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

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?
2. Si suponemos que queremos predecir el último atributo a partir de los anteriores, ¿estaríamos ante un problema de clasificación o 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*?
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?
2. ¿Cuántas instancias del conjunto de entrenamiento clasifica bien?
3. ¿Qué porcentaje de instancias clasifica bien?
4. ¿Qué crees que indica la matriz de confusión?

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?
2. ¿Cuántas instancias del conjunto de entrenamiento clasifica bien?
3. ¿Qué porcentaje de instancias clasifica bien?
4. Analizar la matriz de confusión: ¿qué ha clasificado mal?
5. Pulsar el botón de *More Options* y seleccionar la opción de *Output predictions*. ¿En qué instancias se ha equivocado?
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 ID3

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 ID3 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?
4. ¿Qué porcentaje de éxito obtienes si utilizas como mecanismo de evaluación la validación cruzada?
5. ¿Qué porcentaje de éxito estimas que obtendrás en el futuro sobre nuevos datos con el árbol generado con ID3?

2.4. Árboles de Regresión

Cargar el conjunto de datos *cpu.arff*. Entre los algoritmos *ID3*, *J48* y *M5P*, elegir uno de ellos para aproximar el atributo *class* sin que sea necesario tratar los datos de entrada de ninguna forma.

1. ¿Qué algoritmo has elegido? ¿por qué?
2. ¿Qué porcentaje de error obtienes si utilizas como mecanismo de evaluación la validación cruzada?
3. ¿Por qué no disponemos ahora de una matriz de confusión?

3. Agrupación

El objetivo de este ejercicio es familiarizarse con algunas técnicas de agrupación. Para ello, vamos a utilizar también el conjunto de datos *iris.arff*.

- Cargar el conjunto de datos *iris.arff*.
- Eliminar el atributo *class*
- Ejecutar el algoritmo *SimpleKMeans* para generar 3 grupos. ¿Qué medida de rendimiento genera Weka? ¿Qué valor proporciona?
- Ejecutar el algoritmo *SimpleKMeans* para generar 5 grupos. ¿Cómo mejora la medida de rendimiento?
- Utilizar la herramienta de visualización de grupos para comparar los dos resultados. ¿Puedes obtener alguna conclusión?
- Ejecutar el algoritmo *EM* con los parámetros por defecto. ¿Cuántas distribuciones genera? ¿Hay alguna relación con alguno de los resultados generados con *SimpleKMeans*?

4. El *Experimenter*

El objetivo de este ejercicio es familiarizarse con una herramienta avanzada de análisis de datos integrada en Weka, denominada *Experimenter*. Esta herramienta permite ejecutar distintos algoritmos de minería de datos sobre distintos conjuntos de datos, de forma que su ejecución secuencial hace más rápida su ejecución, así como la evaluación de los resultados.

Para ello, seguir los siguientes pasos:

- Pulsar el botón *New* para generar un nuevo experimento
- Seleccionar los conjuntos de datos: *iris.arff*, *soybean.arff* y *labor.arff*
- Seleccionar los clasificadores: *J48*,*IBK* con $K = 1$, *IBK* con $K = 3$, *IBK* con $K = 5$, y *SVO*
- En el apartado *Results Destination* seleccionar *CSV file* y utilizar el botón de *Browse* para elegir el fichero
- Pulsar la pestaña *Run* y pulsar el botón de *Start*

- Una vez finalizado el proceso, abrir una hoja de cálculo, y cargar el fichero CSV.
- En ese fichero, se muestra en cada fila los datos de cada ejecución, incluyendo el conjunto de datos, el clasificador utilizado con sus parámetros, así como datos sobre sus resultados
- Localizar la columna que mide el porcentaje de éxito
- Obtener la media del porcentaje de éxito para cada clasificador y conjunto de datos

Una vez realizados los pasos anteriores, responder a las siguientes preguntas:

- ¿Qué resultados ha obtenido cada clasificador en cada conjunto de datos?
- ¿Qué algoritmo ha obtenido mejores resultados en cada conjunto de datos?
- ¿Son los resultados del mejor algoritmo mucho mejores que los del resto?