Universidad Carlos III de Madrid

Aprendizaje Automático

Computer Science Engineering

Tutorial 2: Introducción a Weka

Authors:
Daniel Medina García
Alejandro Rodríguez Salamanca

February 18, 2016

Contents

1 Los ficheros de datos

1.1 ¿Cuántos atributos de entrada tiene el fichero de datos? ¿De qué tipo son?

Uno. Decir tipo y demás y por qué el de clase no es de entrada.

1.2 ¿Podría un algoritmo de aprendizaje automático identificar esa función con los datos que hay en ese fichero? ¿Por qué?

No porque mi respuesta es de libro.

2 Clasificar con ZeroR

2.1 ¿Qué resultado en términos de instancias correctas ofrece el algoritmo ZeroR?

Respuesta

2.2 ¿Qué ocurre si se selecciona otro algoritmo de clasificación permitido para ese conjunto de datos?

Respuesta

2.3 ¿Cuáles son las diferencias al repetir los pasos anteriores con el otro fichero `badges_plain.arff`?

Respuesta

2.4 ¿Qué ocurre si seleccionamos el algoritmo trees / ID3 en el segundo fichero?

Respuesta

3 Generando nuevos atributos

3.1 Propón 6 nuevos atributos y explica por qué los has elegido.

Respuesta

3.2 ¿Cuántos atributos tiene el fichero `badges1.arff` y de qué tipo son?

Respuesta

3.3 ¿Qué otro tipo de información estadística se muestra sobre los atributos? Tras pulsar el botón "Visualize all" indica qué se muestra y si hay algún atributo que no se visualice.

Respuesta

3.4 Genera un clasificador con ZeroR, ¿qué ocurre? Compara los resultados con los obtenidos en el ejercicio anterior.

Respuesta

3.5 Genera un clasificador con trees / ID3, ¿qué ocurre? ¿Qué se podría hacer para solucionar este problema?

No te deja ejecutarlo, por qué? Creo que tenía que ver con el tipo de datos

4 Clasificar con ID3: resolviendo problemas

4.1 ¿Qué información aparece en el desplegable tras abrir la pestaña *Capabilities*? ¿Qué información proporciona "more"?

Respuesta

Los atributos de entrada pueden modificarse a través de tareas de preprocesamiento. En los siguientes pasos vamos a modificar ciertos atributos de `badges1.arff` para que pueda clasificarse con ID3.

4.2 ¿Qué efecto tiene el filtro de discretización sobre el conjunto de datos con *bins* igual a 5?

Respuesta

4.3 ¿Cuántas instancias clasifica bien cuando marcamos *Use training set*? ¿Qué porcentaje representa? ¿Qué crees que indica la "matriz de confusión"? ¿Cuántas instancias de cada tipo se han clasificado mal?

Respuesta

Ahora se selecciona la opción *Outuput predictions*

4.4 ¿Cuál es la primera instancia del conjunto de entrenamiento que se clasifica mal? ¿Por qué?

Respuesta

4.5 ¿Cómo se clasificaría la instancia "Eloisa Figueroa"? ¿Cuáles son los atributos de este nombre? ¿Qué ocurre con los valores de esta instancia si utilizas el filtro usado anteriormente?

Respuesta

A continuación modificamos el fichero original introduciendo el nombre anterior y cambiamos la clase a "positiva", teniendo en cuenta que si contiene enumerados y se introduce un nuevo valor hay que especificarlo también en la definición de los valores posibles del enumerado. Después volvemos a generar el clasificador con ZeroR y training set seleccionado.

4.6 ¿Cómo se clasifica la instancia nueva?

Respuesta

5 Clasificar con J48 (C4.5)

Volvemos a la pestaña de preproceso para cargar `badges1.arff` y volver a generar el clasificador usando la opción de *training set.*

5.1 ¿Cuántas hojas tiene el árbol generado con J48?

Respuesta

5.2 ¿Cuántas instancias del conjunto de entrenamiento clasifica bien? ¿Qué porcentaje representa? ¿Cuántas instancias de cada tipo se han clasificado mal? ¿Cómo se clasificaría la instancia "Eloisa Figueroa"?

Respuesta

5.3 ¿Elegirías este modelo o el generado por ID3? ¿Por qué?

Respuesta

5.4 ¿Hemos encontrado la función exacta para generar las etiquetas? ¿Por qué lo sabes?

Respuesta

6 Utilizando más atributos con J48 (C4.5)

Es momento de volver a la pestaña de preproceso y generar un nuevo atributo que calcule el número de vocales. Después se grabará el conjunto de datos como `badges1-2.arff`, y con él se construirá un clasificador con J48. Una vez generado, se anotan el porcentaje de instancias bien clasificadas y la matriz de confusión, tras lo cual visualizaremos el árbol generado.

6.1 ¿Qué indican los números que aparecen en las hojas del árbol?

Respuesta

6.2 ¿Qué efecto tiene aumentar el valor de "Jitter" en la gráfica que relaciona el nuevo atributo con la clase?

Respuesta

6.3 ¿Podrías decir cuál es el rango de vocales más común en el fichero proporcionado? ¿Se te ocurre algún otro atributo relacionado que pueda aportar información?

Respuesta

6.4 Tras todos estos resultados, ¿qué características o cualidades crees que deben tener los atributos para maximizar el éxito de los algoritmos de aprendizaje automático?

Respuesta

7 Balanceado de datos, selección de características y otros filtros

Para este apartado se ha de cargar en Weka `adult-data.arff`.

7.1 ¿Cuántos atributos de entrada tiene este fichero? ¿Cuántas instancias de entrenamiento?

Respuesta

Ahora se ejecuta el clasificador ZeroR con *cross-validation*.

7.2 ¿Qué resultados aparecen? Explica el resultado.

Respuesta

A continuación se evalua el clasificador solamente con las instancias que figuren en el fichero `adult-test.arff`.

7.3 ¿Qué resultados aparecen? ¿Son estos resultados comparables a los anteriores? ¿Por qué?

Respuesta

Se repite este procedimiento con el J48 (en lugar de ZeroR) usando *cross-validation* y *Supplied test set*.

7.4 ¿Qué resultados aparecen? ¿Qué porcentaje de mejora ha obtenido respecto a los resultados del ZeroR?

Respuesta

7.5 ¿Qué proporción de datos hay de cada clase? ¿Crees que este porcentaje es apropiado para que un algoritmo de aprendizaje automático aprenda bien?

Respuesta

Se procede a modificar las instancias de entrenamiento para que tengan un porcentaje similar entre las dos clases.

7.6 ¿Qué ocurre con el atributo de salida? ¿Ha descendido el número de ejemplos de entrenamiento?

Respuesta

Tras aplicar este filtro, se evalúa de nuevo con *cross-validation* y supplied test set los algoritmos ZeroR y J48.

7.7 ¿Qué resultados dan los algoritmos? ¿Qué resultado crees que es mejor? ¿Por qué?

Respuesta

Se eliminan uno o dos atributos que no se creen útiles para el algoritmo de aprendizaje.

7.8 ¿Cuáles has eliminado? ¿Por qué? ¿Qué es lo que ocurre al repetir la evaluación anterior?

Respuesta

Por último se aplica el filtro de normalización para los atributos numéricos.

7.9 ¿Qué resultados se obtienen?

Respuesta

7.10 Después del procesamiento de datos que has realizado en este apartado, ¿crees que esto ayuda al proceso de aprendizaje? ¿Por qué? ¿Cuál es el mejor resultado obtenido? Justifícalo.

Respuesta