Remuestreo basado en *coverage*: construyendo árboles de decisión consolidados robustos

Igor Ibarguren, Jesús M. Pérez, Javier Muguerza, Olatz Arbelaitz e Ibai Gurrutxaga

Departamento de Arquitectura y Tecnología de Computadores, Universidad del País Vasco UPV/EHU, Manuel Lardizabal 1, 20018 Donostia, España
{igor.ibarguren,txus.perez,j.muguerza,olatz.arbelaitz,i.gurrutxaga}@ehu.eus,
sitio web: http://www.aldapa.eus/

Abstract. Este artículo es un resumen del trabajo publicado en la revista Knowledge-Based Systems [2] en el que se presenta una nueva estrategia de remuestreo ligado al desbalanceo presente en la muestra original y se aplica a la construcción de árboles de decisión consolidados.

Keywords: comprensibilidad, árboles de decisión consolidados, desbalanceo de clases, remuestreo

El desbalanceo de clases es un problema presente en los problemas de clasificación donde una o varias de las clases tienen una representación muy baja en comparación con el resto. Un ejemplo usado comúnmente es el diagnóstico de enfermedades poco frecuentes donde la mayoría de casos que se tiene corresponde a pacientes sanos. Es un problema que suscita gran interés en la comunidad. Una de las maneras de afrontar el desbalanceo de clases es el remuestreo de la muestra de entrenamiento. El algoritmo CTC (Consolidated Tree Construction) fue creado para solucionar un problema donde existía desbalanceo de clases y requiere múltiples muestras para crear un árbol consolidado. En el pasado, el algoritmo CTC se ha utilizado realizando un barrido de números de submuestras prefijados. Últimamente se ha trabajado con submuestras balanceadas. Incluso utilizando el mayor número de submuestras del barrido, para alguna clase de algunas bases de datos se obtenía una representación baja en las submuestras. Este trabajo presenta una manera de ajustar el número de submuestras utilizado en cada problema a la distribución de clases presente en la muestra. Se utilizan suficientes submuestras para asegurar que cada ejemplo de la muestra original tiene una probabilidad mínima de estar presente en al menos una de las submuestras. A esta probabilidad la llamamos cobertura (*coverage*). Problemas más desbalanceados requieren más muestras para obtener el mismo coverage.

La experimentación se ha realizado sobre 96 bases de datos repartidas en tres contextos de clasificación: un conjunto de 30 bases de datos estándares (la mayoría multi-clásicas), un conjunto de 33 bases de datos bi-clásicas desbalanceadas y el mismo conjunto de 33 bases de datos desbalanceadas preprocesadas con SMOTE hasta balancear las dos clases. La metodología utilizada es

un *5x5-fold cross validation* y se han utilizado tests estadísticos para comparar los resultados de los diferentes algoritmos. Las métricas elegidas para evaluar los clasificadores son las mismas que se utilizaron en [1]: kappa y la tasa de acierto (accuracy) para el conjunto estándar y la media geométrica para el conjunto de bases de datos desbalanceadas. En una primera fase de la experimentación, se realiza un análisis interno del algoritmo CTC, observando su comportamiento para un conjunto determinado de valores para el coverage. Se observa que la mayoría de las métricas utilizadas (kappa, tasa de aciertos, TNrate, MCC y F1-Score) aumentan a la vez que se incrementa el valor del coverage por lo que se elige un valor de coverage alto, el 99%, como representativo del algoritmo CTC. En una segunda fase, CTC se compara con los resultados publicados en [1][1] donde se comparaban 16 algoritmos genéticos y los 6 clásicos para inducción de reglas. Para cada uno de los 3 contextos, CTC se compara contra los 5 algoritmos genéticos elegidos en [1] como más competitivos para cada contexto y los 6 algoritmos clásicos. CTC se clasifica primero para kappa y cuarto para accuracy para las bases de datos estándares, primero para las bases de datos desbalanceadas, y tercero para las bases de datos desbalanceadas preprocesadas con SMOTE. Sin embargo, en una comparativa global, teniendo en cuenta las 96 bases de datos de los 3 contextos, CTC se clasifica primero con diferencias significativas contra la mayoría de algoritmos. Esto se debe a que la posición de la mayoría de los algoritmos fluctúa de manera considerable entre diferentes contextos, mientras que CTC se clasifica en las primeras posiciones, incluso primera, en las tres. Por lo tanto se observa una gran robustez en los árboles de decisión consolidados. Una comparativa posterior compara los resultados de CTC en las bases de datos desbalanceadas frente a un conjunto de mejores propuestas para combatir el desbalanceo de clases encontradas en varias publicaciones de la literatura. En este caso CTC no es capaz de superar a sus competidores aunque sí que mejora el resultado del algoritmo en el que se basa: C4.5.

Acknowledgement

Este trabajo fue financiado por el Gobierno Vasco (IT-395-10 y PRE-2013-1-887, BOPV/2013/128/3067) y por el Ministerio de Economía y Competitividad del Gobierno de España, cofinanciado por el FEDER (TIN2014-52665-C2-1-R).

Referencias

1. Fernández, A., García, S., Luengo, J., Bernadó-Mansilla, E., Herrera, F.: Genetics-based machine learning for rule induction: State of the art, taxonomy, and comparative study. Evolutionary Computation, IEEE Transactions on 14(6), 913–941 (2010)
2. Ibarguren, I., Pérez, J.M., Muguerza, J., Gurrutxaga, I., Arbelaitz, O.: Coverage-based resampling: Building robust consolidated decision trees. Knowledge-Based Systems 79(0), 51 – 67 (2015)

[1] http://sci2s.ugr.es/gbml