uc3m | Universidad Carlos III de Madrid

OPENCOURSEWARE

APRENDIZAJE AUTOMÁTICO PARA EL ANÁLISIS DE DATOS

GRADO EN ESTADÍSTICA Y EMPRESA

Ricardo Aler

PRÁCTICA 2: Clasificación con datos desbalanceados (puntuación máxima **3.0 puntos**).

1. Descripción del problema

Vamos a trabajar con un problema de clasificación con muestras desbalanceadas. El problema consiste en clasificar imágenes con y sin microcalcificaciones en mamografía [1]. Cada imagen viene representada por 6 atributos, además de la clase. Cada atributo representa alguna propiedad de la imagen (brillo, forma, tamaño, …).

2. (0.5 puntos) Preparación de los datos y exploración del problema

- Leed los datos *"mammography.arff"* con *read.arff* del paquete *foreign*.
- Usad *head* y *summary* para conocer los datos: de qué tipo son los atributos y si existen NA's.
- La columna que contiene la clase es la última. Determinad cuántos datos hay de cada clase, cuál es la minoritaria y si estamos en presencia de una muestra desbalanceada. Transformad la columna de la clase, que contiene los valores -1 y +1 (por tanto es numérica), en otra de tipo factor (*as.factor*), que contenga los valores NORMAL y MICRO (donde MICRO es la clase minoritaria).

Como sabemos, en problemas desbalanceados, puede ocurrir que el clasificador aprenda bien la clase mayoritaria a costa de ignorar la minoritaria. Comprobad si esto ocurre en este problema mediante máquinas de vectores de soporte con kernel radial e hiper-parámetros por omisión. Como método de evaluación, usad train/test (Holdout). Las medidas para evaluar van a ser la tasa de aciertos o accuracy (*acc*), el *balanced accuracy* (*bac*), el true positive rate (*tpr*), y el true negative rate (*tnr*). *bac* = ½*(*tpr*+*tnr*). Recordad que en problemas de muestras desbalanceadas, es importante que las particiones estén estratificadas. Comentad los resultados.

Repetid el punto anterior, pero construyendo directamente la matriz de confusión, como en el tutorial. ¿Se observan similares resultados?

3. (1.0 puntos) Curva ROC y cambio de *threshold* para mejorar BAC

Generad una curva ROC y justificad si se podría mejorar el TPR sin empeorar demasiado el TNR simplemente cambiando el *threshold*.

Si la respuesta a la cuestión anterior es positiva, vamos a ajustar automáticamente el threshold (ver en el apartado "Ajustando el threshold como cualquier otro hiper-parámetro" del tutorial). Vamos a ajustar **únicamente** el *threshold* (dejando fijos el resto de hiper-parámetros a sus valores por omisión). O sea, el resto de valores de hiper-parámetros deben quedar con sus valores por omisión. Usaremos grid-search, evaluación con 3 folds y optmizando *bac*. ¿Se consiguen mejorar TPR y BAC del apartado anterior, sin empeorar demasiado TNR, simplemente ajustando el threshold?

Si quisiéramos, también podríamos ajustar el resto de los hiper-parámetros de la SVM (*cost* y *gamma*), además del *threshold*, pero en este caso no vamos a hacerlo para ahorrarnos un poco de trabajo.

4. (1.5 puntos) Usando *SMOTE* para rebalancear la muestra

Una segunda manera de mejorar resultados con muestras desbalaceadas es el rebalanceo de la muestra. En concreto, *SMOTE* suele dar buenos resultados.

El hiper-parámetro más importante de *SMOTE* es el ratio que queremos usar para sobre-muestrear la clase minoritaria (*sw-rate*). Comenzaremos usando un valor razonable como rate = 1/(número datos MICRO/número datos NORMAL). Comparad los resultados con los mejores obtenidos hasta el momento.

En segundo lugar, usad algún método de ajuste automático de hiper-parámetros que permitan probar unos pocos valores de *sw-rate*, concretamente rate/8, rate/4, rate/2 y rate. Tienen que ser pocos porque SMOTE es algo lento. ¿Se mejoran los resultados? ¿Qué *sw-rate* es el mejor?

5. Referencias

[1] Woods, K., Doss, C., Bowyer, K., Solka, J., Priebe, C., & Kegelmeyer, P. (1993). Comparative Evaluation of Pattern Recognition Techniques for Detection of Microcalcifications in Mammography. International Journal of Pattern Recognition and Artificial Intelligence, 7(6), 1417–1436.