



Prácticas de Minería de Datos

Grado en Ingeniería Informática

Curso 2013-14

PRÁCTICA 2

Conocimiento del entorno WEKA

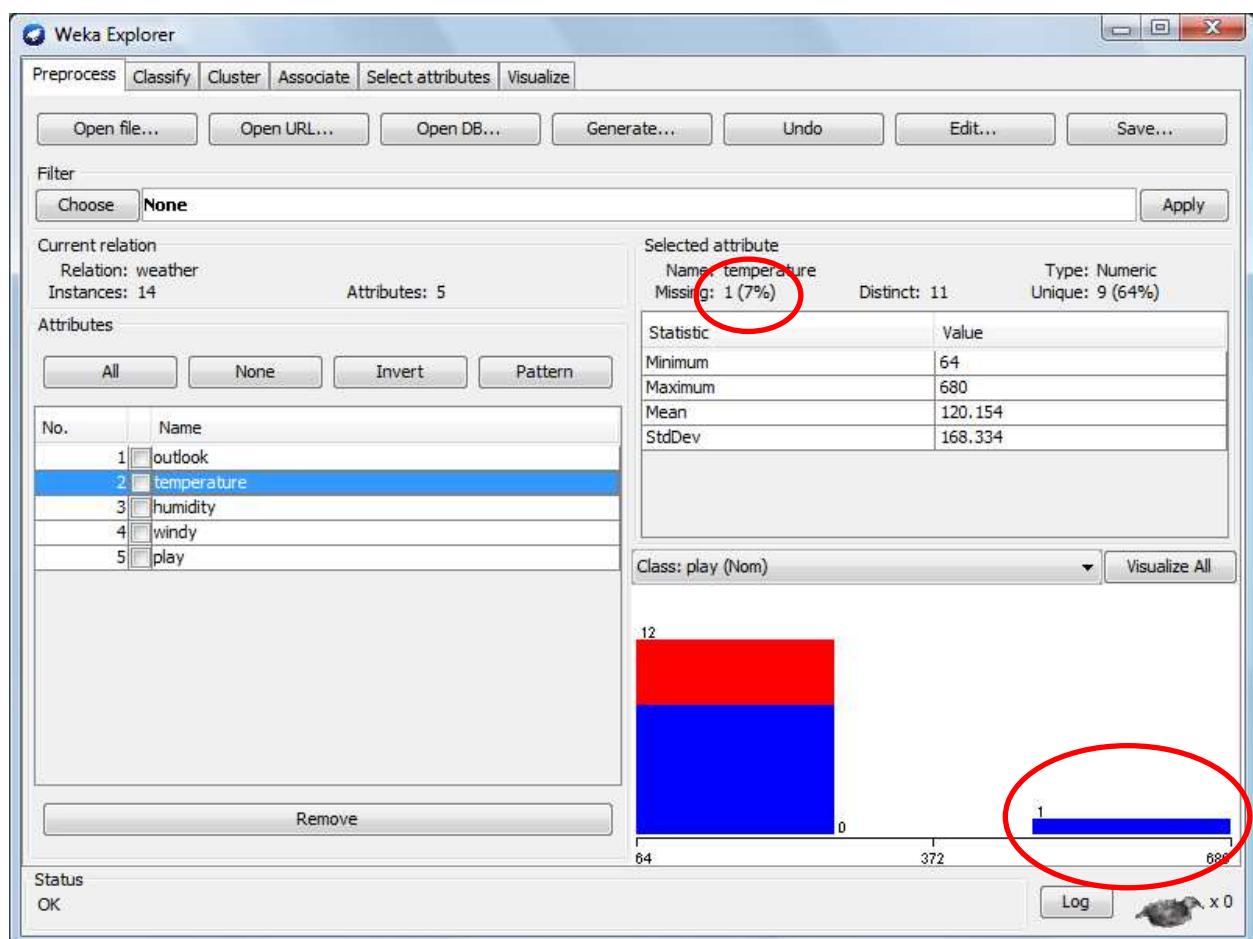


OBJETIVOS

- Familiarizarse con el entorno de minería de datos WEKA
- Utilizar las principales funciones de visualización de datos para detectar valores anómalos y ausentes

1. Explorer

Cuando abrimos un fichero en Weka nos aparece la ventana de preprocesamiento desde donde se puede obtener mucha información acerca del fichero que acabamos de cargar. Analizando los datos estadísticos y los histogramas podemos detectar valores erróneos:

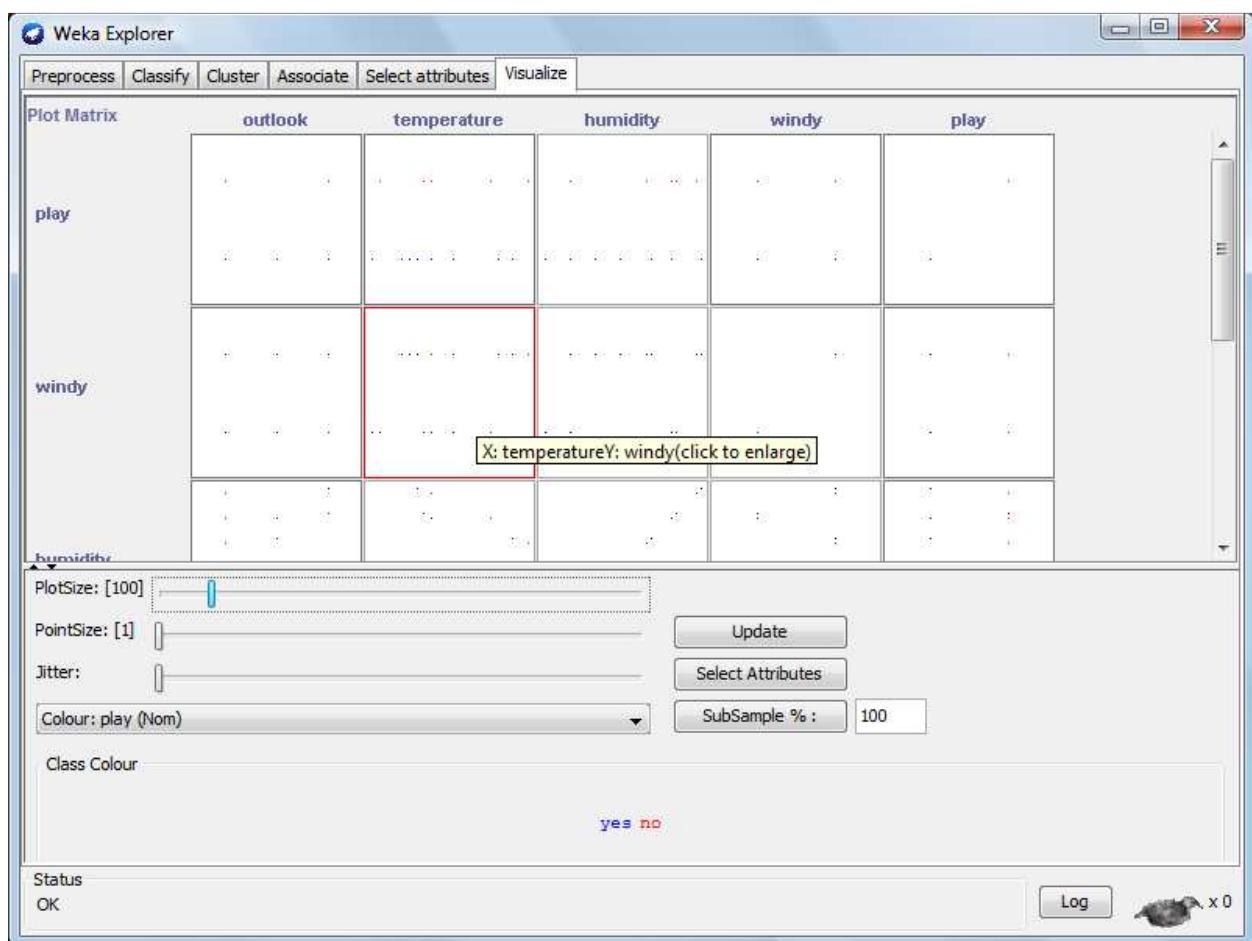


Por ejemplo, al seleccionar el atributo temperatura del fichero weather, y observar los datos estadísticos podemos ver que de las 14 instancias hay 1 con valor desconocido (*missing values*), y analizando los **valores máximos, mínimo, medio y desviación estándar** (sólo en atributos numéricos), junto con el **histograma** con información sobre la distribución de los ejemplos para ese atributo, podemos detectar un valor extraño que es mucho mayor que el resto, lo que hace pensar que se trata de un dato erróneo.

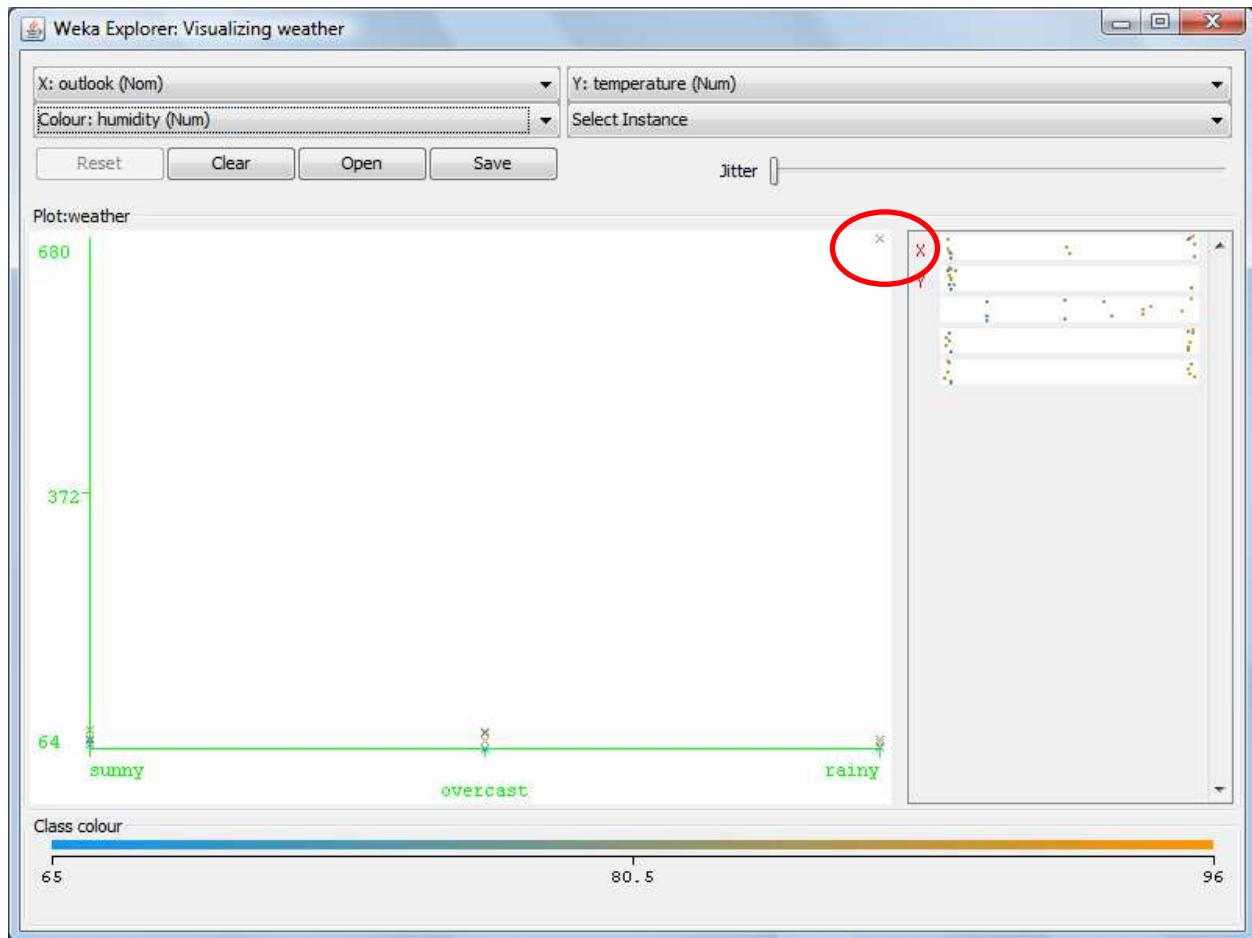
2. Visualización

Podemos investigar los valores anómalos con más detalle, mediante el uso de gráficas por campos. Nos interesa ver los datos numéricos, ya que en estos casos es más fácil detectar outliers (datos anómalos) gráficamente. Para ello podemos usar la herramienta de visualización de WEKA que permite presentar gráficas 2D que relacionen pares de atributos para comparar diferentes valores, con la opción de utilizar colores para añadir información de un tercer atributo. Por ejemplo, podemos querer visualizar la temperatura respecto al tiempo (meteorología) y además mostrando la humedad.

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.



Al igual que antes, analizando dichas gráficas es posible detectar valores erróneos (obsérvese el valor que aparece en la parte superior derecha)

3. Edición de datos

Una vez detectados los valores erróneos o faltantes podemos editar los datos para corregirlos manualmente o automatizar parte de dicho proceso aplicando filtros.

Ejercicios

1. Supongamos que tenemos una base de datos almacenada en un fichero de texto plano (los campos están separados por tabuladores). Convertir el fichero “**empleadosP2.txt**” en formato “**arff**” para que se puede trabajar con Weka.
2. Examine detenidamente la base de datos anterior y responda a las siguientes cuestiones:
 - ¿Qué almacena la base de datos?
 - Indique el significado de cada uno de los campos
 - Número de instancias y número de atributos
 - ¿Cuáles son los atributos que tienen más valores ausentes?, ¿cuántos y qué porcentaje?
 - ¿Hay alguna instancia que se pueda eliminar? ¿Por qué? Elimina dichas instancias de forma manual editando los datos
 - Investiga si hay alguna otra forma de eliminar dichas instancias de forma automática con `weka.filters.unsupervised.instance.RemoveWithValues`
 - Indica al menos 2 formas diferentes de hacerlo con este filtro:
 - Enumera los atributos con valores ausentes, indicando cuantos tiene y en qué porcentaje (responde a esta pregunta una vez eliminadas las instancias no válidas anteriores)
 - ¿Cuál es el atributo que tiene más valores únicos?, ¿cuáles son esos valores?
 - Estudiando la información estadística de los atributos:
 - ¿Cuántos valores tiene el atributo que hace de ‘clase’?, ¿cuáles son? ¿Qué conclusiones se pueden obtener de los valores del atributo de clase?
 - ¿Qué campos tiene atributos erróneos?, ¿por qué?
 - Estudiando los histogramas de todos los atributos:
 - ¿Qué campos parecen tener atributos erróneos?, ¿por qué?
 - Para confirmar las sospechas de datos erróneos utiliza la herramienta de visualización 2-D para mostrar el atributo antigüedad respecto al sueldo y además mostrando los estudios.
3. Una vez identificado los campos con atributos erróneos, verás que algunos son nominales y otros son numéricos. En el caso de los atributos numéricos erróneos identifica las instancias que los tienen y justifica la acción a realizar (elige entre eliminar dichas instancias o asignarle un valor correcto).
 - Si piensas que es mejor eliminar dichas instancias indica cuantas has eliminado, en caso contrario indica qué valores le has asignado a los atributos erróneos.
4. Edita los datos y corrige los datos nominales erróneos.
 - Indica para cada atributo con valores nominales erróneos que valor has considerado como erróneo y a qué valor lo has convertido.

5. Salva los datos en disco con el nombre “**empleadosP2_v1.arff**”. Edita el fichero arff y elimina los valores de los atributos erróneos anteriores.

- Carga el fichero **empleadosP2_v1.arff** y comprueba que está todo correcto.

6. Una vez resuelto el tema de los datos erróneos vamos a ocuparnos ahora de los datos con valores ausentes. Edita los datos y corrige los datos nominales erróneos.

- Para el campo Estudios suponed que un dato ausente indica que sólo tiene estudios obligatorios. A fin de conservar los datos originales del campo Estudios realiza los cambios en una copia de dicho campo llamada CEstudios. Utiliza para ello el filtro:

```
weka.filters.unsupervised.attribute.Copy
```

- Para el campo Hijos indica cuál es el criterio más apropiado y en base a ello rellena los valores ausentes con dicho valor.
- Para el resto de campos con valores ausentes rellenarlo con el valor más frecuente para dicho atributo (el campo Estudios debe conservar los valores perdidos). Utiliza para ello el filtro:

```
weka.filters.unsupervised.attribute.ReplaceMissingValues
```

7. Salva los datos en disco con el nombre “**empleadosP2_ok.arff**”.

- Utilizando la herramienta de visualización 2-D responde a las siguientes cuestiones:

- Visualiza la gráfica antigüedad - sueldo en función de los estudios. ¿Influye la antigüedad en el sueldo o este solo viene determinado por el nivel de estudios?
- Visualiza la gráfica Sueldo – Estudios ¿Qué estudios tienen los que ganan más? ¿y los que ganan menos?
- Visualiza la gráfica antigüedad - sueldo en función del estado civil (es decir, si están casados o no).

¿Cómo entregar la práctica?

- Utilizar un documento de texto para responder a las cuestiones y subirlo a través de la plataforma web, junto con los ficheros .arff pedidos.
- Responde a cada una de las preguntas, utilizando pantallas gráficas (histogramas, gráficas 2-D, etc.) para complementar las respuestas cuando sea necesario.