

MLDR

PAQUETE R PARA EXPLORACIÓN MULTIETIQUETA

David Charte Francisco Charte

11 nov 2015 – TAMIDA (Retos) – CAEPIA '15



Soft Computing and Intelligent Information Systems – Universidad de Granada

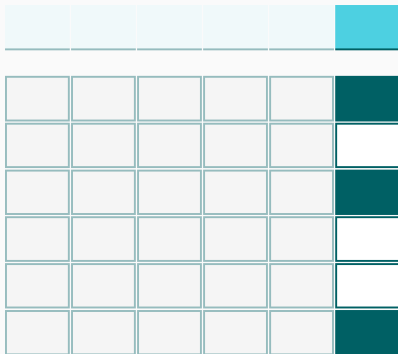
CLASIFICACIÓN DE DATOS

Aplicaciones:

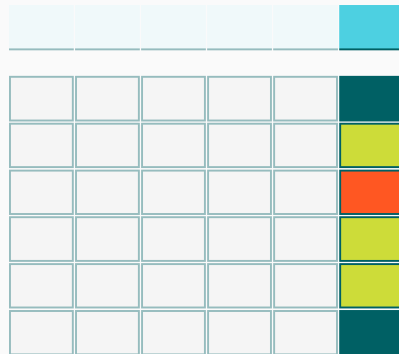
- Detección de spam
- Diagnóstico de enfermedades
- Detección de fraude
- Predicción de riesgos
- ...

CLASIFICACIÓN TRADICIONAL

Clasificación binaria



Clasificación multiclase

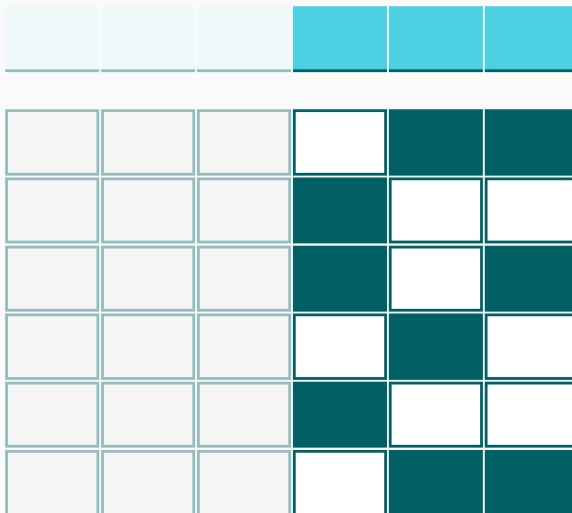


INFORMACIÓN NO BINARIA/MULTICLASE

- Escenas/elementos en fotografías
- Publicaciones de texto
- Contenido multimedia
- ...

Categorías no excluyentes \Rightarrow Etiquetas

CLASIFICACIÓN MULTIETIQUETA



CLASIFICACIÓN MULTIETIQUETA

- Instancia: $(X, \mathbf{Y}) \in X^1 \times X^2 \times \dots \times X^f \times \mathcal{P}(L)$
- Para cada instancia hay $2^{|L|}$ posibles predicciones
- ¿Dependencia entre etiquetas?

CLASIFICACIÓN MULTIETIQUETA

- Instancia: $(X, \mathbf{Y}) \in X^1 \times X^2 \times \dots \times X^f \times \mathcal{P}(L)$
- Para cada instancia hay $2^{|L|}$ posibles predicciones
- ¿Dependencia entre etiquetas?
- Adaptación de algoritmos
- Transformación de datos
 - Binary Relevance: 1 problema multietiqueta $\sim |L|$ problemas binarios
 - Label Powerset: 1 problema multietiqueta \sim 1 problema multiclase con $2^{|L|}$ clases

CLASIFICACIÓN MULTIETIQUETA

- Instancia: $(X, \mathbf{Y}) \in X^1 \times X^2 \times \dots \times X^f \times \mathcal{P}(L)$
- Para cada instancia hay $2^{|L|}$ posibles predicciones
- ¿Dependencia entre etiquetas?
- Adaptación de algoritmos
- Transformación de datos
 - Binary Relevance: 1 problema multietiqueta $\sim |L|$ problemas binarios
 - Label Powerset: 1 problema multietiqueta \sim 1 problema multiclase con $2^{|L|}$ clases
- Nuevas métricas para obtener más información acerca de los datos:
Card

ÍNDICE

1. Introducción

2. El paquete mldr

3. Análisis exploratorio

4. Clasificación

MOTIVACIÓN

- Necesidad de una herramienta accesible para exploración de datos multietiqueta
- Potencial de R para manejo de datos: estructuras de datos, instrucciones vectorizadas...
- Paquetes de gráficos disponibles para R
- Facilidad de interacción desde la consola interactiva de R

INSTALACIÓN Y CARGA

Disponible en CRAN

```
install.packages("mlr")  
library(mlr)
```

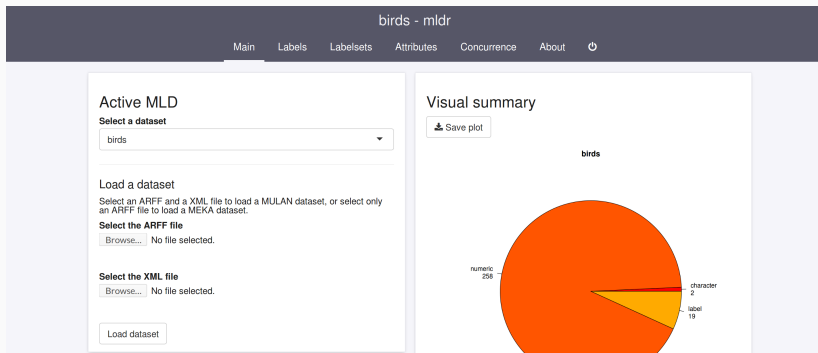
INSTALACIÓN Y CARGA

Disponible en CRAN

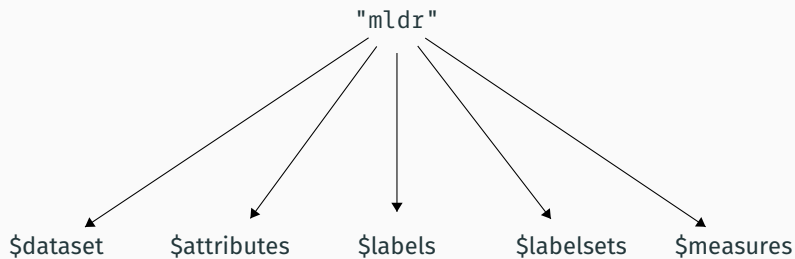
```
install.packages("mldr")
```

```
library(mldr)
```

```
mldrGUI()
```



ESTRUCTURA DE UN OBJETO MLDR



LECTURA Y CREACIÓN DE DATASETS

- Datasets en formato ARFF de Mulan y MEKA:

```
emotions <- mldr("emotions")
```

```
enron <- mldr("ENRON-F", use_xml = FALSE)
```

LECTURA Y CREACIÓN DE DATASETS

- Datasets en formato ARFF de Mulan y MEKA:

```
emotions <- mldr("emotions")
```

```
enron <- mldr("ENRON-F", use_xml = FALSE)
```

- Datasets de ejemplo: *emotions*, *birds*, *genbase*

LECTURA Y CREACIÓN DE DATASETS

- Datasets en formato ARFF de Mulan y MEKA:

```
emotions <- mldr("emotions")  
enron <- mldr("ENRON-F", use_xml = FALSE)
```

- Datasets de ejemplo: *emotions*, *birds*, *genbase*

- Creación de nuevos datasets desde `data.frames`:

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

LECTURA Y CREACIÓN DE DATASETS

- Datasets en formato ARFF de Mulan y MEKA:

```
emotions <- mldr("emotions")  
enron <- mldr("ENRON-F", use_xml = FALSE)
```

- Datasets de ejemplo: *emotions*, *birds*, *genbase*

- Creación de nuevos datasets desde `data.frames`:

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

- Filtrado de datasets

OBTENCIÓN DE MEDIDAS

```
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
```

OBTENCIÓN DE MEDIDAS

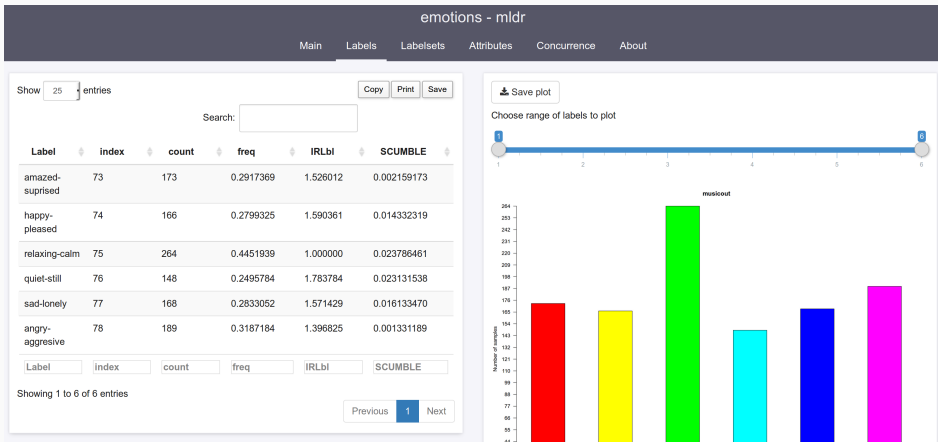
```
summary(emotions)
```

```
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
```

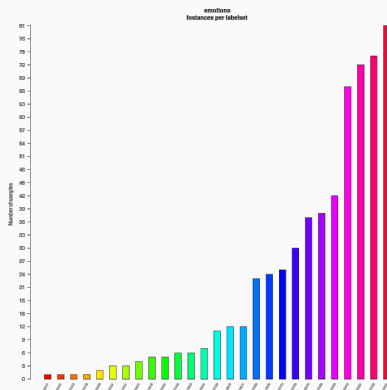
	index	count	freq	IRLbl	SCUMBLE
amazed-surprised	73	173	0.2917369	1.526012	0.002159173
happy-pleased	74	166	0.2799325	1.590361	0.014332319
relaxing-calm	75	264	0.4451939	1.000000	0.023786461
quiet-still	76	148	0.2495784	1.783784	0.023131538
sad-lonely	77	168	0.2833052	1.571429	0.016133470
angry-aggressive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

OBTENCIÓN DE MEDIDAS

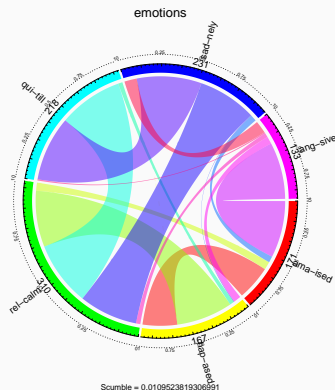


GENERACIÓN DE GRÁFICOS

```
plot(emotions, type = "LSB")
```

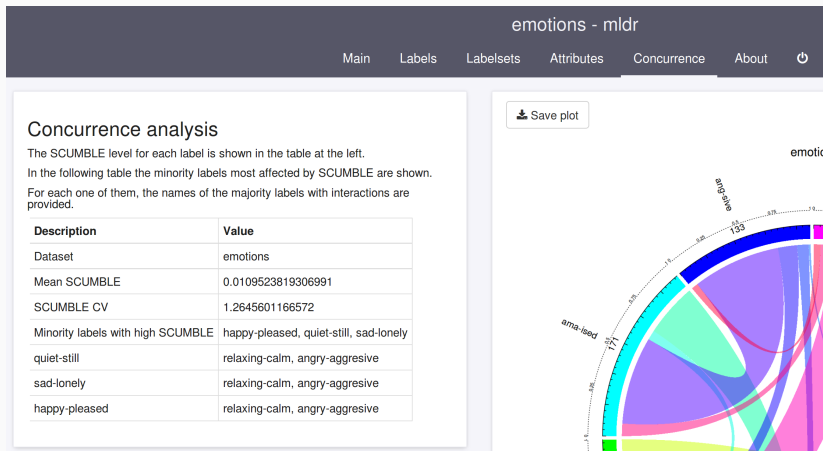


```
plot(emotions, type = "LC")
```



INFORME DE CONCURRENCIA

Búsqueda de etiquetas difíciles



CLASIFICACIÓN CON MLDR

- Transformaciones Label Powerset y Binary Relevance
- 19 métricas de evaluación de resultados
- (*Pronto*) Interfaz común para implementación de clasificadores

```
mldr_evaluate(emotions, predictions)
```

```
List of 20
```

```
$ Accuracy      : num 0.912
$ AUC           : num 0.916
$ AveragePrecision: num 0.669
$ Coverage      : num 2.72
$ FMeasure     : num 0.942
$ HammingLoss   : num 0.0883
$ MacroAUC     : num 0.919
$ MacroFMeasure : num 0.865
$ MacroPrecision: num 0.805
$ MacroRecall   : num 0.936
$ MicroAUC     : num 0.918
$ MicroFMeasure : num 0.868
$ MicroPrecision: num 0.811
$ MicroRecall   : num 0.935
$ OneError     : num 0.111
$ Precision    : num 0.927
$ RankingLoss  : num 0.508
$ Recall       : num 0.927
$ SubsetAccuracy: num 0.831
$ ROC          :List of 15
```


GRACIAS POR SU ATENCIÓN

MLDR

PAQUETE R PARA EXPLORACIÓN MULTIETIQUETA

Repositorio CRAN: <http://cran.r-project.org/web/packages/mldr/>

Proyecto en GitHub: <https://github.com/fcharte/mldr>

Aplicación en ShinyApps: <https://fdavidcl.shinyapps.io/mldr>