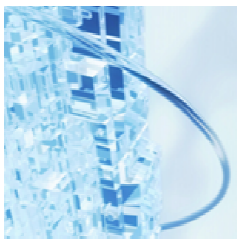


Prácticas de Minería de Datos

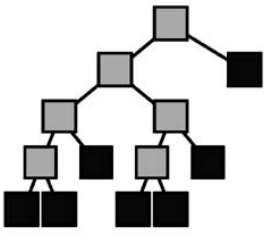
Grado en Ingeniería Informática

Curso 2013-14

PRÁCTICA 4

Clasificación

Almacenamiento de modelos y predicción de nuevos casos



OBJETIVOS

- Aprender a construir diferentes clasificadores e interpretar los resultados obtenidos.
- Aprender a guardar un modelo para poder predecir nuevas instancias sin clasificar.

1. Introducción

La utilización del **aprendizaje automático** para la tarea de **clasificación** comprende dos pasos:

- **Entrenamiento**: se parte de unos datos conocidos o ejemplos de entrenamiento que se utilizan para construir un modelo (por ejemplo, un árbol de decisión)

- **Clasificación**: una vez creado el modelo con los datos conocidos, se utiliza dicho modelo para clasificar nuevos datos, para los que no se conoce la clase

En esta práctica generaremos modelos de clasificación y aprenderemos a interpretar los resultados y, además, almacenaremos los modelos creados y los utilizaremos para clasificar nuevos ejemplos.

2. Primera parte: interpretación de la evaluación

Utilizaremos la base de datos ***vehicle.arff***.

Vamos a construir 3 modelos de clasificación, utilizando 10-validación cruzada:

- Un árbol de decisión usando el algoritmo **C4.5** (J48 en Weka)

- Un segundo **árbol** de decisión, **a seleccionar** de entre los ofertados por Weka, que cumpla los requisitos para la base de datos ***vehicle.arff***.

- La regla generada por el algoritmo **1R** (OneR en Weka).

Estudiando el informe de evaluación, responder a las siguientes preguntas:

1. ¿Qué modelo obtiene mejor tasa de acierto?

    - Indica para cada modelo cuántas instancias clasifica correctamente, cuantas erróneamente, la tasa de acierto y la de error

2. ¿Qué clasificador obtiene mejores resultados para la clase “saab”?, ¿cuál es el indicador?

    - Explica qué significa cada uno de las siguientes medidas y la fórmula para calcularla: TP Rate, FP Rate, Precision, Recall, F-Measure.

- Realiza los cálculos de forma manual de las medidas anteriores para la marca “opel” en el modelo **C4.5** (J48 en Weka)

3. Para el modelo surgido del J48, ¿cuántos coches de tipo “van” son clasificados de tipo “opel” y de tipo “saab”?

   - Indica para cada marca de coche cuántos son correctamente clasificados en cada modelo de clasificación

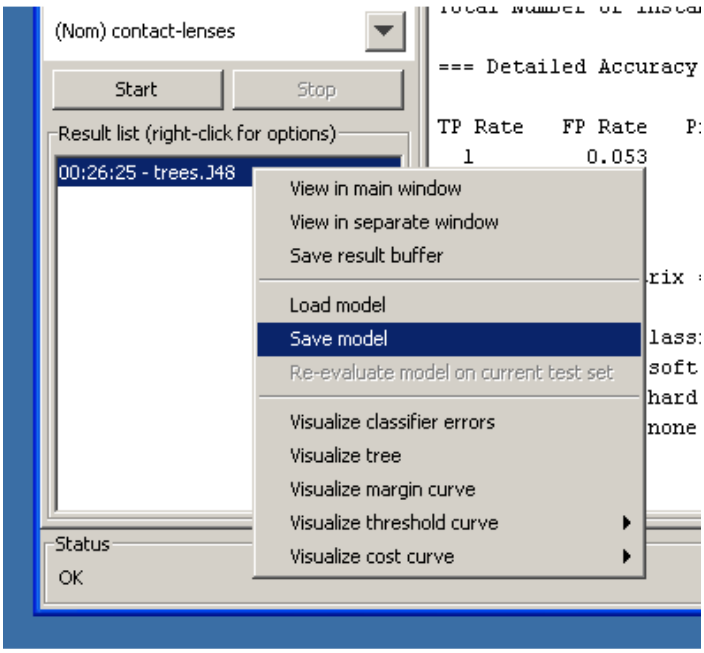
4. Para todos los modelos, ¿qué porcentaje de errores cometen para la clase “opel”?

3. Segunda parte: almacenamiento del modelo y predicción de nuevos casos

Veamos el problema de recomendar un tipo de lentillas dependiendo ciertas variables de entrada (edad, tipo de problema visual, existencia de astigmatismo y nivel de producción de lágrimas). Esta información se encuentra almacenada en la base de datos ***contact-lenses.arff***.

El objetivo es generar un modelo que se ajuste a los ejemplos de entrenamiento y, posteriormente, usar dicho modelo para clasificar nuevos ejemplos.

Para ello, una vez generado el modelo, lo almacenaremos pulsando con el botón derecho del ratón en la ventana que muestra la lista de resultados (*Result list*) y eligiendo la opción “*Save model*”



Una vez almacenado el modelo, podemos clasificar nuevos ejemplos. En primer lugar, se debe crear un fichero **.arff** con el ejemplo (o los ejemplos) a clasificar.

Dado que la clase de estos nuevos ejemplos es desconocida, se indica con una interrogación.

Ejemplo:

```
@relation test
@attribute age {young, pre-presbyopic, presbyopic}
@attribute spectacle-prescrip {myope, hypermetrope}
@attribute astigmatism {no, yes}
@attribute tear-prod-rate {reduced, normal}
@attribute contact-lenses {soft, hard, none}

@data
young, hypermetrope, no, normal, ?
presbyopic, myope, yes, normal, ?
```

Es conveniente guardar este fichero en la misma ruta donde se encuentre el modelo.

Para obtener la predicción de nuestro fichero de test, debemos realizar los siguientes pasos:

1. Cargar el modelo pulsando con el botón derecho del ratón en la ventana que muestra la lista de resultados (*Result list*) y eligiendo la opción "*load model*"

2. Seleccionar "*Supplied test set*" en la ventana de "*Test options*" y elegir el fichero .arff con las instancias a clasificar.

3. Pulsar el botón "*More options…*"

4. Deseleccionar todas las opciones menos la de "*Output predictions*".

5. Pulsando con el botón derecho del ratón en la ventana que muestra la lista de resultados (*Result list*) sobre el modelo cargado, elegir la opción "*Re-evaluate model on current test set*"

La salida mostrada debe ser algo parecido a esto:

```
=== Predictions on test set ===

 inst#,    actual, predicted, error, probability distribution
     1          ?     1:soft      +  *0.833  0      0.167
     2          ?     2:hard      +   0     *1      0

=== Evaluation on test set ===
=== Summary ===

Total Number of Instances               0
Ignored Class Unknown Instances                 2
```

Este informe muestra que la primera instancia la ha clasificado como "*soft*" y la segunda como "*hard*". La columna *error* muestra un símbolo "+" cuando hay error entre la clase predicha y la real (en nuestro caso siempre será "+" puesto que la clase real es desconocida).

Ejercicio para la segunda parte

La base de datos ***credit-g_simple.arff*** contiene 1000 ejemplos que representan clientes de una entidad bancaria que demandaron un crédito. Tiene 7 atributos numéricos y 13 nominales. La clase es binaria e indica si el cliente puede ser considerado como fiable para concederle el crédito o no.

1. Para poder predecir si un cliente es fiable sólo disponemos de cierta información, así que nos interesa construir un modelo en base a dicha información. En concreto, conocemos:

   - duración del crédito (en meses)
   - propósito del crédito
   - cantidad que solicita
   - estado civil
   - edad
   - situación laboral

2. Construir tres modelos de clasificación usando los clasificadores de la primera parte. Utilizar 10-validación cruzada para construir el modelo.

3. Almacenar los modelos con nombres distintos

4. Utilizando los tres modelos, predecir si sería conveniente concederles un crédito a estos 3 clientes:

12	Vacaciones	1500	Hombre soltero	Trabajador no cualificado	37	¿?
48	Electrónica de consumo	3000	Hombre casado	Trabajador no cualificado	25	¿?
6	Negocios	6500	Mujer soltera	Trabajador cualificado	45	¿?

Para: J48

12	Vacaciones	1500	Hombre soltero	Trabajador no cualificado	37	
48	Electrónica de consumo	3000	Hombre casado	Trabajador no cualificado	25	
6	Negocios	6500	Mujer soltera	Trabajador cualificado	45	

Para: 1R

12	Vacaciones	1500	Hombre soltero	Trabajador no cualificado	37	
48	Electrónica de consumo	3000	Hombre casado	Trabajador no cualificado	25	
6	Negocios	6500	Mujer soltera	Trabajador cualificado	45	

Para:

12	Vacaciones	1500	Hombre soltero	Trabajador no cualificado	37	
48	Electrónica de consumo	3000	Hombre casado	Trabajador no cualificado	25	
6	Negocios	6500	Mujer soltera	Trabajador cualificado	45	

5. Realizar un pequeño informe con las conclusiones e indica, en base, a las predicciones anteriores a quien o quienes les concederías el crédito y por qué.

¿Cómo entregar la práctica?

- Utilizar un documento de texto para responder a las cuestiones y subirlo a través de la plataforma web