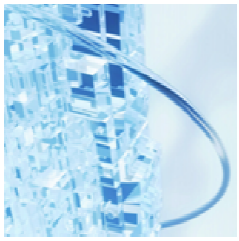


Prácticas de Minería de Datos

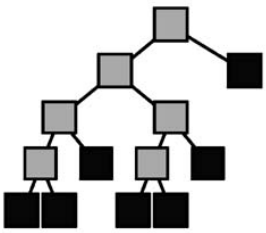
Grado en Ingeniería Informática

Curso 2013-14

PRÁCTICA 6

Clasificación

Combinación de modelos y métodos



OBJETIVO

- Familiarizarse con el uso de las principales técnicas de combinación de modelos y métodos de aprendizaje.

1. Introducción

La calidad de las predicciones obtenidas con una determinada técnica o técnicas aplicadas a un modelo o problema concreto puede mejorarse mediante la combinación de diferentes modelos y/o diferentes métodos de aprendizaje sobre dicho problema, dando lugar a dos técnicas principales:

- **Combinación de modelos**: a partir de un problema se generan diferentes modelos que son aprendidos con el **mismo** método de aprendizaje y se realiza una predicción conjunta resultante de la combinación de todos ellos.

- **Combinación de métodos**: a partir de un problema se generan diferentes modelos que son aprendidos con **diferentes** métodos de aprendizaje y se realiza una predicción conjunta resultante de la combinación de todos ellos.

En esta práctica generaremos modelos de clasificación haciendo uso de las principales técnicas de combinación de modelos y métodos de aprendizaje que implementa Weka.

2. Combinación de modelos (Bagging y Boosting)

Los métodos de combinación de modelos (metamodelos o multiclasificadores) se basan en la idea de usar un conjunto de modelos diferentes **aprendidos con el mismo método de aprendizaje** y combinarlos para crear así un nuevo modelo con el que realizar una predicción conjunta, que mejora los resultados obtenidos con los modelos de forma individual, mejorando de esta forma la calidad de las predicciones obtenidas mediante la combinación de las predicciones de varios modelos.

Los métodos de combinación de modelos implementan la estrategia de “usar todas las hipótesis posibles” o “dos ven más que uno”, aunque también es cierto que para que esta afirmación se cumpla los dos deben tener buena vista, y además no deben tener un comportamiento idéntico, ya que en ese caso no habría mejora.

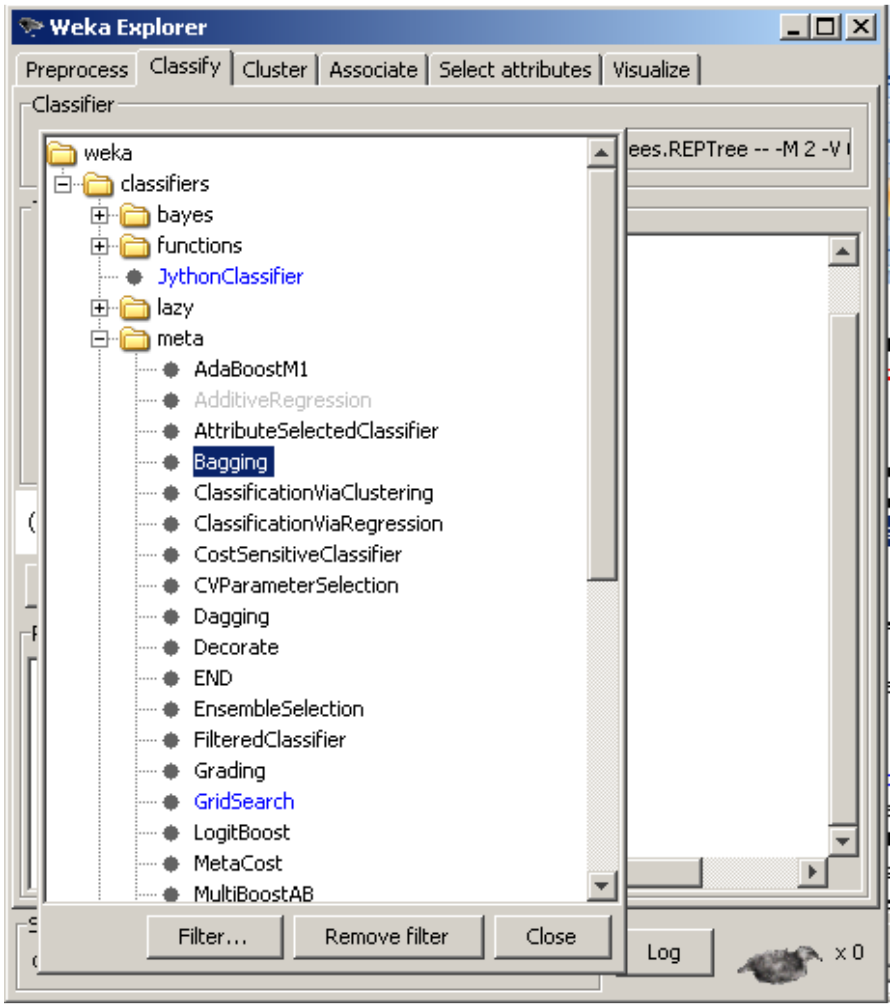
La calidad del metamodelo obtenido viene determinada por dos factores:

- Los modelos a combinar (deben ser precisos y suficientemente diferentes entre sí).
- La forma de combinar los modelos

La ventaja que tiene esta técnica es que por regla general se mejora la precisión obtenida por cada uno de los modelos de forma individual, aunque también tiene desventajas. Principalmente, el modelo se complica, perdiendo comprensibilidad. Además, necesitamos más recursos (tiempo y memoria) para el aprendizaje y utilización de modelos combinados.

Uno de los aspectos más destacados de Weka es la gran cantidad de métodos de combinación de modelos que posee.

Para combinar modelos accedemos en WEKA al explorador y seleccionamos la pestaña *Classify*.



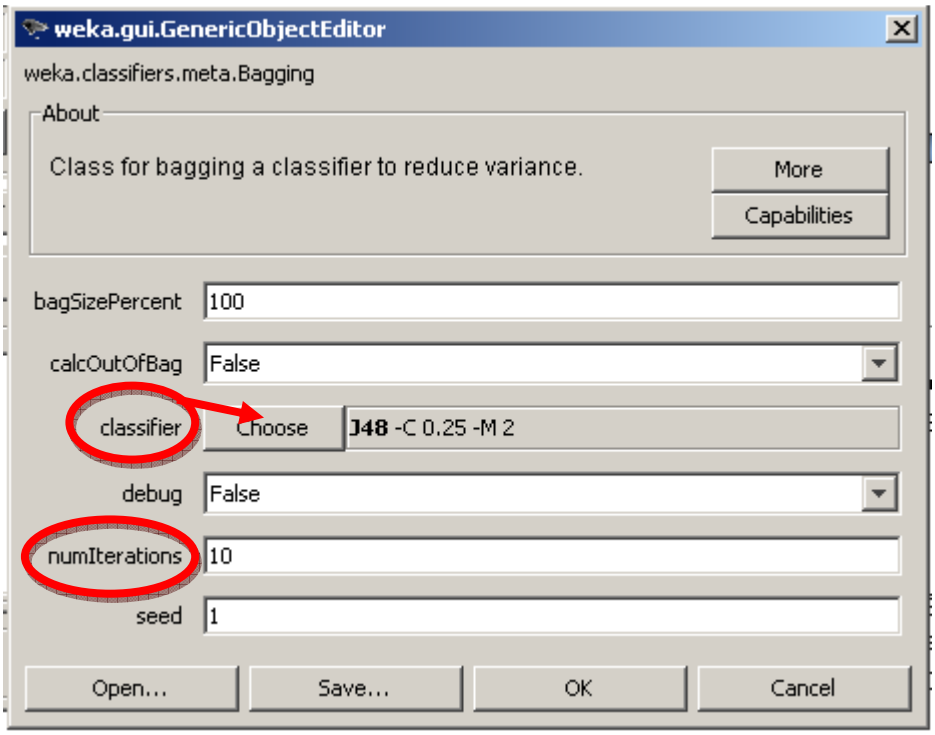
2.1 Bagging

Vamos a empezar con el método *Bagging*. Esta técnica se fundamenta en construir un conjunto de *n* modelos mediante el aprendizaje desde *n* conjuntos de datos. Cada conjunto de dato se construye realizando un muestreo con repetición del conjunto de datos de entrenamiento.

Es decir, a la hora de aprender un modelo a partir de un conjunto de datos de entrenamiento, *Bagging* genera n subconjuntos de entrenamientos aleatorios con reposición a partir del conjunto de entrenamiento original y para cada uno de los n subconjuntos genera un modelo distinto utilizando el mismo método, los cuales son combinados para realizar una predicción conjunta.

Es útil para algoritmos de aprendizaje débiles (redes neuronales, arboles de decisión, inducción de reglas), que son aquellos que son muy sensibles al conjunto de entrenamiento (generan modelos muy diferentes ante pequeñas alteraciones en el conjunto de entrenamiento).

Para seleccionar *Bagging*, pulsamos *Choose* y en *Meta*, tendremos el método *Bagging*. En este caso las opciones más importantes son *numIterations* donde marcamos el número de modelos base que queremos crear, y *Classifier*, donde seleccionamos el método base con el cual deseamos crear los modelos base.



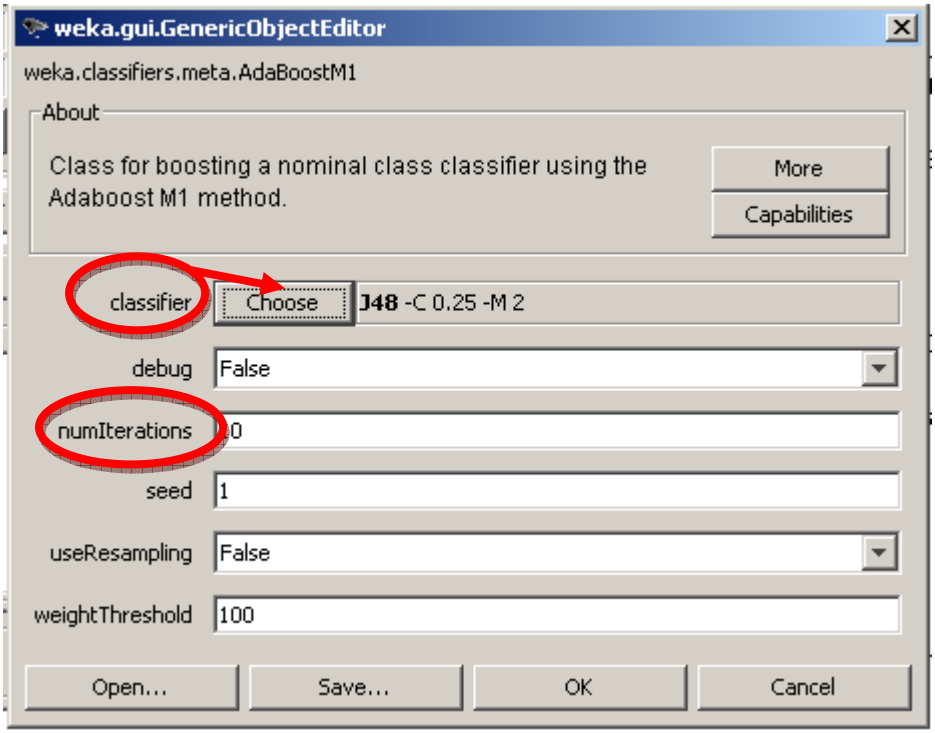
Si lanzamos WEKA, se observa que obviamente le cuesta más realizar el aprendizaje, en realidad tantas veces más como *numIterations* hayamos indicado.

2.2 Boosting

Veamos qué pasa con *Boosting*. Esta técnica es bastante parecida a *Bagging* (genera n subconjuntos de entrenamientos y para cada uno genera un modelo distinto utilizando el mismo método), aunque utiliza una estrategia más ingeniosa ya que cada iteración intenta corregir los errores cometidos anteriormente dando más peso a los datos que se han clasificado erróneamente.

Boosting asocia un peso a cada ejemplo del conjunto de entrenamiento y en cada iteración el modelo generado minimiza el número de ejemplos mal clasificados incrementando el peso de esos ejemplos y reduciendo el de los ejemplos bien clasificados con idea de que se tenga más en cuenta esos ejemplos incorrectamente clasificados a la hora de realizar el aprendizaje.

Para seleccionar *Boosting*, seleccionamos *AdaBoostM1* en *Meta*. En este método las opciones más importantes son también *numIterations* donde marcamos el número de iteraciones máximas (es decir de modelos base), y *Classifier*, donde seleccionamos el método base con el cual deseamos crear los modelos base.



Al lanzar el aprendizaje volvemos a comprobar que el proceso es más costoso, en tiempo, que los métodos simples.

3. Combinación de métodos (Métodos híbridos o Meta aprendizaje)

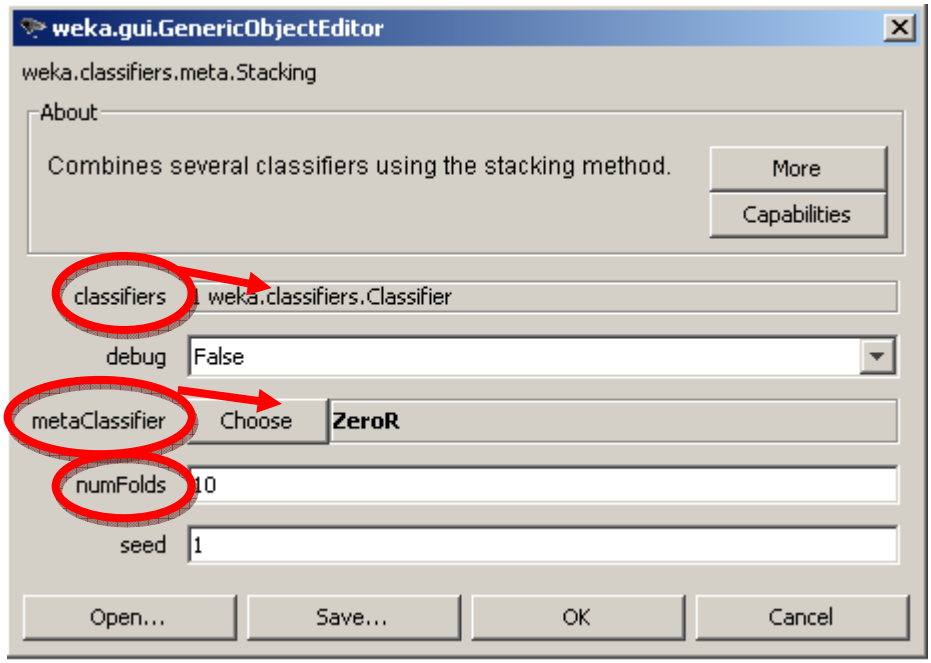
Las técnicas de combinación de modelos se centran en combinar modelos aprendidos con el mismo método de aprendizaje para realizar una predicción conjunta. Es decir, generan a partir del conjunto de entrenamiento original, n subconjuntos de datos de entrenamientos aleatorios con reposición y utilizan **un mismo método de aprendizaje** para generar los diferentes modelos a partir de dichos subconjuntos, para combinarlos en una nueva predicción conjunta.

El problema que presentan es que si el método empleado no es apropiado al problema, por más combinaciones que hagamos el resultado no mejorará ostensiblemente.

Como alternativa surgen los métodos híbridos o de **combinación de métodos**. La idea al igual que antes es generar a partir del conjunto de entrenamiento original, n subconjuntos de datos de entrenamientos aleatorios con reposición y utilizar **distintos métodos de aprendizaje** para generar los diferentes modelos a partir de dichos subconjuntos, para combinarlos en una nueva predicción conjunta.

La **combinación de métodos** busca combinar las predicciones alcanzadas con distintos modelos, generados por métodos diferentes, mediante un nuevo clasificador que realiza meta-aprendizaje con las predicciones de cada modelo, seleccionando para cada clase, aquella generada por el método más preciso.

Para hacer esto en WEKA, una vez seleccionado el explorador y la pestaña *Classify*, seleccionados en *meta* el método *Stacking*.

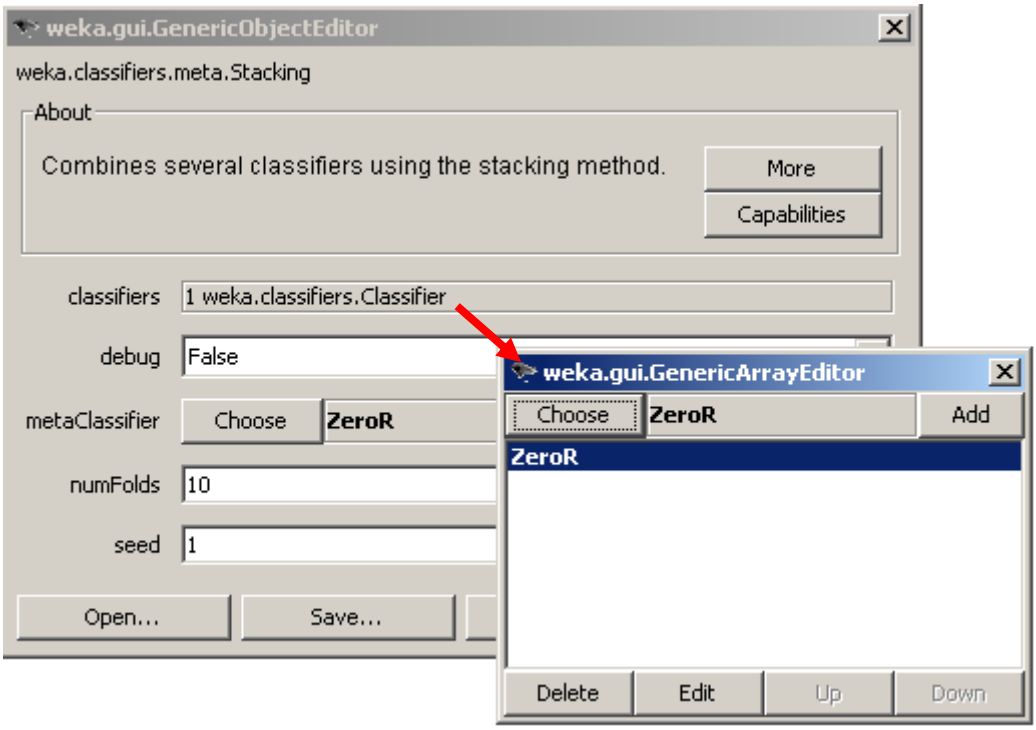


En este método las opciones más importantes son:

- *classifiers*, donde deberemos ir añadiendo (*Add*) y configurando los distintos métodos que se pretenden combinar. Para ello una vez hacemos click en *classifiers* aparece una ventana donde podremos ir eligiendo (*Choose*) los distintos métodos que pretendemos combinar (una vez elegido un método, lo añadimos a la lista de métodos a combinar pulsando el botón *Add* y lo configuramos con los parámetros apropiados pulsando el botón *Edit*).

- *metaClassifier*, que será el método empleado para aprender la predicción combinada (si queremos que la clasificación final se realice por votación mayoritaria basta elegir como metaclasificador ZeroR, que es la opción por defecto).

- *numfolds*, donde indicaremos el número de particiones empleadas para la validación cruzada (por defecto 10).



Ejercicios

1. Haciendo uso del *Explorer*, comparar los resultados obtenidos para el *dataset* "credit-g.arff" usando como algoritmos de clasificación: *J48*, *Bagging* sobre *J48* y *AdaboostM1* sobre *J48*. Probar con 10, 20 y 40 repeticiones tanto para *bagging* como para *boosting*. Usar como medidas para la comparativa el porcentaje de acierto, la media F y el Área Bajo la Curva ROC (AUC).

2. Realizar la comparativa anterior, pero en esta ocasión sobre el *dataset* "contact-lenses.arff".

3. Elaborar un pequeño informe con los resultados obtenidos en los puntos anteriores en forma tabular y las conclusiones a las que llega.

4. Usar el *Experimenter* para realizar una comparativa de los resultados obtenidos al clasificar los *datasets* "credit-g.arff", "weather.arff" e "iris.arff" con los siguientes métodos:
   - K-NN con valor de k=1.
   - K-NN con valor de k=3.
   - K-NN con valor de k=5.
   - OneR
   - SMO
   - Stacking, tomando como clasificadores base los tres K-NN anteriores y el OneR y usando J48 como algoritmo para hacer el meta-aprendizaje.

   Para que el experimento no ocupe excesivo tiempo limitar el número de repeticiones a 5.

5. Una vez creado el fichero con los datos del experimento, salva dicho fichero de configuración, así como los resultados.

6. Elaborar un pequeño informe con los resultados obtenidos y las conclusiones a las que llega.

¿Cómo entregar la práctica?

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

- Sube también el fichero de configuración y el fichero de resultados obtenidos con el *Experimenter.*