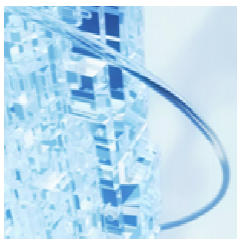


Prácticas de Minería de Datos

Grado en Ingeniería Informática

Curso 2013-14

PRÁCTICA 1

Conocimiento del entorno WEKA

OBJETIVOS

- Familiarizarse con el entorno de minería de datos WEKA
- Conocer el formato de datos utilizado por Weka.
- Conocer las principales funciones de visualización de datos

1. ¿Qué es WEKA?

WEKA es una herramienta gratuita para experimentación de datamining, escrita en lenguaje Java, que permite aplicar, analizar y evaluar las técnicas más relevantes de análisis de datos, principalmente las provenientes del aprendizaje automático, sobre cualquier conjunto de datos del usuario.

Para ello únicamente se requiere que los datos a analizar se almacenen con un cierto formato, conocido como arff.

2. Formato de fichero de Weka

Los datos de entrada de la herramienta Weka deben estar codificados en un formato específico, denominado Attribute-Relation File Format (extensión "arff"), aunque también admite otros formatos de entrada de datos. La herramienta permite cargar los datos en tres soportes: fichero de texto, acceso a una base de datos y acceso a través de Internet mediante una dirección URL. En nuestras prácticas utilizaremos ficheros de texto.

El formato "arff" está compuesto por una estructura claramente diferenciada en tres partes:

1. Cabecera. Se define el nombre de la relación. Su formato es el siguiente:

```
@relation <nombre-de-la-relación>
```

2. Declaraciones de atributos. En esta sección se declaran los atributos de la base de datos junto a su tipo. La sintaxis es la siguiente:

```
@attribute <nombre-del-atributo> <tipo>
```

Weka admite atributos numéricos y nominales. Para los primeros, es necesario añadir la palabra real o integer (o simplemente numeric) tras el nombre del atributo. Para los atributos nominales se especifican, entre llaves, los posibles valores que puede tomar.

```
@attribute tiempo {soleado, lluvioso, nublado}
@attribute nota real
```

Además de estos tipos de atributos, Weka admite datos de tipo cadena (string) y de tipo fecha (date)

```
@attribute nombre string
@attribute hoy fecha
```

3. Sección de datos. Cada instancia debe estar en una fila y los atributos se separan con comas.

```
@data
4,3.2
```

También es posible definir los datos de una forma abreviada (*sparse data*). Si tenemos una muestra en la que hay muchos datos que sean 0 podemos expresar los datos prescindiendo de estos elementos, poniendo cada una de las filas entre llaves y situando delante de cada uno de los datos el número de atributo (hay que tener en cuenta que la numeración de los atributos comienza en 0).

```
@data
0, 26, 0, 0, 0, 0, 63, 0, 0, 0, "class A"
0, 0, 0, 42, 0, 0, 0, 0, 0, 0, "class B"

@data
{1 26, 6 63, 10 "class A"}
{3 42, 10 "class B"}
```

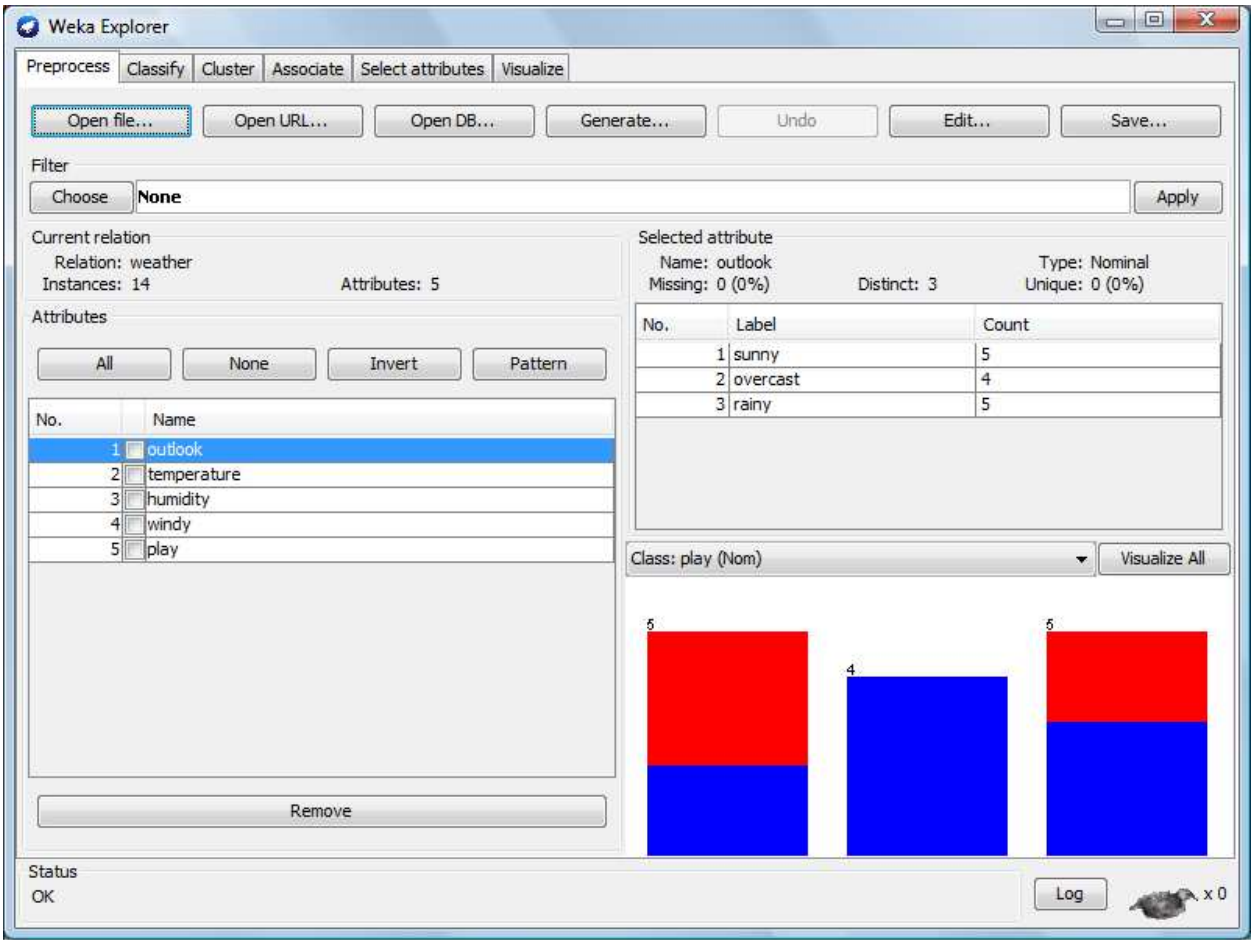
Los **valores desconocidos** (*missing values*) se representan con el símbolo '?'

```
María,?,?,23,5.5
```

3. Explorer

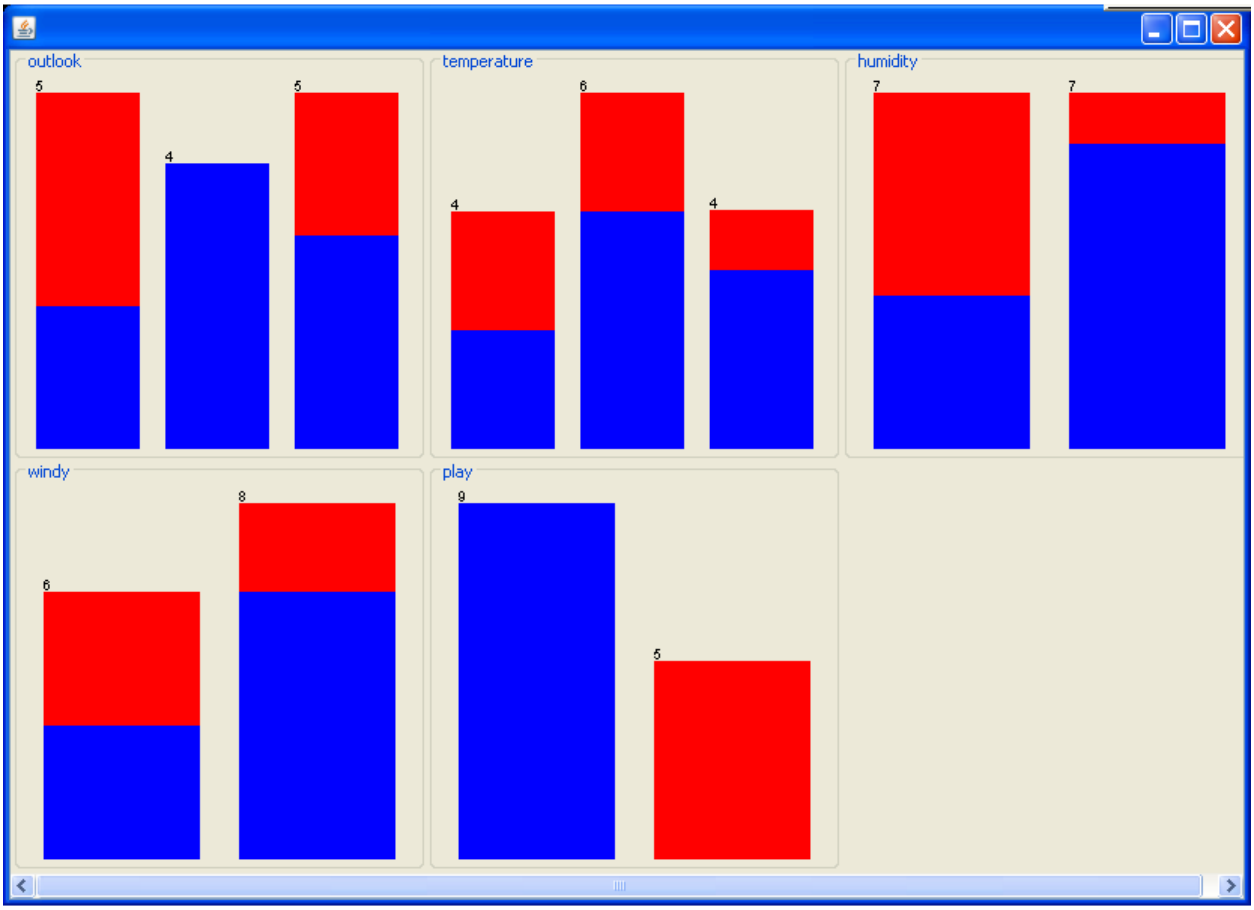
En esta apartado se describe someramente el entorno Explorer, ya que permite el acceso a la mayoría de las funcionalidades integradas en Weka de una manera sencilla.

Veamos qué ocurre cuando abrimos un fichero. El fichero weather.arff contiene información acerca de los días que se ha podido jugar al tenis, dependiendo de diversos aspectos meteorológicos. Un posible objetivo podría ser determinar o predecir si hoy podremos jugar al tenis. En la siguiente figura se puede apreciar la ventana que aparece al abrir el fichero:



Desde esta ventana se puede obtener mucha información acerca del fichero que acabamos de cargar. Por ejemplo, el sistema indica que hay 14 instancias o registros con 5 atributos. Al seleccionar cada uno de los atributos, se puede obtener más información de dicho atributo: **tipo** (nominal o numérico), **número de valores distintos** (*distinct*), número y porcentaje de instancias con valor desconocido (*missing values*), los **valores máximos, mínimo, medio** y **desviación estándar** (sólo en atributos numéricos), valores que solamente se dan en una instancia (*unique*), y un **histograma** con información sobre la distribución de los ejemplos para ese atributo, reflejando con el uso de colores la distribución de clases de cada uno de los registros. Por ejemplo, el atributo Outlook tiene tres valores diferentes (*Sunny*, *Overcast* y *Rainy*) siendo la distribución de [5, 4, 5]. Esto significa que de los 5 registros donde el atributo **Outlook=sunny**, 3 tienen clase **no** y 2 tienen clase **yes**; cuando **Outlook=overcast**, los 4 registros tienen clase **yes**, y cuando **Outlook=rainy**, hay 3 con clase **yes** y 2 con clase **no**.

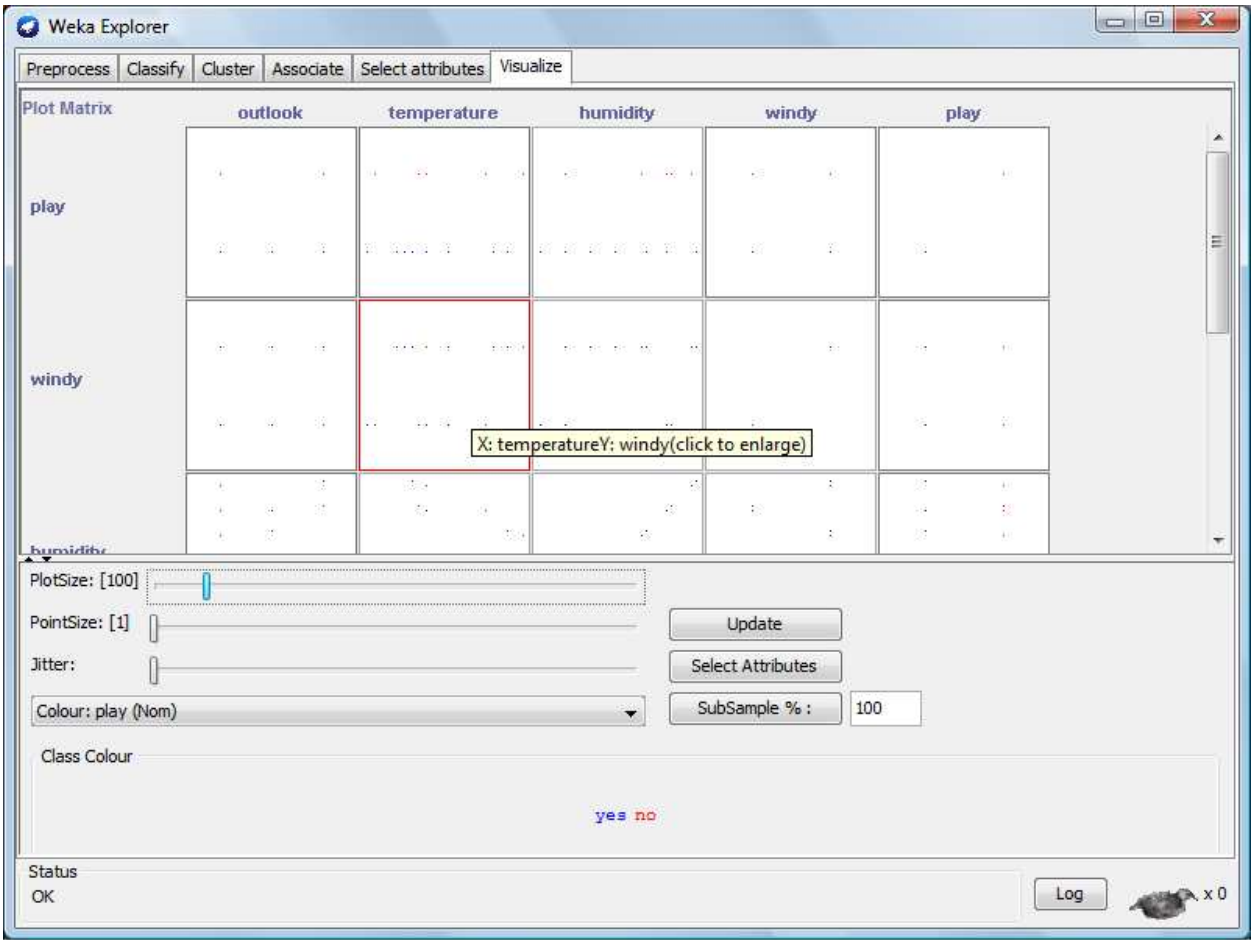
Si queremos ver todos los histogramas en una sola pantalla, podemos pulsar "*Visualize all*"



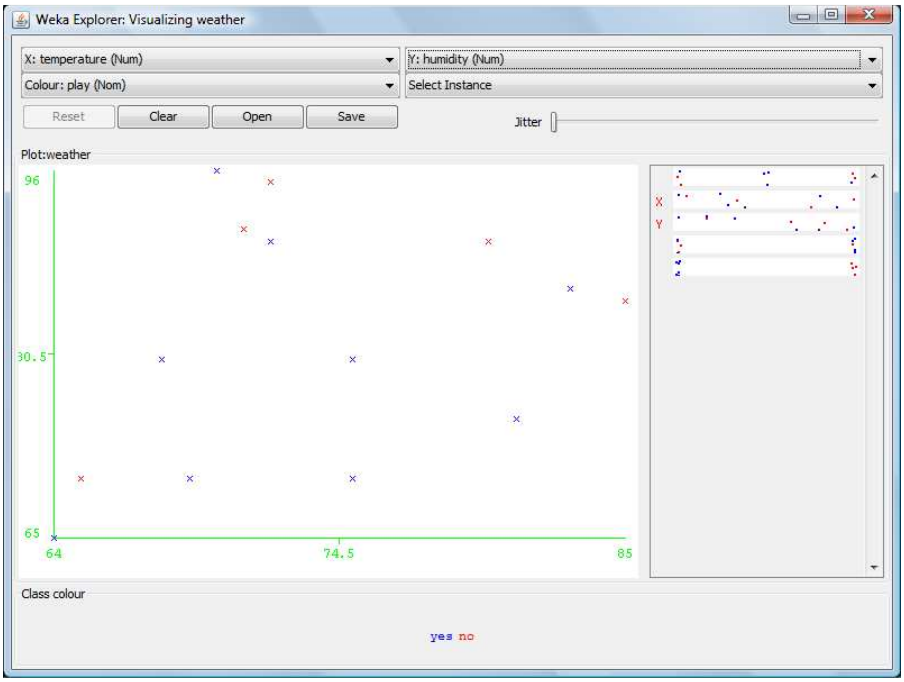
4. Visualización

Una de las primeras etapas del análisis de datos puede ser simple análisis visual de éstos, en ocasiones de gran utilidad para desvelar relaciones de interés. La herramienta de visualización de WEKA permite presentar gráficas 2D que relacionen pares de atributos, con la opción de utilizar colores para añadir información de un tercer atributo. Además, tiene incorporada una facilidad interactiva para seleccionar instancias con el ratón.

Al seleccionar la opción ***Visualize*** del Explorer aparecen todos los pares posibles de atributos en las coordenadas horizontal y vertical.



Se puede seleccionar la gráfica deseada para verla en detalle en una ventana nueva.



Ejercicios

1. Supongamos que tenemos una base de datos almacenada en una hoja de cálculo. Convertir el fichero **"miopía.xls"** en formato **"arff"** para que se puede trabajar con Weka.

2. Actualmente, existen muchos ficheros en la red en formato **"arff"**, sobre todo en el repositorio de la Universidad de California – Irvine (*UCI Repository*). Su página web es: http://archive.ics.uci.edu/ml/datasets.html. Para cada base de datos se muestra una breve descripción con cierta información sobre los atributos, el uso que se le puede dar, etc. Para este ejercicio usaremos una base de datos almacenada en el fichero llamado **"labor.arff"** (se encuentra dentro de la carpeta de Weka).
   Responder a las siguientes cuestiones:

   - ¿Qué almacena la base de datos?
   - Número de instancias y número de atributos
   - ¿Cuántos valores tiene el atributo que hace de 'clase'?, ¿cuáles son?
   - ¿Cuál es el atributo que tiene más valores ausentes?, ¿cuántos y qué porcentaje?
   - ¿Cuál es el atributo que tiene más valores únicos?, ¿cuáles son esos valores?
   - Estudiando los histogramas de todos los atributos:
     - ¿qué conclusiones se pueden obtener de los valores del atributo "*working-hours*"?
     - ¿cuál es el atributo que mejor divide a la clase? Escribe varias reglas del tipo **"si A es X entonces la clase es Y"**, donde **X** es una etiqueta que puede tomar los valores: *muy bajo*, *bajo*, *medio*, *alto* o *muy alto*, e **Y** es un valor de la clase
     - ¿cuál es el atributo que peor divide a la clase?, ¿por qué?

   - Utiliza la herramienta de visualización 2-D para mostrar el atributo "wage-increase-first-year" en el eje X.
     - ¿Con qué otro atributo en el eje Y se divide mejor a la clase? Escribe varias reglas del tipo **"si wage-increase-first-year es X y A es Y entonces la clase es Z"**, donde **X** e **Y** son etiquetas que indica si el valor es muy bajo, bajo, medio, alto o muy alto, y **Z** es un valor de la clase.

¿Cómo entregar la práctica?

- Utilizar un documento de texto para responder a las cuestiones y subirlo a través de la plataforma web
- Para el ejercicio 1, escribir el contenido del fichero "arff" que habéis creado.
- Para el ejercicio 2, responder a las preguntas. Podéis utilizar pantallas gráficas (histogramas, gráficas 2-D, etc.) para complementar las respuestas