \in L$, mientras que en MLC se pretende predecir un subconjunto $Y \in \mathcal{P}(L)$. Esto implica que en el primer caso se escogerá una de entre $|L|$ posi-

bilidades ($|L| = 2$ en clasificación binaria, y $|L| > 2$ en clasificación multiclase), y en el último de entre $2^{|L|}$ posibles elecciones.

En algunos casos el número de etiquetas de un MLD es de varios cientos o incluso miles, siendo en ocasiones mayor que el número de atributos de entrada. Otro escenario común es que se produzca un desequilibrio en la aparición de las etiquetas: mientras que unas etiquetas están presentes en gran parte de las instancias, otras podrían aparecer en muchas menos ocasiones. Las etiquetas frecuentes y las menos frecuentes pueden aparecen a la vez en algunas instancias, un caso que es interesante estudiar.

Los algoritmos de MLC se suelen basar en uno de dos enfoques [2]: adaptación de algoritmos, mediante la cual se toma un algoritmo o una técnica conocida y se altera para ajustarse al problema de multietiqueta, o transformación de los datos, de forma que se puedan tratar con algoritmos de clasificación binaria [9] o multiclase [10] y después combinar los resultados obtenidos.

2.1. Características de los MLDs

Para poder analizar las diversas situaciones que se pueden encontrar en un problema de MLC, se han de estudiar medidas específicas a este tipo de datasets [8], que dan una idea del comportamiento de los datos en el mismo.

Sea $D \subset X^1 \times X^2 \times \ldots X^f \times \mathcal{P}(L)$ un conjunto finito de instancias del espacio de características. Cada instancia, por tanto, es $(X, Y) \in D$ con $X \in \prod_{i=1}^{f} X^i$, $Y \subset L$. La información más básica que se puede obtener son los cardinales de algunos conjuntos a partir de esta definición: la cuenta de atributos f, el número de instancias $|D|$, el número de etiquetas existentes $|L|$, y el número de combinaciones distintas de etiquetas que aparecen en el MLD $|\{Y : (X, Y) \in D\}|$.

Algunas medidas que ofrecen una idea acerca de la distribución de etiquetas en el MLD son el número medio de etiquetas activas por instancia, *Card*, y la relativización de esta medida respecto al número total de etiquetas, *Dens* (1).

$$Card = \frac{1}{|D|} \sum_{(X,Y)\in D} |Y|, \; Dens = \frac{1}{|D|} \sum_{(X,Y)\in D} \frac{|Y|}{|L|} \; . \tag{1}$$

Para ahondar en más detalle acerca del nivel de desequilibrio en MLDs, se han definido distintas métricas en [11]. El *IRLbl* (2) es un cálculo para una etiqueta que indica su nivel de desequilibrio, es decir, lo lejos que está una etiqueta de ser la más frecuente.

$$IRLbl(y) = \frac{\max\limits_{y' \in L} \left(\sum\limits_{(X,Y)\in D} h(y', Y) \right)}{\sum\limits_{(X,Y)\in D} h(y, Y)}, \; h(y, Y) = \begin{cases} 1 & y \in Y \\ 0 & y \notin Y \end{cases} \; . \tag{2}$$

A partir de esta medida se definen otras que se refieren al MLD en general, como es la media de los *IRLbl*, *MeanIR*, acompañada del coeficiente de variación

media correspondiente, *CVIR*, dados por su definición (3).

$$MeanIR = \frac{1}{|L|}\sum_{y\in L} IRLbl(y),\ \ CVIR = \frac{IRLbl\sigma}{MeanIR}\ \ , \tag{3}$$

donde $IRLbl\sigma$ es la desviación típica muestral corregida de los valores de *IRLbl* para las etiquetas de L.

Por último, existen diversas medidas para tratar de cuantificar las relaciones entre etiquetas. Estas suelen estar relacionadas con las combinaciones de etiquetas o labelsets existentes. Por ejemplo, el número de labelsets que cuentan con una sola ocurrencia en el MLD, o el número de labelsets distintos que se dan. Una medida más avanzada es *SCUMBLE*, definida en [12] como se muestra en la expresión (4).

$$SCUMBLE = \frac{1}{|D|}\sum_{i=1}^{|D|}\left(1 - \frac{1}{\overline{IRLbl_i}}\left(\prod_{y\in L} IRLbl_{iy}\right)^{(1/|L|)}\right)\ \ , \tag{4}$$

donde $IRLbl_{iy} = IRLbl(y)$ si y es una etiqueta activa en la i-ésima instancia, y $IRLbl_{iy} = 0$ en otro caso; $\overline{IRLbl_i}$ es el nivel promedio de desequilibrio para las etiquetas activas en la i-ésima instancia.

Todas estas medidas son útiles para obtener información acerca de la distribución de los datos en el conjunto de instancias y del posible comportamiento de un algoritmo de clasificación o el nivel de éxito de una técnica de preprocesamiento. Sin embargo, es deseable una herramienta que ofrezca una visión general del MLD a partir de estos datos, y que los muestre de forma organizada.

3. El Paquete mldr

R dispone de una gran variedad de herramientas para manejo, tratamiento y visualización de datos. Sin embargo, hasta ahora no contaba con una herramienta capaz de manejar MLDs, es decir, con el soporte necesario para cargar, crear y escribir datasets de este tipo, identificar las etiquetas y proporcionar el cálculo de medidas útiles para la exploración de los datos.

El paquete mldr pretende ocupar ese lugar, proporcionando un conjunto de funciones que habilitan la lectura y escritura de MLDs mediante archivos de tipo ARFF, la creación de estructuras de datos de clase `"mldr"`, el cálculo de las principales medidas de exploración y la generación de gráficos para la visualización de las características del MLD.

Frente a las herramientas ya existentes para el tratamiento de MLDs, como MULAN o MEKA, mldr facilita el acceso a los datos de forma interactiva desde la línea de comandos de R, soporta ambos formatos de archivo e incluso permite crear nuevos MLDs desde otras estructuras de datos como los `data.frame`, sin necesidad de cargarlos de un archivo. Gracias a las amplias funcionalidades estadísticas de R, la exploración y recopilación de medidas de los MLDs es más extensa, y la generación de gráficos más cómoda y ajustable. En suma, mldr ofrece mayor funcionalidad de análisis exploratorio que MULAN y MEKA.

3.1. Instalación

R es una plataforma que cuenta con un intérprete interactivo disponible para los principales sistemas operativos. A través del propio intérprete se pueden descargar e instalar los paquetes disponibles en CRAN mediante un comando sencillo:

```
> install.packages("mldr")
```

Una vez realizado este paso, es necesario ejecutar el comando de carga del paquete para disponer de su funcionalidad:

```
> library(mldr)
```

Al ejecutar esa línea, el intérprete importará todas las funciones que mldr permite utilizar y, además, pondrá a disposición del usuario tres MLDs de ejemplo: birds [13], emotions [14] y genbase [15].

Una llamada a la función `mldrGUI()` lanzará la interfaz de usuario web que se incorpora junto con el paquete, abierta en una nueva pestaña de navegador. Esta interfaz gráfica da acceso a gran parte de la funcionalidad del paquete, centrándose en la exploración y la visualización de las características de los MLDs, sin necesidad de generar estadísticos y gráficos mediante código.

3.2. Lectura y Escritura de MLDs

La función `mldr` es la encargada de construir las estructuras de datos con las que manejar MLDs a partir de archivos ARFF. Estos archivos son leídos en forma de tablas, y son necesarios datos adicionales que indiquen cuáles de las columnas corresponden a etiquetas. Para ello existen diversas soluciones, siendo las más comunes las soportadas por MULAN y MEKA.

La primera de ellas es aportar un archivo XML indicando los nombres de las columnas que son etiquetas, mientras que la segunda es utilizar una cabecera especial en la línea del archivo que indica el nombre del MLD, junto a un parámetro que especifica el número de atributos que son de salida. Ambas posibilidades son válidas para mldr, basta con indicar el parámetro adecuado al llamar a la función:

```
> emotions <- mldr("emotions") # Usando XML tipo MULAN
> enron <- mldr("ENRON-F", use_xml = FALSE) # Usando cabecera MEKA
```

Otros parámetros permiten la lectura de MLDs que no cuenten con archivo XML ni cabecera especial, necesitando por tanto el número de columnas que corresponden a etiquetas, la lista de índices de etiquetas o bien los nombres de cada una de ellas.

Desde la interfaz gráfica es sencillo cargar un MLD: se utilizan los botones de subida de archivos en la pestaña Main para transferir a la aplicación el archivo ARFF y el XML correspondientes, como se observa en la Fig. 1.

Además de leer MLDs de un archivo, mldr también permite crearlos a partir de un `data.frame` de R y escribirlos de nuevo a un fichero ARFF. Para ello

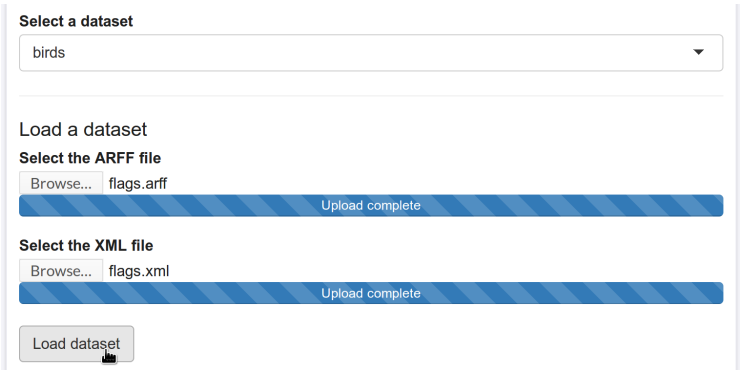

Figura 1. Carga de un MLD desde la interfaz de mldr

cuenta con las funciones `mldr_from_dataframe` y `write_arff`, respectivamente. Por ejemplo, las siguientes líneas de código generarían un MLD a partir de datos obtenidos de una distribución normal, y después lo guardarían a un archivo:

```
> ejemplo <- data.frame(matrix(rnorm(1000), ncol = 10))
> ejemplo$label1 <- c(sample(c(0,1), 100, replace = TRUE))
> ejemplo$label2 <- c(sample(c(0,1), 100, replace = TRUE))
> mld <- mldr_from_dataframe(df, labelIndices = c(11, 12))
> write_arff(mld, "ejemplo_mld", write.xml = TRUE)
```

3.3. Medidas Exploratorias

Como se ha descrito en la Sección 2.1, las medidas diseñadas para MLDs permiten dar ideas acerca de distintos aspectos que caracterizan los datos, tanto las instancias como las etiquetas. El paquete calcula estas medidas para el usuario y las almacena en el miembro `measures` del objeto de clase `"mldr"`, pudiendo consultarse directamente o mediante la función `summary`. Las medidas específicas para etiquetas estarán disponibles en el miembro `labels` del objeto, que además incluirá los nombres de las etiquetas y sus índices en el MLD. Asimismo, mldr también incorpora en el miembro `labelsets` un vector con la cuenta de ocurrencias de cada labelset del MLD:

```
> summary(emotions)
  num.attributes num.instances num.labels num.labelsets
              78           593          6            27
  num.single.labelsets max.frequency cardinality   density
                     4            81    1.868465 0.3114109
    meanIR    scumble
  1.478068 0.01095238

> emotions$labels
```

```
                 index count      freq     IRLbl      SCUMBLE
amazed-suprised     73   173 0.2917369 1.526012 0.002159173
happy-pleased       74   166 0.2799325 1.590361 0.014332319
relaxing-calm       75   264 0.4451939 1.000000 0.023786461
quiet-still         76   148 0.2495784 1.783784 0.023131538
sad-lonely          77   168 0.2833052 1.571429 0.016133470
angry-aggresive     78   189 0.3187184 1.396825 0.001331189

> emotions$labelsets
000111 001101 010010 010100 101000 001001 001011 100011 000100
     1      1      1      1      2      3      3      4      5
010001 011100 100010 110001 111000 000010 000011 010000 100000
     5      6      6      7     11     12     12     23     24
001010 001100 000110 110000 001000 001110 000001 011000 100001
    25     30     37     38     42     67     72     74     81
```

En la pestaña Main de la interfaz de usuario se muestran datos generales sobre el MLD junto a algunos estadísticos básicos obtenidos a partir de las medidas para etiquetas y labelsets. Las pestañas Labels y Labelsets contienen tablas con datos específicos, similares a la de la Fig. 2.

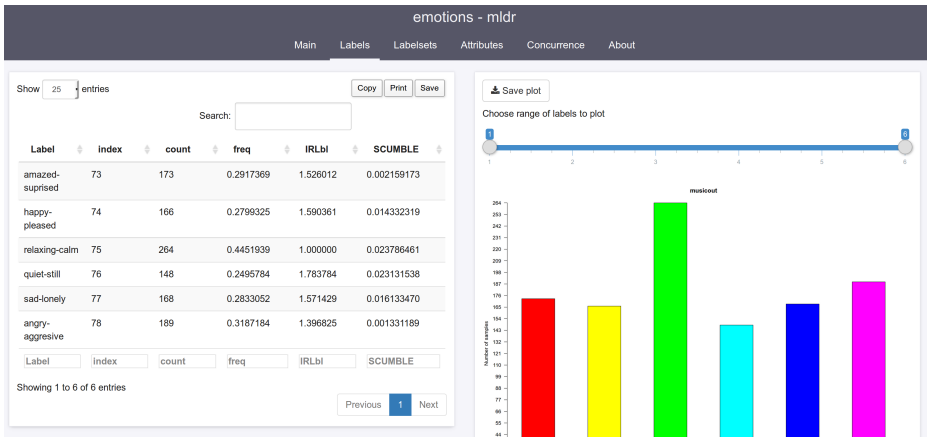

Figura 2. Consulta de información de las etiquetas en la interfaz gráfica

3.4. Generación de Gráficos

El paquete mldr incorpora una versión personalizada del método `plot` para la clase `"mldr"`. Esta permite al usuario entregar como parámetro un objeto de esta clase junto con ajustes para generar gráficos personalizados. La función es capaz de generar siete tipos distintos de gráficos, en concreto tres tipos de

histogramas para representar distintas relaciones entre los datos y las etiquetas, dos gráficos de barras con propósito similar, un gráfico de sectores que detalla los tipos de atributo y un último gráfico que muestra la concurrencia entre etiquetas.

Para generar cualquiera de estos gráficos el prodecimiento a seguir será muy similar: basta llamar a la función `plot` con el parámetro `type` adecuado:

```
> plot(emotions) # Por defecto: gráfico de concurrencia
> plot(birds, type = "LH") # Histograma de etiquetas
```

El histograma de etiquetas (tipo `"LH"`) muestra el número de etiquetas según el número de instancias en las que están activas. Esto permite visualizar la dispersión de las etiquetas en las instancias: si el gráfico se acumula a la izquierda, entonces la mayoría de etiquetas aparecen en pocas instancias y existe una gran dispersión, mientras que en el caso contrario, el gráfico se acumulará a la derecha.

El histograma de labelsets (tipo `"LSH"`) tiene un significado similar pero orientado a labelsets, es decir, da una idea de la concentración de los labelsets en las instancias: si hay muchos labelsets que se repiten o por el contrario la mayoría de ellos aparecen en muy pocas instancias.

El tercer histograma es el de cardinalidad (tipo `"CH"`), y muestra el número de instancias según su cardinalidad, es decir, una acumulación a la izquierda indicará que una gran cantidad de instancias tienen pocas etiquetas, y una acumulación a la derecha señalará que muchas instancias tienen muchas etiquetas.

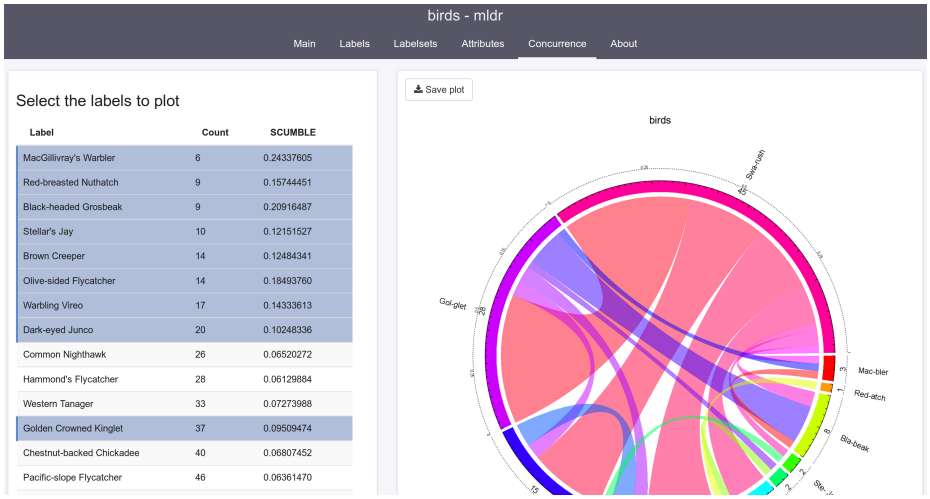

Figura 3. Ajuste del gráfico de concurrencia de etiquetas

Los gráficos de barras para etiquetas (tipo `"LB"`) y para labelsets (tipo `"LSB"`) tienen una función simple: para cada etiqueta y para cada labelset indican en cuántas instancias están presentes. El gráfico de sectores sobre tipos de atributos (tipo `"AT"`) indica los tipos y la cantidad de atributos de cada tipo.

Por último, el gráfico de concurrencia entre etiquetas (tipo `"LC"`) da idea de las coocurrencias que se dan en las instancias entre distintas etiquetas. Este gráfico tiene forma circular dividida en arcos, donde cada arco representa una etiqueta. El ancho de estos es proporcional al número de instancias en que aparece la etiqueta correspondiente. De estos arcos parten distintas bandas, cuya anchura es proporcional al número de instancias en que las dos etiquetas conectadas aparecen juntas. La concurrencia entre etiquetas es un aspecto importante en el estudio de los MLDs, ya que puede determinar el éxito de una técnica de preprocesamiento [12].

Todos estos gráficos pueden ser visualizados y ajustados desde la interfaz gráfica en las pestañas correspondientes. La pestaña Main contiene los tres histogramas y el gráfico de sectores, mientras que la pestañas Labels y Labelsets muestran los gráficos de barras asociados. El gráfico de concurrencia se encuentra en la pestaña Concurrence, donde se pueden restringir la cantidad de etiquetas a considerar para generar de nuevo el gráfico, como se observa en la Fig. 3.

3.5. Transformación y Filtrado

Las instancias en un objeto de clase `"mldr"` se pueden filtrar fácilmente mediante el operador `[` de R. Por ejemplo, la siguiente línea de código crearía en la variable `happy.music` un nuevo MLD con las instancias de emotions en las que la etiqueta *happy-pleased* está activa:

```
> happy.music <- emotions[emotions$dataset$"happy-pleased" == 1]
```

Asimismo, la función `mldr_transform` proporciona la funcionalidad necesaria para aplicar las transformaciones Binary Relevance [9] y Label Powerset [10] utilizadas para MLDs:

```
> emotionsbr <- mldr_transform(emotions, type = "BR")
> emotionslp <- mldr_transform(emotions, type = "LP")
```

Estas funcionalidades pueden ser de gran ayuda para implementar nuevos algoritmos de clasificación de forma nativa en R, ya que varios de los algoritmos ya existentes se basan en las transformaciones antes mencionadas.

4. Comentarios Finales

En este artículo se ha presentado una herramienta novedosa para el tratamiento de MLDs, el paquete mldr para R. La MLC difiere fuertemente de las clasificaciones binaria y multiclase en varios aspectos, que pueden provocar ciertas situaciones que conviene estudiar, como el desequilibrio de las etiquetas y la concurrencia entre estas. El paquete presentado ayuda a explorar las características de los MLDs y a visualizar estas situaciones, permitiendo además realizar transformaciones conocidas de los datos. La interfaz gráfica incorporada facilita el acceso a la funcionalidad a cualquier usuario.

El paquete mldr puede ya servir de base para la implementación de nuevos algoritmos de clasificación, y en futuras versiones se ampliará la funcionalidad para que facilite aún más la tarea de otros desarrolladores al incorporar sus algoritmos al lenguaje R.

Agradecimientos: Este trabajo es parcialmente financiado por el Ministerio de Educación bajo el proyecto TIN2012-33856 y los proyectos regionales de Andalucía P10-TIC-06858 y P11-TIC-9704.

Referencias

1. Kotsiantis, S.B.: Supervised machine learning: A review of classification techniques (2007)
2. Gibaja, E., Ventura, S.: Multi-label learning: a review of the state of the art and ongoing research. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 4(6) (2014) 411–444
3. Ihaka, R., Gentleman, R.: R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics **5**(3) (1996) 299–314
4. Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J., Vlahavas, I.: MULAN: A Java Library for Multi-Label Learning. Journal of Machine Learning Research 12 (2011) 2411–2414
5. Read, J., Reutemann, P.: MEKA: A Multi-label Extension to WEKA
6. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann (2005)
7. Hornik, K., Buchta, C., Zeileis, A.: Open-source machine learning: R meets Weka. Computational Statistics 24(2) (2009) 225–232
8. Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining Multi-label Data. In Maimon, O., Rokach, L., eds.: Data Mining and Knowledge Discovery Handbook. Springer US, Boston, MA (2010) 667–685
9. Godbole, S., Sarawagi, S.: Discriminative Methods for Multi-Labeled Classification. In: Advances in Knowledge Discovery and Data Mining. Volume 3056. (2004) 22–30
10. Boutell, M., Luo, J., Shen, X., Brown, C.: Learning multi-label scene classification. Pattern Recognition 37(9) (2004) 1757–1771
11. Charte, F., Rivera, A.J., del Jesus, M.J., Herrera, F.: Addressing imbalance in multilabel classification: Measures and random resampling algorithms. Neurocomputing 163(0) (2015) 3–16
12. Charte, F., Rivera, A., Jesus, M.J., Herrera, F.: Concurrence among imbalanced labels and its influence on multilabel resampling algorithms. In: Proc. 9th International Conference on Hybrid Artificial Intelligent Systems, Salamanca, Spain, HAIS'14. Volume 8480 of LNCS. (2014)
13. Briggs, F., Huang, Y., Raich, R., Eftaxias, K., et al.: The 9th annual mlsp competition: New methods for acoustic classification of multiple simultaneous bird species in a noisy environment. In: Machine Learning for Signal Processing (MLSP), 2013 IEEE International Workshop on. (Sept 2013) 1–8
14. Trohidis, K., Tsoumakas, G., Kalliris, G., Vlahavas, I.P.: Multi-label classification of music into emotions. In: ISMIR. Volume 8. (2008) 325–330
15. Diplaris, S., Tsoumakas, G., Mitkas, P., Vlahavas, I.: Protein Classification with Multiple Algorithms. In: Proc. 10th Panhellenic Conference on Informatics, Volos, Greece, PCI'05. (2005) 448–456