Off-centered entropies to deal with class imbalance for decision trees

Philippe Lenca*
Institut TELECOM, TELECOM Bretagne, Lab-STICC, Brest, France
philippe.lenca@telecom-bretagne.eu

Information Systems & Data Analysis, Vannes, March 27-28 2008

In supervised learning, the data set is said imbalanced if the class prior probabilities are highly unequal [5]. In the case of two-class problems, the larger class is called the majority class and the smaller the minority class. Real-life two-class problems -especially bank, insurance and finance data- have often minority class prior under 0.10 (e.g. fraud detection, credit scoring, extreme events which are events that have a high impact and a low frequency, etc.). In such a case the performances of data mining algorithms are lowered, especially the error rate corresponding to the minority class, even though this class corresponds to positive cases and the cost of misclassifying the positive examples is higher than the cost of misclassifying the negative examples. This problem gave rise to many papers (e.g. [6, 3, 15]) and dealing with imbalanced and cost-sensitive data has been recognized as one of the 10 most challenging problems in data mining [18]. Solutions to this problems are proposed both at the data and algorithmic level (e.g. [4, 2, 17, 16]).

Our presentation will focus on decision trees that are one of the most used data mining models ([1, 13]). One of the advantage of decision trees is that they produce intelligible results. This point could become more and more important due to Basel II and Solvency II accords.

At the data level, the proposed solutions change the class distribution. They include different forms of re-sampling, such that over-sampling or under-sampling on a random or a directed way. At the algorithmic level, a first solution is to re-balance the error rate by weighting each type of error with the corresponding cost. In decision trees learning, other algorithmic solutions consist for example in adjusting the probabilistic estimates at the tree leaf or adjusting the decision thresholds, the use of a criterion of minimal cost, or pre-pruning strategies for the cost-sensitive decision tree algorithm to avoid overfitting. At both levels, some researchers studied three issues (quality of probabilistic estimates, pruning, and effect of preprocessing the imbalanced data set), usually considered separately, concerning C4.5 decision trees and imbalanced data sets.

Our presentation will focus on algorithmic solutions for decision trees. To deal with the class imbalance problem, two non-centered entropies (off-centered entropy [8, 7, 10, 9] and asymmetric entropy [12, **?**, 19, 14]) have been proposed. Non-centered entropies have the particularity of taking their maximum value for a distribution fixed by the user. This distribution can be the a priori distribution of the class variable modalities or a distribution taking into account the costs of misclassification. In this presentation we will present the concepts of the different entropies and compare their effectiveness on imbalanced data sets. This presentation will show the interest of off-centered entropies to deal with the problem of class imbalance [11]. We will also present in a more general way decision trees features, like pruning and decision rules, for imbalanced data sets.

*Joint work with Stéphane Lallich, Thanh-Nghi Do, Nguyen-Khang Pham and Benoît Vaillant.

References

[1] L. Breiman, J. H. Friedman, R. A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Wadsworth International, 1984.

[2] N. Chawla. C4.5 and imbalanced datasets: Investigating the effect of sampling method, probalistic estimate, and decision tree structure. In *ICML'Workshop on Learning from Imbalanced Data Sets*, 2003.

[3] N. Chawla, N. Japkowicz, and A. Kolcz, editors. *Special Issue on Class Imbalances*, volume 6 of *SIGKDD Explorations*, 2004.

[4] P. Domingos. Metacost: A general method for making classifiers cost sensitive. In *Int. Conf. on Knowledge Discovery and Data Mining*, pages 155–164, 1999.

[5] N. Japkowicz. The class imbalance problem: Significance and strategies. In *International Conference on Artificial Intelligence*, volume 1, pages 111–117, 2000.

[6] N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5):429–450, 2002.

[7] S. Lallich, P. Lenca, and B. Vaillant. Construction of an off-centered entropy for supervised learning. In C. H. Skiadas, editor, *The XIIth International Symposium on Applied Stochastic Models and Data Analysis*, page 8 p., Chania, Crete, Greece, 2007.

[8] S. Lallich, B. Vaillant, and P. Lenca. Parametrised measures for the evaluation of association rule interestingness. In J. Janssen and P. Lenca, editors, *The XIth International Symposium on Applied Stochastic Models and Data Analysis*, pages 220–229, Brest, France, 2005.

[9] S. Lallich, B. Vaillant, and P. Lenca. Construction d'une entropie décentrée pour l'apprentissage supervisé. In *Atelier Qualité des Données et des Connaissances (associé à Extraction et Gestion des Connaissances 2007)*, pages 45–54, Namur, Belgium, 2007.

[10] S. Lallich, B. Vaillant, and P. Lenca. A probabilistic framework towards the parameterization of association rule interestingness measures. *Methodology and Computing in Applied Probability*, 9:447–463, 2007.

[11] P. Lenca, S. Lallich, T.-N. Do, and N.-K Pham. A comparison of different off-centered entropies to deal with class imbalance for decision trees. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Osaka, Japan, 2008.

[12] S. Marcellin, D. A. Zighed, and G. Ritschard. An asymmetric entropy measure for decision trees. In *IPMU 2006*, pages 1292–1299, Paris, France, 2006.

[13] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993.

[14] G. Ritschard, D. A. Zighed, and S. Marcellin. Données déséquilibrées, entropie décentrée et indice d'implication. In *Rencontres Internationales Analyse Statistique Implicative*, pages 315–327, Castellón, Spain, 2007.

[15] S. Visa and A. Ralescu. Issues in mining imbalanced data sets - A review paper. In *Midwest Artificial Intelligence and Cognitive Science Conf.*, pages 67–73, Dayton, USA, 2005.

[16] G. M. Weiss, K. McCarthy, and B. Zabar. Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs? In R. Stahlbock, S. F. Crone, and S. Lessmann, editors, *International Conference on Data Mining*, pages 35–41, Las Vegas, Nevada, USA, 2007. CSREA Press.

[17] G. M. Weiss and F. Provost. Learning when training data are costly: The effect of class distribution on tree induction. *J. of Artificial Intelligence Research*, 19:315–354, 2003.

[18] Q. Yang and X. Wu. 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making*, 5(4):597–604, 2006.

[19] D. A. Zighed, S. Marcellin, and G. Ritschard. Mesure d'entropie asymétrique et consistante. In *Extraction et Gestion des Connaissances*, pages 81–86, Namur, Belgium, 2007.