Gestión y Tecnología del Conocimiento

Minería de Datos

Agosto - Septiembre 2008

Ejercicios de Weka

Comentarios generales sobre los ejercicios

- Asumiendo que se conocen los contenidos teóricos, el tiempo estimado para realizar los ejercicios es de **2 horas**
- Describir las soluciones a los ejercicios de una manera lo más formal posible

Nombre: Franklin Saúl Gancino Mejía.

1. Análisis de los datos:

El objetivo de este ejercicio es familiarizarse con el entorno de Weka, y estudiar algunas de las funcionalidades de análisis de datos. Estas funcionalidades incluyen análisis estadístico, visualización, etc. Recordad que el manual de Weka está disponible en http://www.cs.waikato.ac.nz/ml/weka/index documentation.html.

1.1. Obtención de los datos

Descargar el siguiente conjunto de datos:

iris data set: iris.arff

Abrir el fichero de datos con un editor, y estudiar su contenido:

1. ¿Cuántos atributos caracterizan los datos de esta tabla de datos?

   Existen 5 atributos.

2. Si suponemos que queremos predecir el último atributo a partir de los anteriores, ¿estaríamos ante un problema de clasificación o de regresión?

   Estaríamos ante un problema de regresión.

1.2. Estudio estadístico de los datos

Lanzar la herramienta weka

Lanzar el Explorer

Abrir el fichero iris.arff

Una vez cargado el conjunto de datos, en la sección attributes se puede pinchar sobre cada atributo para obtener información estadística de ellos. Contestad a las siguientes preguntas:

1. ¿Cuál es el rango de valores del atributo petalwitdth?

   Mínimo = 0.1

   Máximo = 2.5

   Media = 1.199

   Desviación estándar = 0.763

2. Con la información que puedes obtener visualmente, ¿qué atributo/s crees que son los que mejor permitirán predecir el atributo class?

1.3.Aplicación de filtros

1. Aplicar el filtro filters/unsupervised/attribute/normalize sobre el conjunto de datos. ¿Qué efecto tiene este filtro?
2. Aplicar el filtro filters/unsupervised/instance/RemovePercentage sobre el conjunto de datos. ¿Qué efecto tiene este filtro?
3. Grabar el conjunto de datos como iris2.arff.
4. Aplicar el filtro filters/unsupervised/attribute/Discretize sobre el conjunto de datos. ¿Qué efecto tiene este filtro?

1.4.Visualización

Volver a cargar el conjunto de datos iris2.arff Pulsar la pestaña Visualize. Aumentar Point Size a 5 para visualizarlos datos mejor.

1. Aumentar el valor de Jitter: ¿qué efecto tiene?

2. Clasificación

El objetivo de este ejercicio es familiarizarse con las primeras técnicas de análisis de datos. En concreto, con los árboles de decisión.

2.1.Clasificador ZeroR

Cargar el conjunto de datos iris.arff. En la pestaña Classify, seleccionar el clasificador ZeroR. En las Test Options seleccionar Use trainning set, y pulsar el botón de Start para que genere el clasificador. En un instante, en la ventana de salida aparecerán los datos de la clasificación realizada. Analizar esta salida.

1. ¿Qué modelo genera el clasificador ZeroR?

   El modelo Full-training-set.

2. ¿Cuántas instancias del conjunto de entrenamiento clasifica bien?

   50 instancias.

3. ¿Qué porcentaje de instancias clasifica bien?

   33.3333 %.

4. ¿Qué crees que indica la matriz de confusión?

   El desempeño de algoritmo aplicado.

2.2.Clasificador J48

Cargar el conjunto de datos iris.arff. En la pestaña Classify, seleccionar el clasificador trees/j48. En las Test Options seleccionar Use trainning set, y pulsar el botón de Start para que genere el clasificador.

1. ¿Cuántas hojas tiene el árbol generado con J48?

   El árbol tiene 8 hojas.

2. ¿Cuántas instancias del conjunto de entrenamiento clasifica bien?

   147 instancias.

3. ¿Qué porcentaje de instancias clasifica bien?

   98 %.

4. Analizar la matriz de confusión: ¿Qué ha clasificado mal?

   Iris-versicolor.

   Iris-virginica.

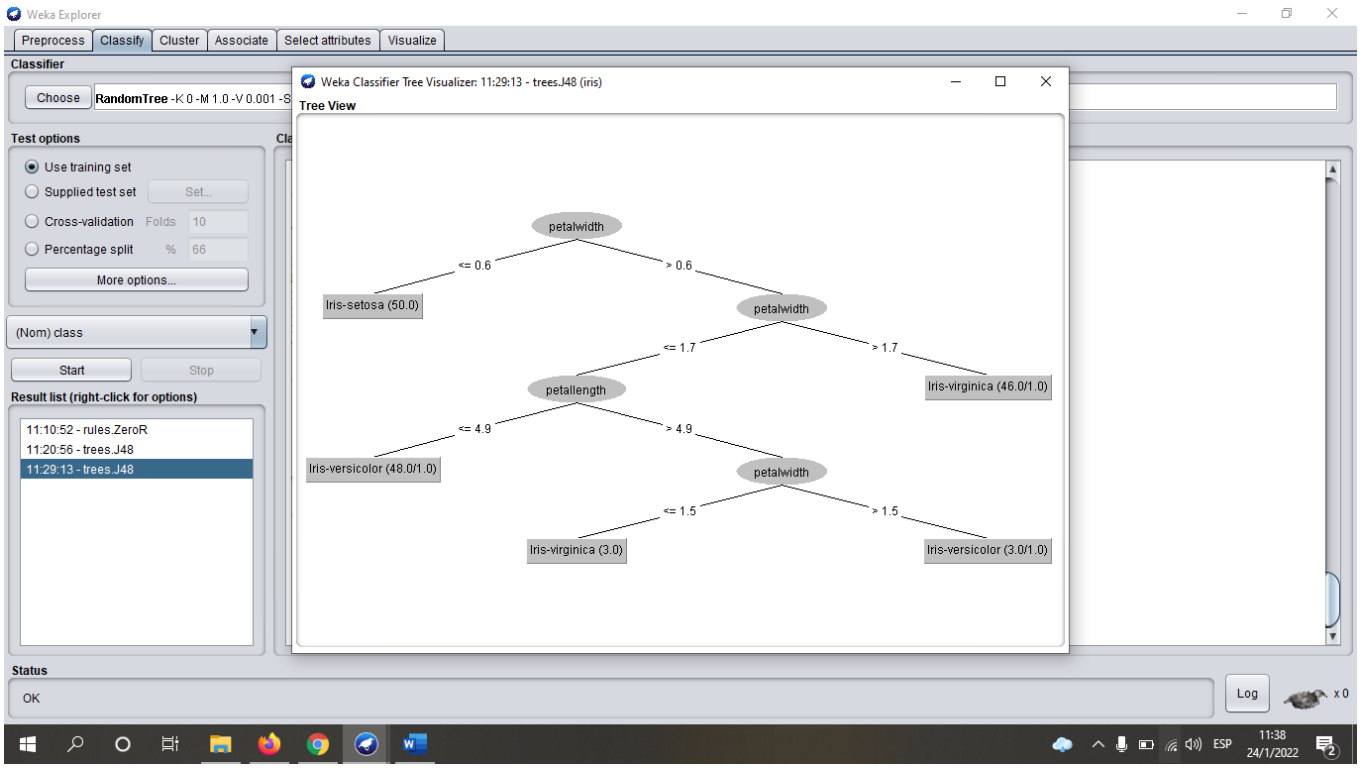
5. Pulsar el botón de More Options y seleccionar la opción de Output predictions. ¿En qué instancias se ha equivocado?

   Desde la instancia 51 hasta la 150.

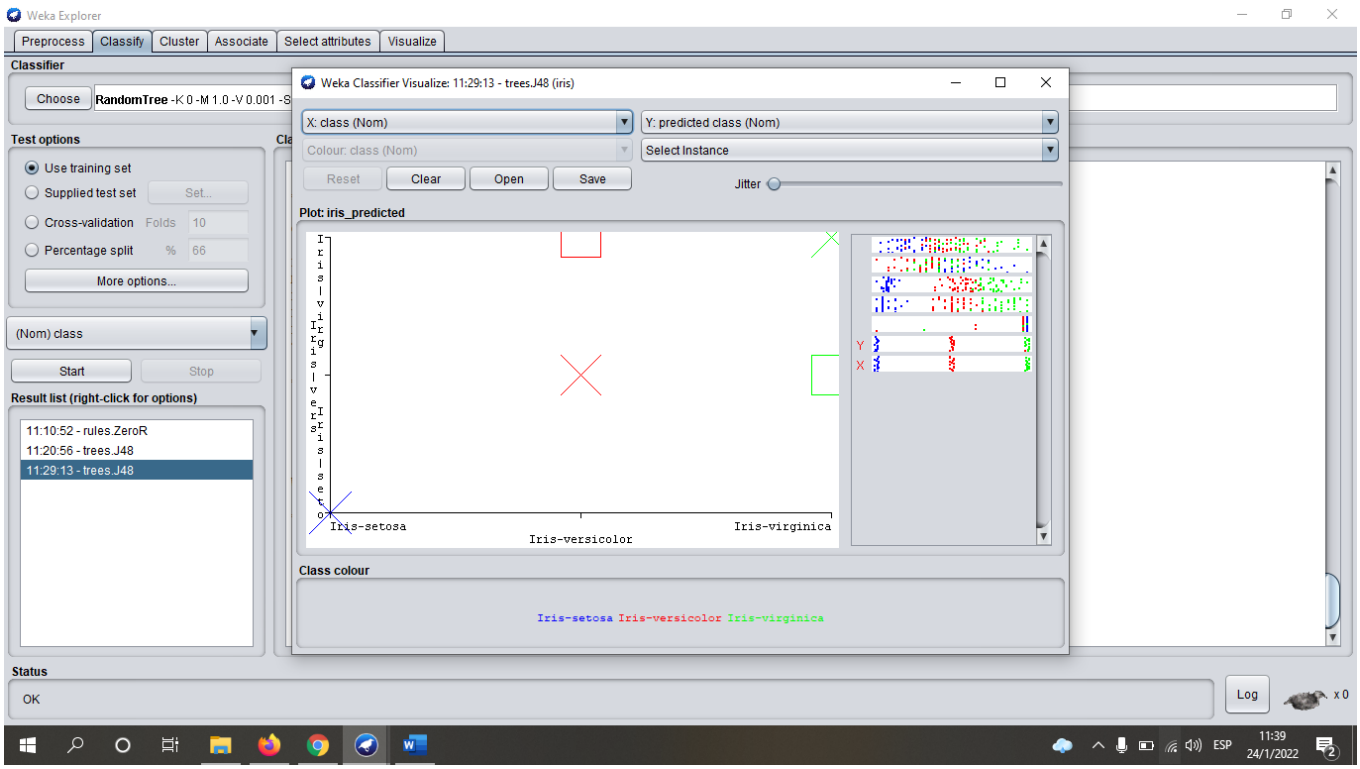
6. Elegir una instancia que J48 haya clasificado erróneamente y a analizar por qué.

Además, utiliza alguna de las herramientas de visualización de Weka:

En la ventana de Result list, pulsa en el botón derecho sobre el modelo generado con J48 para desplegar las opciones. Pulsa sobre Visualize Tree



En la ventana de Result list, pulsa en el botón derecho sobre el modelo generado con J48 para desplegar las opciones. Pulsa sobre Visualize Errors



2.3.Clasificador RandomTree

Cargar el conjunto de datos iris.arff. Seleccionar el clasificador ID3 y utilizarlo para generar un árbol de decisión.

1. ¿Has podido ejecutar el algoritmo RandomTree sobre el conjunto de datos directamente? ¿Por qué?
2. ¿Qué acciones has llevado a cabo para poder ejecutarlo?
3. ¿Qué porcentaje de éxito sobre el conjunto de entrenamiento has obtenido?

   Un porcentaje del 100 %.
4. ¿Qué porcentaje de éxito obtienes si utilizas como mecanismo de evaluación la validación cruzada?

   Del 1.
5. ¿Qué porcentaje de éxito estimas que obtendrás en el futuro sobre nuevos datos con el árbol generado con RandomTree?