

Práctica 2 de MINERIA DE DATOS 09/10

Clasificación mediante la herramienta WEKA

Duración: 2 Sesiones

Fecha límite de entrega: 30 de Noviembre

José A. Gámez, M. Julia Flores & Pablo Bermejo

16/11/2009

1. Objetivo de la práctica

En la práctica anterior se ha hecho especial énfasis en el preprocesamiento, considerando sólo dos clasificadores y además tratándolos como cajas negras, es decir, sin considerar sus parámetros de configuración (si es que los había). En esta práctica movemos el foco de atención hacia la tarea de clasificación. Así, el **objetivo de la práctica** es aplicar las técnicas de clasificación estudiadas en clase sobre una base de datos (ver la siguiente sección), trabajando los distintos parámetros de los clasificadores para obtener un mejor resultado, pero obviamente, no nos olvidaremos del preprocesamiento que como ya hemos visto puede mejorar el rendimiento de un clasificador de forma notoria.

2. Base de datos considerada

Como base de datos sobre la que se experimentarán distintos clasificadores utilizaremos la misma base de datos que en la Práctica 1. Más concretamente, cada grupo utilizará la base de datos preprocesada que mejores resultados le haya dado. Así, en la memoria no olvidéis describir la base de datos utilizada: número de atributos, tipo de atributos, número de instancias, ratio instancias positivas/negativas, valores perdidos, atributos construidos, ...

3. ¿Qué hay que hacer?

Básicamente probar los clasificadores vistos en clase más uno extra como se detallará a continuación:

1. Listas de decisión. Se usará el algoritmo visto en clase que en WEKA aparece como PRISM (carpeta *rules*).
2. Árboles de clasificación. Se usará C4.5 o J48 como se denomina en WEKA (carpeta *trees*).
Para usarlo tal y como se ha estudiado en clase debes configurar *subtreeraising = false*, *binarySplit = true* y poner a 1 *minNumObj* y *numFolds*.
3. Naïve Bayes (carpeta Bayes) con las opciones por defecto. Como comprobarás usa la suavización por Laplace.

4. Clasificador bayesiano TAN. Para seleccionar este método ve a la carpeta Bayes y elige BayesNet. Haz click en la caja correspondiente a BayesNet para abrir sus opciones, y en la pestaña de `SearchAlgorithm` selecciona `local->TAN`.
5. Métodos perezosos o basados en instancias. La implementación del método de los vecinos más próximos (kNN) disponible en WEKA la encontrarás en la carpeta `Lazy ->IBk`. Como verás está parametrizado el número de vecinos a considerar (k) pero también la forma en que se pesa la distancia (*distanceWeighting*).
6. El algoritmo de redes neuronales estudiado en clase (backpropagation) lo encontrarás en la carpeta `function` bajo el nombre de `MultilayerPerceptron`.
Inicialmente usálo con los parámetros por defecto excepto quizás el número de epoch (*trainingTime*). Después usa la ayuda (botón *more*) para realizar algunas pruebas variando la estructura de la red o algunos parámetros.
7. Aparte de los clasificadores anteriores debes elegir otro clasificador no estudiado en clase y usarlo, deberás describir su fundamento e ideas principales usando unas dos páginas en la memoria.
8. Adicionalmente puedes considerar los multclasificadores **Bagging** (implementa la técnica de bagging) y **AdaBoostM1** (implementa la técnica de boosting), ambos en la carpeta `meta`.

Como se puede comprobar, muchos de estos clasificadores son configurables en función de una serie de parámetros. Es parte del objetivo de esta práctica *jugar* con ellos para ver si se mejora la configuración inicial.

La idea es realizar una comparativa sobre los resultados obtenidos para cada base de datos, comparando entre los distintos clasificadores, pero también entre las distintas configuraciones para cada clasificador. Como medida de evaluación principal se usará la tasa de acierto en una validación cruzada de cinco folds (5cv), aunque también se debe tener en cuenta la complejidad de los modelos obtenidos, el tiempo necesario para aprender, para validar, etc.

Como resumen final debe aparecer una tabla para cada base de datos como la siguiente:

Clasificador	Preprocesamiento	Parámetros	Tiempo CPU	Tasa de Acierto (5cv)
PRISM				
J48				
NB				
TAN				
IBK				
NN				
Algoritmo-elegido				
Bagging (opcional)				
Boosting (opcional)				
MEDIA				

donde preprocesamiento debe indicar brevemente si se ha hecho algún preprocesamiento (aparte del preprocesamiento inicial de vuestra base de datos, que deberéis describir antes de los experimentos), por ejemplo, (dis.4bins, 7vars-filter) indicaría que se ha realizado una discretización en cuatro bins, y que se han seleccionado 7 variables usando un método filter. Como esta entrada debe corresponder a la mejor para el algoritmo concreto (p.e. J48) su descripción completa se habrá realizado ya previamente. Parámetros debe contener los valores a los que se han fijado los parámetros configurables del algoritmo, si es que tiene.

Por último, como veis aparece una fila al final para la media, esto es por que queremos que trabajeis SOBRE TODOS los clasificadores bien mediante sus parámetros o bien mediante preprocesamiento, y por eso le daremos mucho valor al rendimiento medio obtenido, no a que se trabaje solo en las opciones de c4.5 y se aplique NB tal cual.