

PRÁCTICA 1

Selección de Variables con Weka



Introducción

Weka es un conjunto de librerías JAVA para la extracción de conocimientos desde bases de datos. Es un software que ha sido desarrollado bajo licencia GPL por la universidad de Waikato, lo cual ha impulsado que sea una de las suites más utilizadas en el área en los últimos años.

Tiene un formato específico: .arff, para guardar las bases de datos. Una vez cargado un .arff, el Explorador de Weka nos ofrece resúmenes estadísticos sobre el contenido de la base de datos, y nos permite aplicar sobre ésta diferentes procesos relacionados con Descubrimiento del Conocimiento:

- Selección de variables
- Selección de registros
- Discretizar variables numéricas
- Construir modelos predictivos
- Evaluar los modelos predictivos
- Visualizar los registros con técnicas de brushing

En esta práctica nos centraremos en el primer y último punto, los cuales como hemos visto en clase están relacionados con la Visualización de la Información.

Paso 0: Antes de trabajar con una base de datos, conviene conocer su procedencia y descripción de sus variables. Trabajaremos con la base de datos ‘Pima Indians Diabetes’, obtenida del repositorio de la UCI en esta dirección: <http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes>.

Paso 1: De Excel a Weka

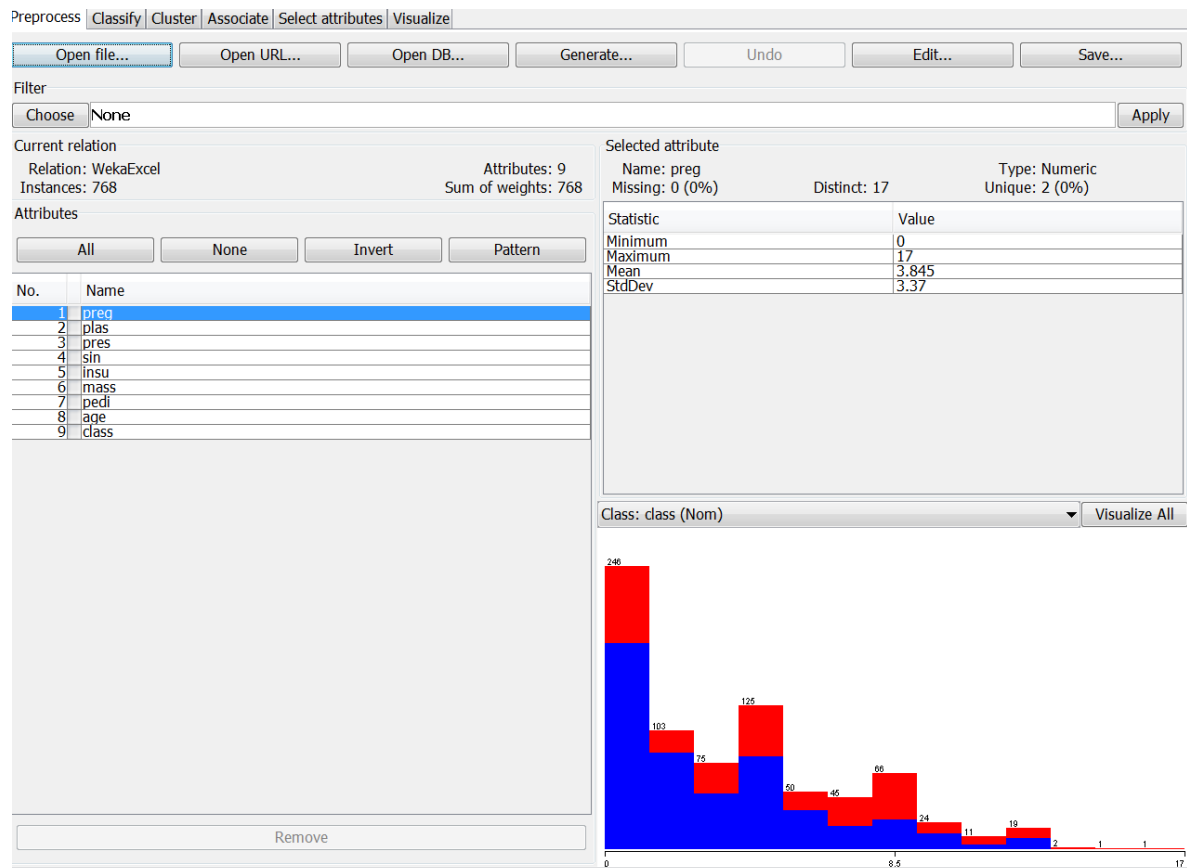
Normalmente trabajaremos con bases de datos en formato Excel. Así que Weka permite importar los archivos .xls y .xlsx a .arff, siempre que sigan 2 condiciones:

- La primera fila del archivo Excel contiene 1 nombre de variable por columna.
- Las siguientes filas contienen el valor, numérico o texto, para cada variable. Si el valor es perdido, se indica con el símbolo ‘?’.

Convierte a .arff el archivo *diabetes.xlsx*. Para ello:

- 1) Ejecuta Weka
- 2) Selecciona la aplicación *Explorer*
- 3) En el panel *Preprocess*, pulsa el botón *Open file...*, y busca el archivo Excel.
- 4) Guárdalo como .arff, y ábrelo con el Block de Notas para estudiar su formato interno.

Paso 2: Familiarízate con la información que nos da el panel *Preprocess*



- Número de registros o instancias
- Número de atributos
- Porcentaje de valores perdidos del atributo seleccionado
- Tipo del atributo seleccionado (Numérico o Nominal)
- Valores mínimo, máximo, media y desviación estándar del atributo seleccionado
- Histograma de valores del atributo seleccionado

Paso 3: Pulsa el botón “Choose” y busca el filtro necesario para discretizar los atributos numéricos.

- Filtro supervisado (que tenga en cuenta que hay una variable clase) de atributos: *Discretize*
- Observa cómo ha cambiado el histograma de valores de los atributos

Paso 4: Selección de atributos.

-Prueba distintos métodos de selección de atributos en la pestaña *Select Attributes*

- Ranking de atributos
- Selección de subconjuntos de atributos
- ¿Coinciden los atributos seleccionados por las distintas técnicas?

Paso 5: Clasificación

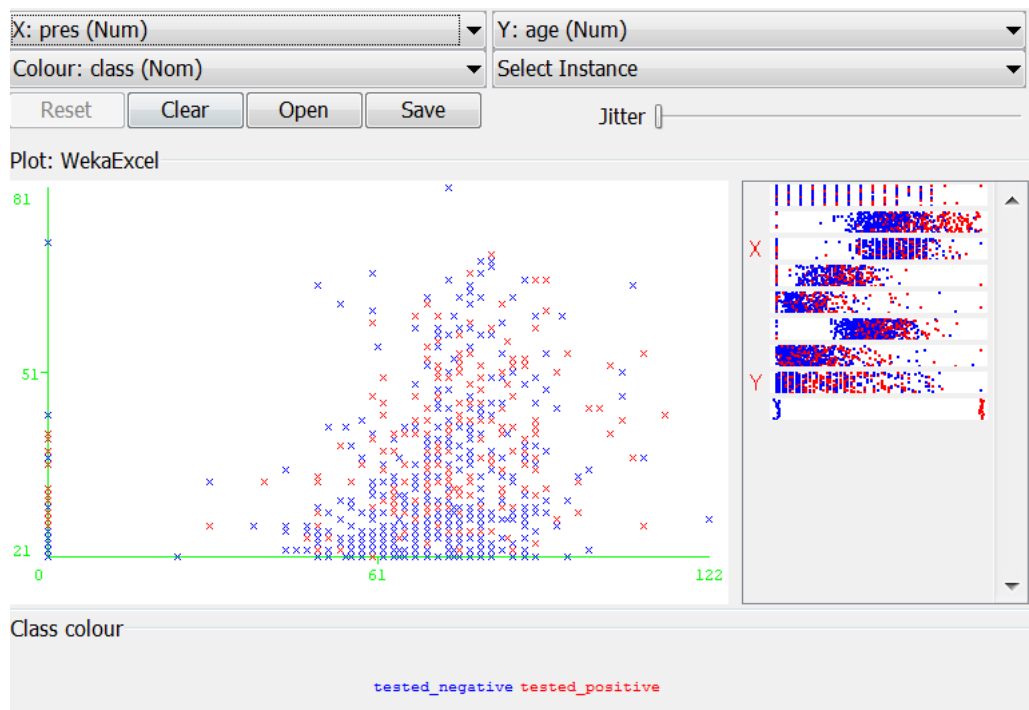
1) En la pestaña *Classify*, usa el clasificador *bayes.NaiveBayes* y el *trees.j48*. Observa los resultados que estos modelos predictivos te dan en Tasa de Aciertos (Accuracy).

2) Ahora, para comprobar si los modelos predictivos mejoran realizando selección de variables, utiliza el meta-clasificador *meta.AttributeSelectedClassifier* y prueba con los mismos clasificadores otra vez, con diferentes técnicas de selección de atributos

Paso 6: Visualización e Interacción

En la pestaña *Visualize* podemos ver una Matriz de Dispersión de la base de datos, y Weka nos permite interactuar con esta para personalizar la visualización.

1) Haz click en una casilla, para poder ver el Diagrama de Dispersión de las 2 variables que correspondan a esa celda.



2) Cambia el color con el que se representan los registros de una clase

3) Juega con “Jitter”. ¿Para qué crees que sirve?

4) Comprueba cómo puedes ir cambiando las variables representadas en cada eje. Así, puedes fijar una en el eje X, y ver cómo varía la dispersión de los puntos para diferentes variables en el eje Y. Encuentra clusters de casos negativos y positivos jugando con los pares de variables. Si no encuentras, compara *mass* vs. *plas*.

5) Con la herramienta *Select Instance*, selecciona un conjunto de registros. Pulsa *Submit* y ahora cambia la variable de un eje, o las dos. ¿Qué ocurre?

6) En el botón *Colour*, puedes seleccionar a qué atributo se refiere un color. Así, selecciona alguna variable distinta a la clase y que sea numérica.

- ¿Suele tomar valores pequeños o grandes?
- ¿Para qué valores de las variables X,Y, la variable seleccionada toma mayores valores?