Una herramienta para analizar conjuntos de datos multi-etiqueta

Jose M. Moyano, Eva L. Gibaja, and Sebastián Ventura

Universidad de Córdoba. Departamento de Informática y Análisis Numérico

Resumen El aprendizaje multi-etiqueta, a diferencia del aprendizaje clásico, permite que cada patrón del conjunto de datos tenga asociadas simultáneamente varias clases o etiquetas. Los conjuntos de datos multi-etiqueta presentan características como desbalanceo, gran dimensionalidad o dependencias entre etiquetas, que pueden tener un gran efecto en el rendimiento de los algoritmos desarrollados para resolver este tipo de problemas. El objetivo principal de nuestro trabajo es presentar una herramienta con interfaz gráfica e intuitiva para la exploración y análisis de datos multi-etiqueta, que permita extraer y guardar las características de los *datasets*. Por otro lado, permite extraer características de varios *datasets* simultáneamente, incluye tareas de preprocesado, como particionado, selección de características e instancias, y conversión de *datasets* entre los principales formatos de datos para aprendizaje multi-etiqueta.

Keywords: Aprendizaje multi-etiqueta, software, exploración de datos, preprocesado

1. Introducción

El aprendizaje multi-etiqueta (*Multi-Label Learning*, MLL) es una tarea en la cual, a diferencia del aprendizaje clásico, un patrón puede tener asociadas varias clases (etiquetas) simultáneamente. Este tipo de aprendizaje ha sido aplicado con éxito en problemas de gran actualidad como clasificación de datos multimedia [9], diagnóstico médico [11] o *marketing* directo [15]. Por ejemplo, en *marketing* directo, a un cliente potencial se le pueden sugerir diversos productos que le puedan parecer interesantes.

La representación más extendida de los *datasets* multi-etiqueta es la que se puede apreciar en la Figura 1. En ella se puede observar como, en lugar de una única clase, aparece un conjunto de clases o etiquetas binarias, cuyo valor será de 1 o 0 indicando si el patrón tiene asociada o no dicha etiqueta.

Actas de la XVII Conferencia de la Asociación Española para la Inteligencia Artificial, pp. 857-866

#	Características	Clase
1	$\overline{x_1}$	0
2	$\overline{x_2}$	1
3	$\overline{x_3}$	0
4	$\overline{x_4}$	0
5	$\overline{x_5}$	1

(a) *Dataset* tradicional

#	Características	λ_1	λ_2	λ_3	λ_4
1	$\overline{x_1}$	1	0	0	0
2	$\overline{x_2}$	0	1	1	0
3	$\overline{x_3}$	1	1	1	0
4	$\overline{x_4}$	1	0	0	1
5	$\overline{x_5}$	0	1	1	0

(b) *Dataset* multi-etiqueta

Figura 1. Diferencia entre conjuntos de datos tradicional y multi-etiqueta

En MLL, muchos problemas presentan características como desbalanceo (algunas etiquetas y/o combinaciones de etiquetas son muy frecuentes mientras que otras apenas están presentes), alta dimensionalidad (en número de patrones, atributos y etiquetas) o relaciones entre etiquetas [7]. Además, las características del *dataset* pueden tener efecto directo en el rendimiento de los algoritmos, de ahí la importancia de caracterizar y analizar los datos [5]. El problema de determinar qué algoritmo funciona mejor en función de las características del *dataset* ha sido abordado en [5] utilizando un enfoque de *meta-learning*, basándose en las características de los *datasets*. Existen también dos librerías para aprendizaje multi-etiqueta, Mulan [14] y Meka [3], con distintos algoritmos y con formatos de *datasets* diferentes. Por esto, también es interesante poder realizar la conversión entre ambos formatos de datos.

El principal objetivo del trabajo es presentar una herramienta gráfica para exploración y análisis de *datasets* multi-etiqueta, que permita representar de forma gráfica las características más representativas y almacenar los valores de un amplio conjunto de características para su posterior análisis. Además, la herramienta permite realizar un análisis del desbalanceo y de las relaciones entre etiquetas, aspectos considerados de gran influencia en el rendimiento de los algoritmos de aprendizaje multi-etiqueta. Por otro lado, también permitirá realizar tareas de preprocesado, como particionado o selección de atributos e instancias y conversión del formato de los datos. Además, la herramienta permitirá cargar varios *datasets* para calcular sus características de manera conjunta. Toda esta funcionalidad permitirá analizar los problemas de cara a adaptar mejor la técnica seleccionada para su solución, así como transformar y procesarlos de manera sencilla.

El resto del artículo se organiza de la siguiente manera: la Sección 2 incluye los antecedentes, la Sección 3 presenta y muestra el funcionamiento de la herramienta implementada, y por último, la Sección 4 muestra las conclusiones obtenidas y líneas de trabajo futuro.

2. Antecedentes

Para abordar de manera específica el problema de la caracterización de *datasets* multi-etiqueta, hasta donde conocemos, solamente existe la herramienta

mldr [4]. Se trata de un paquete en R para analizar datos multi-etiqueta, que ofrece tanto métodos para trabajar en R como una interfaz gráfica. Incluye funciones para el cálculo de las principales métricas de caracterización de los *datasets*, generación de gráficos de dichas características o aplicación de transformaciones como *Binary Relevance* (BR) [13] o *Label Powerset* (LP) [10]. Por otro lado, cabe citar Mulan [14] y Meka [3], dos APIs que, aún no siendo herramientas específicas para caracterización de *datasets*, ofrecen funciones para extraer algunas características básicas, como la cardinalidad, densidad o frecuencia de etiquetas. Por último, Chekina realiza en [5] un estudio para crear un conjunto de reglas que ayuden a seleccionar el mejor algoritmo de clasificación multi-etiqueta en función de las características del *dataset*. Para ello, considera un gran número de métricas para caracterización de *datasets* multi-etiqueta.

Nuestra herramienta incluye un amplio catálogo de métricas de caracterización de *datasets* multi-etiqueta (Figura 2). Las métricas se han seleccionado basándonos en la taxonomía propuesta por Charte [4]. Tomando como base esta taxonomía, hemos considerado de interés incluir las métrica básicas de Mulan y Meka, y las métricas propuestas por Chekina, categorizándolas en la taxonomía anterior y creando un nuevo grupo de métricas de atributos. Además, la herramienta añade métricas y gráficos para visualizar el desbalanceo y las relaciones entre etiquetas. Permite también cargar *datasets* en formatos de Mulan y Meka y herramientas para preprocesado de datos, como particionado y selección tanto de características como de instancias, así como obtener métricas para varios *datasets* simultáneamente.

3. Herramienta para analizar conjuntos de datos multi-etiqueta

En esta sección se va a presentar cada una de las funcionalidades de la herramienta implementada: obtención de información del *dataset*, información de las etiquetas, dependencias entre etiquetas, preprocesado o comparación de varios *datasets*. La herramienta está disponible en `http://www.uco.es/grupos/kdis/kdiswiki/MLLResources/Software/Analisis_Datasets_MLL.zip`. Se ha implementado en Java utilizando las librerías Mulan [14] y Weka [8] para la caracterización y el tratamiento de los *datasets* y las librerías JGraphX [2] y JFreeChart [1] para la representación de gráficos.

3.1. Obtener información del dataset

Al ejecutar la herramienta, se permite cargar un *dataset* multi-etiqueta, aceptando *datasets* en formato *.arff* tanto de Mulan como de Meka. Una vez se carga el *dataset*, aparece en la mitad superior de la pantalla un resumen con las métricas más comunes para la caracterización del *dataset*, como número de atributos, instancias y etiquetas, entre otras. En la mitad inferior, aparece un conjunto más amplio de métricas. El usuario puede seleccionar un conjunto de estas métricas y obtener su valor (Figura 3). Además, los resultados se pueden guardar para su posterior procesamiento en diferentes formatos, como *.txt*, *.csv*, *.arff* y *.tex*.

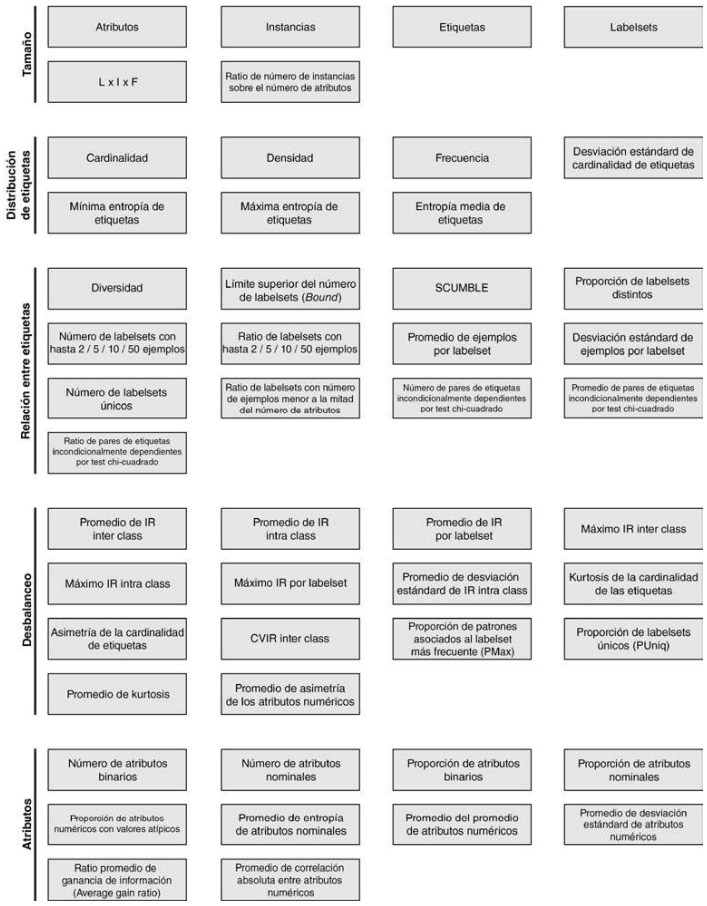


Figura 2. Taxonomía de las métricas para caracterización de *datasets* MLL

3.2. Obtener información de las etiquetas

En la pestaña *Labels* se muestran distintas métricas y gráficos sobre las etiquetas. Se permite guardar tanto los datos en formato *.csv* como las gráficas en formato *.png*. A continuación, se indican las distintas opciones que ofrece la herramienta, mostrando ejemplos para el *dataset* genbase.

- ***Label frequency***: muestra la frecuencia de aparición de cada una de las etiquetas sobre el número de instancias. En la Figura 4(a) se puede observar como existe una etiqueta con una frecuencia mayor al doble que la siguiente, mientras que por otro lado, se observa como existen hasta 10 etiquetas con una frecuencia menor al 1 %, mostrando un claro desbalanceo.
- ***Labelset frequency***: muestra la frecuencia de cada *labelset*, como se puede observar en la Figura 4(b). Igual que en el caso anterior, el *labelset* más

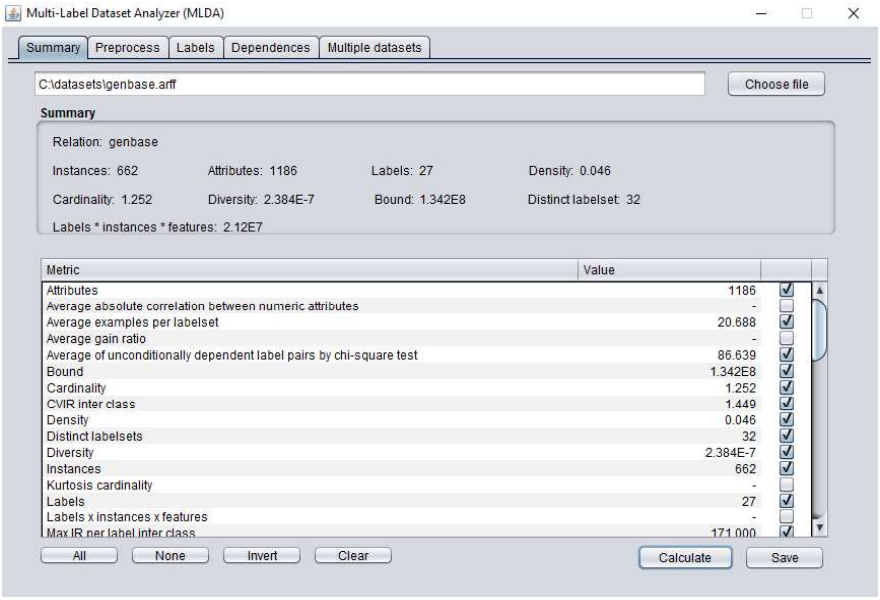
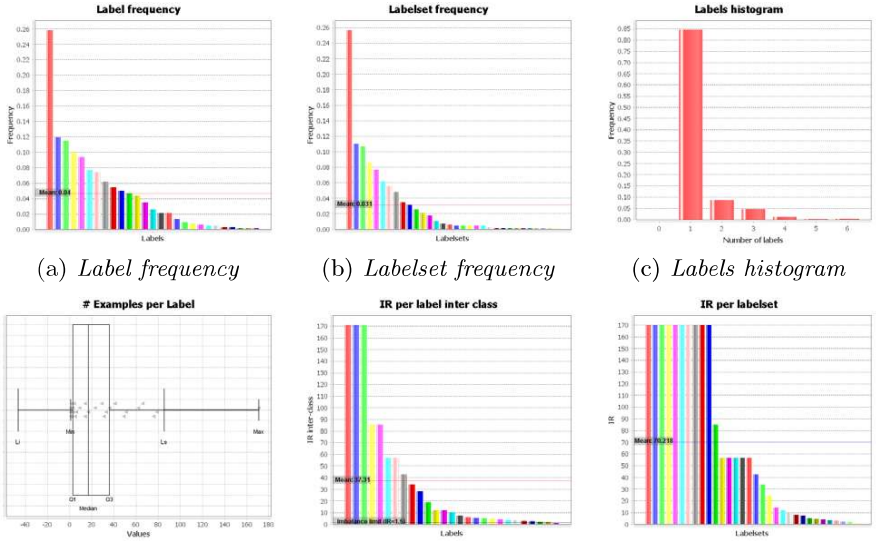


Figura 3. Pestaña principal

frecuente tiene una frecuencia mucho mayor al resto, existiendo un alto grado de desbalanceo entre *labelsets*.

- ***Labels histogram***: muestra un histograma indicando el número de etiquetas por patrón. En la Figura 4(c) se observa que casi el 85 % de los patrones tienen únicamente una etiqueta asociada, mientras que el número de patrones con más de 4 etiquetas asociadas apenas son el 2 %.
- ***Box diagram***: muestra un gráfico de tipo *boxplot* con el número de patrones por etiqueta o el número de patrones por *labelset*. Además, también muestra gráficos *boxplots* para los valores de los atributos numéricos. La Figura 4(d) muestra la distribución del número de patrones por etiqueta.
- ***IR inter class***: muestra los datos de *IR* (*Imbalance Ratio*) *inter-class* para cada etiqueta, indicando el desbalanceo existente entre etiquetas. Si el valor de IR es mayor a 1,5 se considera que dicha etiqueta está desbalanceada. Este valor se calcula dividiendo la cantidad de ejemplos positivos de la etiqueta más frecuente entre la cantidad de ejemplos positivos de la etiqueta actual. En la Figura 4(e) se observa como, excepto para la etiqueta más frecuente, los valores de IR son mayores a 1,5, por lo que todas las etiquetas están desbalanceadas respecto a la más frecuente.
- ***IR intra class***: muestra los datos de *IR intra-class* [12] para cada etiqueta. Indica el desbalanceo dentro de una etiqueta. Se calcula dividiendo el número de ejemplos positivos de dicha etiqueta entre el número de negativos.
- ***IR per labelset***: muestra los datos de *IR* para cada *labelset*. Se calcula dividiendo la cantidad de ejemplos positivos del *labelset* más frecuente entre

la cantidad de ejemplos positivos del *labelset* actual. La Figura 4(f) se puede volver a observar el alto grado de desbalanceo existente.



(a) *Label frequency* (b) *Labelset frequency* (c) *Labels histogram*

(d) *Box diagram* (e) *IR inter-class* (f) *IR per labelset*

Figura 4. Información de etiquetas del *dataset* genbase

3.3. Dependencias entre etiquetas

La herramienta ofrece la posibilidad de visualizar la dependencia entre etiquetas. Para ello incluye los coeficientes Chi y Phi de relación entre etiquetas, grafos de co-ocurrencia de etiquetas y *heatmap*. Se permite guardar tanto las tablas con los valores de cada uno de los coeficientes de dependencia en formato *.csv*, como los gráficos.

Coeficientes Chi y Phi El coeficiente Chi (χ^2) identifica la relación incondicional entre pares de etiquetas utilizando test de independencia Chi-cuadrado entre cada par de etiquetas posible. Así, si el valor de Chi entre dos etiquetas es mayor a 6,635 las etiquetas se considerarán dependientes al 99 % de confianza. Dadas dos etiquetas λ_i y λ_j y la tabla de contingencia para ambas etiquetas como se observa en la Tabla 1, el coeficiente Phi (ϕ) [6] se calcula como:

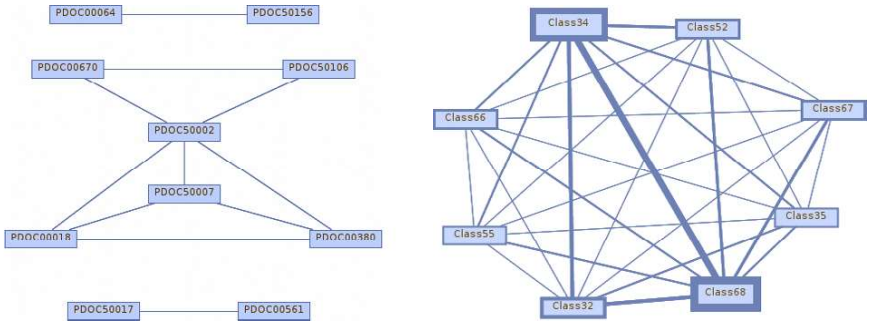
$$\phi = \frac{AD - BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}} \tag{1}$$

Los coeficientes Chi y Phi tienen una relación $\chi^2 = n \cdot \phi^2$, siendo n el número de observaciones o patrones.

Tabla 1. Tabla de contingencia para dos etiquetas

	λ_j	$\neg\lambda_j$
λ_i	A	B
$\neg\lambda_i$	C	D

Grafo de co-ocurrencia Una forma común e interesante de visualizar la dependencia entre pares de etiquetas es mediante un grafo de probabilidades de co-ocurrencia. En este tipo de grafos, el grosor del nodo representa la probabilidad *a priori* de aparición de la etiqueta $P(\lambda_i)$, y el grosor de la conexión entre dos nodos la probabilidad de co-ocurrencia $P(\lambda_i \wedge \lambda_j)$. Se puede seleccionar solamente un subconjunto de etiquetas, o seleccionar automáticamente las n etiquetas más frecuentes o más relacionadas. Además, los valores de co-ocurrencia están disponibles en forma de tabla, que se pueden exportar en formato *.csv*. La Figura 5(a) muestra un grafo de co-ocurrencia con las 10 etiquetas más relacionadas para el *dataset* genbase y la Figura 5(b) uno con las 8 etiquetas más relacionadas del *dataset* mediamill. Para genbase, se observan 3 grupos diferenciados, existiendo dos casos donde únicamente dos etiquetas co-ocurren entre sí, y un grupo de 6 etiquetas que tienen co-ocurrencias entre ellas. Por otro lado, para mediamill todas las etiquetas que se muestran co-ocurren entre sí. Destaca la co-ocurrencia entre las etiquetas *Class34* y *Class68*, siendo también las etiquetas más frecuentes.

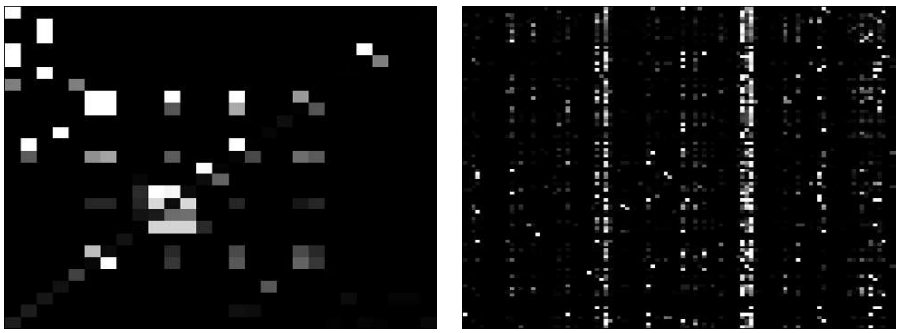


(a) genbase (b) mediamill

Figura 5. Grafos de co-ocurrencia

Heatmap Otra manera de visualizar la relación entre etiquetas son los *hetamap*. Este tipo de gráficos son más útiles que los anteriores cuando el número de etiquetas es más alto. Los *heatmap* representan una matriz $\mathbf{M}$ de dimensiones $q \times q$. Cada celda de la matriz representa la probabilidad condicional, tal que $\mathbf{M}_{ij} = P(\lambda_i|\lambda_j) = P(\lambda_i \wedge \lambda_j)/P(\lambda_j)$, donde i es una fila y j una columna. En el

gráfico, las probabilidades se muestran en escala de grises, donde el color negro será una probabilidad condicional de 0 y el color blanco será una probabilidad condicional de 1. La diagonal que va desde la esquina inferior izquierda a la esquina superior derecha presenta las probabilidades a priori, $\mathbf{M}_{jj} = P(\lambda_j)$. La Figura 6(a) muestra el *heatmap* para el *dataset* genbase y la Figura 6(b) para el *dataset* mediamill.



(a) genbase (b) mediamill

Figura 6. Heatmaps

3.4. Preprocesado y conversión

La herramienta incluye funcionalidades tanto para preprocesado de datos como para conversión a formatos Mulan y Meka (Figura 7). En cuanto al preprocesado de los datos, se puede realizar particionado de los datos tanto en *holdout* como en *k-fold cross-validation*. En ambos casos, el particionado se puede realizar por tres métodos distintos: aleatorio, estratificado iterativo y estratificado mediante LP. Por otro lado, se incluyen dos métodos de selección de características: aleatorio y basado en BR. El método basado en BR necesita tres parámetros, correspondientes a los métodos de combinación (máximo, mínimo o media), normalización (dividir por longitud, dividir por el máximo o no normalizar) y puntuación (por puntuación de evaluación o por *ranking*). Por último, también ofrece la posibilidad de hacer un muestreo aleatorio de instancias.

3.5. Multiple datasets

La pestaña *Multiple datasets* permite cargar varios *datasets* simultáneamente y calcular un conjunto de métricas para todos ellos. Los resultados se pueden guardar como *.txt*, *.csv*, *.arff* y *.tex*. Así, se podrán analizar los resultados de todos los *datasets* conjuntamente. La Tabla 2 muestra la salida de algunas métricas para una división en *5-folds* generada para el *dataset* genbase, incluyendo también el *dataset* original.

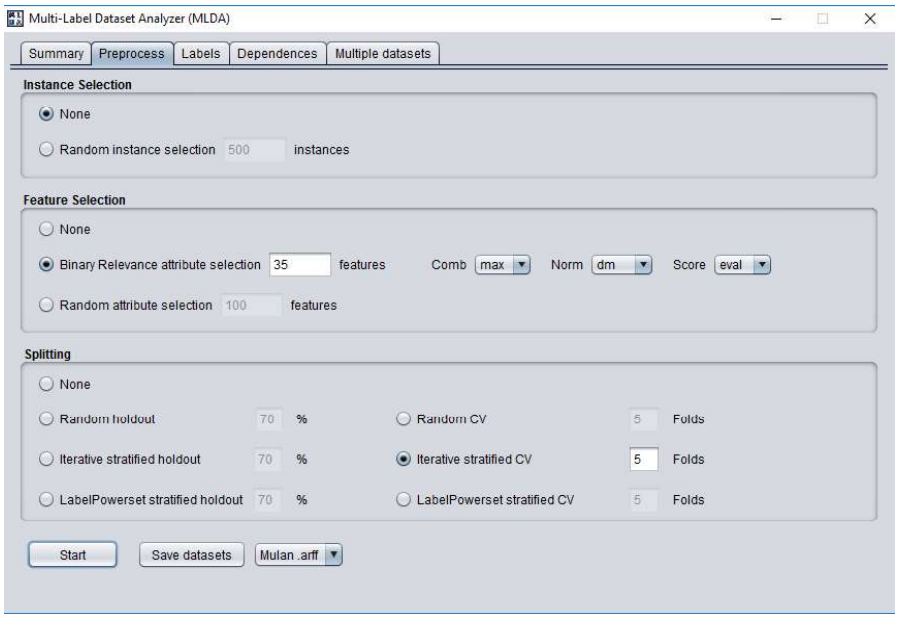


Figura 7. Pestaña de preprocesado de datos

4. Conclusiones

En este trabajo se ha desarrollado una herramienta para la exploración y análisis de *datasets* para aprendizaje multi-etiqueta. Dicha herramienta permite obtener un gran conjunto de métricas de caracterización del *dataset*, información de las etiquetas tales como desbalanceo o dependencias y diversos gráficos representando dicha información. Además de leer los principales formatos de *datasets* multi-etiqueta, proporciona herramientas para conversión de formato, particionado y selección de atributos e instancias. También permite cargar varios *datasets* y calcular un conjunto de métricas para todos ellos.

Como líneas de trabajo futuro, se pretende ampliar los métodos de preprocesado y transformación de los datos, así como proporcionar una API para el cálculo de las distintas métricas de caracterización de los *datasets*.

5. Agradecimientos

Este trabajo ha sido financiado por el proyecto TIN2014-55252-P del Ministerio de Economía y Competitividad y fondos FEDER.

Referencias

1. Jfreechart. `http://www.jfree.org/jfreechart/`, Último acceso: 19-04-2016

Tabla 2. Salida de *multiple datasets*

	genbase	train1	test1	train2	test2	train3	test3	train4	test4	train5	test5
Average examples per labelset	20.688	17.667	6.286	18.276	6.000	17.065	6.650	17.667	6.286	17.633	6.333
Cardinality	1.252	1.253	1.250	1.251	1.258	1.251	1.256	1.255	1.242	1.251	1.256
CVIR inter class	1.449	1.412	0.977	1.358	1.039	1.360	1.010	1.419	0.963	1.353	1.032
Labels	27	27	27	27	27	27	27	27	27	27	27
Mean of IR inter class	37.315	30.976	13.448	34.958	12.175	28.618	11.596	30.627	14.728	35.361	12.260
Mean of IR intra class	143.458	114.432	43.622	134.238	37.699	105.449	36.145	113.973	46.466	135.539	38.261
Mean of IR per labelset	70.219	54.377	16.686	54.041	17.006	57.473	14.302	55.583	16.483	54.764	16.533
Number of labelsets up to 10 examples	19	18	17	17	17	19	16	18	17	18	17
Number of labelsets up to 50 examples	27	27	21	26	22	28	20	27	21	27	21
PMax	0.257	0.257	0.258	0.258	0.250	0.257	0.256	0.255	0.265	0.257	0.256
PUniq	0.015	0.015	0.061	0.015	0.068	0.017	0.038	0.017	0.053	0.015	0.060

2. Jgraphx. `https://www.jgraph.com`, Último acceso: 19-04-2016
3. Meka: A multi-label extension to weka. `http://meka.sourceforge.net/`, Último acceso: 21-04-2016
4. Charte, F., Charte, D.: Working with multilabel datasets in R: The mldr package. The R Journal 7(2), 149–162 (dec 2015)
5. Chekina, L., Rokach, L., Shapira, B.: Meta-learning for selecting a multi-label classification algorithm. pp. 220–227 (2011)
6. Cohen, J., Cohen, P., West, S.G., Aiken, L.S.: Applied Multiple Regression / Correlation Analysis for the Behavioral Sciences (2002)
7. Gibaja, E., Ventura, S.: A tutorial on multilabel learning. ACM Computing Surveys 47(3) (2015)
8. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: An update. SIGKDD Explor. Newsl. 11(1), 10–18 (2009)
9. Nasierding, G., Kouzani, A.: Image to text translation by multi-label classification. In: Advanced Intelligent Computing Theories and Applications with Aspects of Artificial Intelligence. vol. 6216, pp. 247–254 (2010)
10. Sechidis, K., Tsoumakas, G., Vlahavas, I.: On the stratification of multi-label data. Lecture Notes in Computer Science 6913 LNAI(PART 3), 145–158 (2011)
11. Shao, H., Li, G., Liu, G., Wang, Y.: Symptom selection for multi-label data of inquiry diagnosis in traditional chinese medicine. Sci China Ser F-Info Sci (1), 1–13 (2010)
12. Stamatatos, E.: Author identification: Using text sampling to handle the class imbalance problem. Information Processing and Management 44(2), 790–799 (2008)
13. Tsoumakas, G., Katakis, I., Vlahavas, I.: Data Mining and Knowledge Discovery Handbook, Part 6, chap. Mining Multi-label Data, pp. 667–685. Springer (2010)
14. Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J., Vlahavas, I.: Mulan: A java library for multi-label learning. Journal of Machine Learning Research 12, 2411–2414 (2011)
15. Zhang, Y., Burer, S., Street, W., Bennett, K., Parrado-hern, E.: Ensemble pruning via semi-definite programming. Journal of Machine Learning Research 7, 1315–1338 (2006)