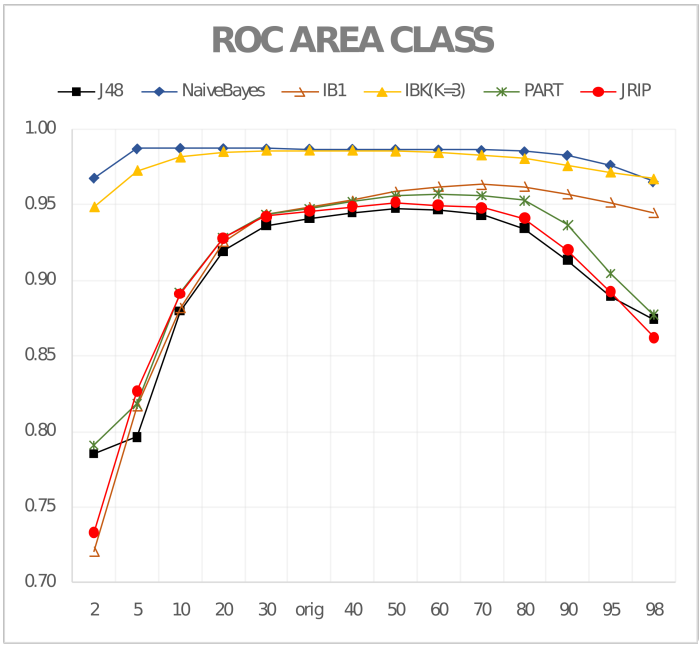
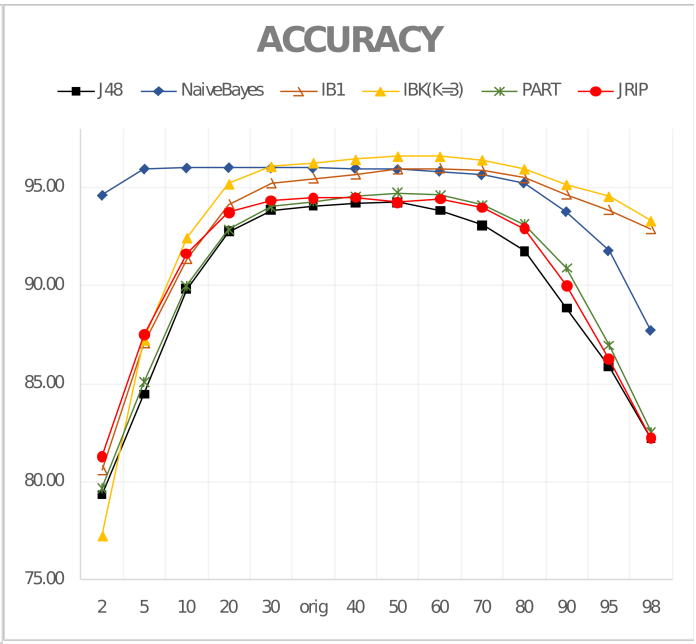
Mejora de la efectividad de la clasificación en la plataforma WEKA en base al uso de métodos de remuestreo sobre la distribución de clases óptima

Se trata de implementar un módulo en la plataforma WEKA que desarrolle la aproximación propuesta por el grupo ALDAPA (http://www.aldapa.eus/) en el artículo "The Quest for the Optimal Class Distribution: an Approach for Enhancing the Effectiveness of Learning via Resampling Methods for imbalanced data sets" [1] (http://link.springer.com/article/10.1007/s13748-012-0034-6).

Esta aproximación está compuesta por dos fases. En una primera fase se determina experimentalmente cuál es la distribución de clases (pseudo-) óptima asociada a un problema de clasificación concreto para un algoritmo de aprendizaje concreto, realizando para ello un barrido de distintos valores de la distribución de clases, de 2% a 98%, usando un método de remuestreo poco costoso computacionalmente como es el submuestreo aleatorio con muestras del tamaño de la clase minoritaria (basado en un trabajo de Weiss y Provost del 2003 [2]).

Esta fase ya está integrada en WEKA y permite obtener resultados como los que se muestran en las gráficas adjuntas, donde se puede observar la evolución de los resultados en base a los criterios de bondad AUC (izquierda) y tasa de acierto (derecha) para la base de datos *breast-w* del repositorio de la UCI [3] (distribución de clases original (orig): 34.48) y un conjunto de seis algoritmos diferentes. Como se puede observar, aunque hay ciertas similitudes en el comportamiento de los distintos algoritmos, la distribución óptima varía de un algoritmo a otro, también de un criterio a otro, y no tiene porqué ser del 50% (clases balanceadas), valor típicamente usado.





En una segunda fase, se usará la distribución de clases óptima obtenida en la fase anterior con un método de remuestreo concreto (de entre los propuestos en la bibliografía para afrontar problemas de *class imbalance*, como SMOTE [4], por ejemplo, ya implementado en WEKA) y se pivotará sobre este valor óptimo (como en el artículo, por ejemplo, con +10 y -10) para buscar la mejor distribución de clases para el método de remuestreo elegido y conseguir así mejorar los resultados para el problema de clasificación y el algoritmo de aprendizaje en consideración.

Además, para corroborar la eficacia de la aproximación propuesta, se realizará una experimentación con un conjunto de bases de datos biclásicas de problemas de clasificación del repositorio de la UCI [3] (http://archive.ics.uci.edu/ml) ya usados en el grupo ALDAPA en otros trabajos, sobre un

conjunto de diferentes algoritmos de clasificación y una serie de métodos de remuestreo de entre los propuestos para afrontar proplemas de *class imbalance*.

Bibliografía

[1] I. Albisua, O. Arbelaitz, I. Gurrutxaga, A. Lasarguren, J. Muguerza, J.M. Pérez. *"The quest for the optimal class distribution: an approach for enhancing the effectiveness of learning via resampling methods for imbalanced data sets"*. Progress in Artificial Intelligence, Vol. 2, Issue 1, 45-63, (2013).

[2] G.M. Weiss, F. Provost *"Learning when training data are costly: The effect of class distribution on tree induction"*. Journal of Artificial Intelligence Research, Vol. 19 , 315–354 (2003).

[3] M. Lichman. *"UCI Machine Learning Repository"* [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, (2013).

[4] N. Chawla, K. Bowyer, L. Hall, W. Kegelmeyer. *"SMOTE: synthetic minority over-sampling technique"*. Journal of Artificial Intelligence Research, Vol. 16 , 321–357 (2004).

Contacto

txus.perez@ehu.eus

http://www.sc.ehu.es/txus