

UNIVERSIDAD CARLOS III DE MADRID

APRENDIZAJE AUTOMÁTICO

COMPUTER SCIENCE ENGINEERING

Tutorial 2: Introducción a Weka

Authors:

Daniel MEDINA GARCÍA

Alejandro RODRÍGUEZ SALAMANCA

February 21, 2016

Contents

1	Los ficheros de datos	3
1.1	¿Cuántos atributos de entrada tiene el fichero de datos? ¿De qué tipo son?	3
1.2	¿Podría un algoritmo de aprendizaje automático identificar esa función con los datos que hay en ese fichero? ¿Por qué?	3
2	Clasificar con ZeroR	4
2.1	¿Qué resultado en términos de instancias correctas ofrece el algoritmo ZeroR? . .	4
2.2	¿Qué ocurre si se selecciona otro algoritmo de clasificación permitido para ese conjunto de datos?	4
2.3	¿Cuáles son las diferencias al repetir los pasos anteriores con el otro fichero, <code>badges_plain</code> ?	4
2.4	¿Qué ocurre si seleccionamos el algoritmo <code>trees</code> / ID3 en el segundo fichero? . . .	4
3	Generando nuevos atributos	5
3.1	Propón 6 nuevos atributos y explica por qué los has elegido.	5
3.2	¿Cuántos atributos tiene el fichero <code>badges1</code> y de qué tipo son?	5
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.	5
3.4	Genera un clasificador con ZeroR, ¿qué ocurre? Compara los resultados con los obtenidos en el ejercicio anterior.	5
3.5	Genera un clasificador con <code>trees</code> / ID3, ¿qué ocurre? ¿Qué se podría hacer para solucionar este problema?	5
4	Clasificar con ID3: resolviendo problemas	6
4.1	¿Qué información aparece en el desplegable tras abrir la pestaña <i>Capabilities</i> ? ¿Qué información proporciona “more”?	6
4.2	¿Qué efecto tiene el filtro de discretización sobre el conjunto de datos con <i>bins</i> igual a 5?	6
4.3	¿Cuántas instancias clasifica bien usando un conjunto de entrenamiento? ¿Qué porcentaje representa? ¿Qué crees que indica la “matriz de confusión”? ¿Cuántas instancias de cada tipo se han clasificado mal?	6
4.4	¿Cuál es la primera instancia del conjunto de entrenamiento que se clasifica mal cuando se usa la opción de predicción de salida? ¿Por qué?	6
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?	6
4.6	¿Cómo se clasifica la instancia nueva?	6
5	Clasificar con J48 (C4.5)	7
5.1	¿Cuántas hojas tiene el árbol generado con J48?	7
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”?	7

5.3	¿Elegirías este modelo o el generado por ID3? ¿Por qué?	7
5.4	¿Hemos encontrado la función exacta para generar las etiquetas? ¿Por qué lo sabes?	7
6	Utilizando más atributos con J48 (C4.5)	8
6.1	¿Qué indican los números que aparecen en las hojas del árbol?	8
6.2	¿Qué efecto tiene aumentar el valor de “Jitter” en la gráfica que relaciona el nuevo atributo con la clase?	8
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?	8
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?	8
7	Balanceado de datos, selección de características y otros filtros	9
7.1	¿Cuántos atributos de entrada tiene este fichero? ¿Cuántas instancias de entrenamiento?	9
7.2	¿Qué resultados aparecen ejecutando ZeroR con validación cruzada? Explica el resultado.	9
7.3	¿Qué resultados aparecen sólo con las instancias del fichero <code>adult-test.arff</code> ? ¿Son estos resultados comparables a los anteriores? ¿Por qué?	9
7.4	¿Qué resultados aparecen repitiendo el proceso con J48? ¿Qué porcentaje de mejora ha obtenido respecto a los resultados del ZeroR?	9
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?	9
7.6	¿Qué ocurre con el atributo de salida cuando se modifican las instancias de entrenamiento para tener un porcentaje similar entre las dos clases? ¿Ha descendido el número de ejemplos de entrenamiento?	9
7.7	¿Qué resultados dan los algoritmos? ¿Qué resultado crees que es mejor? ¿Por qué?	9
7.8	¿Qué atributos consideras inútiles para el algoritmo de aprendizaje y por qué? Elimina un par de ellos. ¿Qué es lo que ocurre al repetir la evaluación anterior?	9
7.9	¿Qué resultados se obtienen aplicando el filtro de normalización para los atributos numéricos?	10
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.	10

1 Los ficheros de datos

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

`badges` Este archivo contiene un atributo de entrada de tipo String que se corresponde con un nombre, y un atributo de salida, la clase a la que pertenece.

`badges_plain` El contenido de este archivo es el mismo que el de `badges` con la diferencia de que los nombres están previamente enumerados. [más?]

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

No. El fichero contiene tan solo un campo de datos, de tipo enumerado, y cada posible valor de dicho campo sólo se encuentra en una instancia. Al ser una clase con dos valores (muchos menos que valores del enumerado) y ser todas las instancias distintas, toda información que saque un algoritmo de aprendizaje automático será meramente memorización, sin valor alguno en futuras instancias con valores distintos a los actuales ya que no existe razonamiento o computación en la resolución de la clase.

2 Clasificar con ZeroR

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

ZeroR indica un resultado del 51.0204% de instancias correctas. Lo único de lo que nos informa este dato es de que las clases están equilibradas dentro de nuestro conjunto de datos.

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

Utilizando meta / MultiScheme obtenemos exactamente el mismo resultado que con ZeroR. Esto es lógico pues no se puede mejorar el resultado al carecer de información suficiente para razonar sobre los datos.

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

Los mismos, pues aunque cambiemos el tipo de campo de enumerado a String se sigue sin tener información para razonar sobre los datos.

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

Si probamos con ID3 podemos observar que el porcentaje de instancias correctamente clasificadas es del 100%. Si usásemos validación cruzada seguramente los resultados se parezcan a los obtenidos con ZeroR, ya que se sigue sin tener información para clasificar más allá de la memorización y por ello no se tiene un criterio razonado para clasificar nuevas instancias.

3 Generando nuevos atributos

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

Al tener como datos de entrada nombres propios, sólo podemos basarnos en la naturaleza de las palabras que los componen para realizar algún tipo de razonamiento. Así, proponemos algunos nuevos campos a partir del original que detallan las características de dichos nombres:

- Número de letras del nombre.
- Número de vocales.
- Número de consonantes.
- Número de espacios.
- Inicial del nombre o del apellido.
- Número de repeticiones de cada una de las letras que aparecen.

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

Este archivo contiene campos que persiguen lo mismo que planteábamos en la pregunta anterior: descubrir más datos sobre los nombres. Así, el archivo contiene siete nuevos campos de entrada además del nombre de tipos numérico y enumerado.

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` 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 usando un conjunto de entrenamiento? ¿Qué porcentaje representa? ¿Qué crees que indica la “matriz de confusión”? ¿Cuántas instancias de cada tipo se han clasificado mal?

Respuesta

- 4.4 ¿Cuál es la primera instancia del conjunto de entrenamiento que se clasifica mal cuando se usa la opción de predicción de salida? ¿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 y volver a generar el clasificador usando un conjunto de entrenamiento

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`, 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 carga en Weka `adult-data`.

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

Respuesta

7.2 ¿Qué resultados aparecen ejecutando ZeroR con validación cruzada? Explica el resultado.

Respuesta

7.3 ¿Qué resultados aparecen sólo con las instancias del fichero `adult-test.arff`? ¿Son estos resultados comparables a los anteriores? ¿Por qué?

Respuesta

7.4 ¿Qué resultados aparecen repitiendo el proceso con J48? ¿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

7.6 ¿Qué ocurre con el atributo de salida cuando se modifican las instancias de entrenamiento para tener un porcentaje similar entre las dos clases? ¿Ha descendido el número de ejemplos de entrenamiento?

Respuesta

Tras aplicar este filtro, se evalúan de nuevo con validación cruzada 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

7.8 ¿Qué atributos consideras inútiles para el algoritmo de aprendizaje y por qué? Elimina un par de ellos. ¿Qué es lo que ocurre al repetir la evaluación anterior?

Respuesta

7.9 ¿Qué resultados se obtienen aplicando el filtro de normalización para los atributos numéricos?

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? Justificalo.

Respuesta