Cost-sensitive Classifier Evaluation using Cost Curves

Robert C. Holte[1] and Chris Drummond[2]

[1] Computing Science Department, University of Alberta, Edmonton, Alberta, Canada, T6G 2E8
holte@cs.ualberta.ca
[2] Institute for Information Technology, National Research Council, Ontario, Canada, K1A 0R6
Chris.Drummond@nrc-cnrc.gc.ca

Abstract. The evaluation of classifier performance in a cost-sensitive setting is straightforward if the operating conditions (misclassification costs and class distributions) are fixed and known. When this is not the case, evaluation requires a method of visualizing classifier performance across the full range of possible operating conditions. This talk outlines the most important requirements for cost-sensitive classifier evaluation for machine learning and KDD researchers and practitioners, and introduces a recently developed technique for classifier performance visualization – the cost curve – that meets all these requirements.

Introduction

Methods for creating accurate classifiers from data are of central interest to the data mining community [2, 15, 16]. The focus of this talk is on binary classification, *i.e.* classification tasks in which there are only two possible classes, which we will call *positive* and *negative*. In binary classification, there are just two types of error a classifier can make: a *false positive* is a negative example that is incorrectly classified as positive, and a *false negative* is a positive example that is incorrectly classified as negative. In general, the cost of making one type of misclassification will be different—possibly very different—than the cost of making the other type.[3]

Methods for evaluating the performance of classifiers fall into two broad categories: numerical and graphical. Numerical evaluations produce a single number summarizing a classifier's performance, whereas graphical methods depict performance in a plot that typically has just two or three dimensions so that it can be easily inspected by humans. Examples of numerical performance measures are accuracy, expected cost, precision, recall, and area under a performance curve (AUC). Examples of graphical performance evaluations are ROC curves [18, 19], precision-recall curves [6], DET curves [17], regret graphs [13], loss difference plots [1], skill plots [4], prevalence-value-accuracy plots [21], and the method presented in this talk, cost curves [7, 11].

Graphical methods are especially useful when there is uncertainty about the misclassification costs or the class distribution that will occur when the classifier is deployed. In this setting, graphical measures can present a classifier's actual performance

[3] We assume the misclassification cost is the same for all instances of a given class; see [12] for a discussion of performance evaluation when the cost can be different for each instance.

for a wide variety of different operating points (combinations of costs and class distributions), whereas the best a numerical measure can do is to represent the average performance across a set of operating points.

Cost curves are perhaps the ideal graphical method in this setting because they directly show performance as a function of the misclassification costs and class distribution. In particular, the x-axis and y-axis of a cost curve plot are defined as follows.

The x-axis of a cost curve plot is defined by combining the two misclassification costs and the class distribution—represented by $p(+)$, the probability that a given instance is positive—into a single value, $PC(+)$, using the following formula:

$$PC(+) = \frac{p(+)\mathtt{C}(-|+)}{p(+)\mathtt{C}(-|+) + (1 - p(+))\mathtt{C}(+|-)} \tag{1}$$

where $\mathtt{C(-|+)}$ is the cost of a false negative and $\mathtt{C(+|-)}$ is the cost of a false positive. $PC(+)$ ranges from 0 to 1.

Classifier performance, the y-axis of a cost curve plot, is "normalized expected cost" (NEC), defined as follows:

$$\mathtt{NEC} = FN * PC(+) + FP * (1 - PC(+)) \tag{2}$$

where FN is a classifier's false negative rate, and FP is its false positive rate. NEC ranges between 0 and 1.

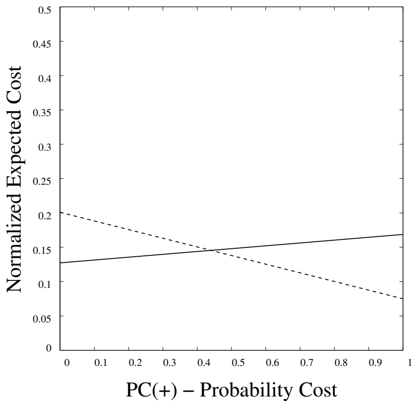

Fig. 1. Japanese credit - Cost curves for 1R (dashed line) and C4.5 (solid line)

To draw the cost curve for a classifier we draw two points, $y = FP$ at $x = 0$ and $y = FN$ at $x = 1$, and join them by a straight line. The cost curve represents the normalized expected cost of the classifier over the full range of possible class distributions and misclassification costs. For example, the dashed line in Figure 1 is the cost curve for the decision stump produced by 1R [14] for the Japanese credit dataset from the UCI repository and the solid line is the cost curve for the decision tree C4.5 [20] learns from the same training data. In this plot we can instantly see the relation between 1R and C4.5's performance across the full range of deployment situations. The vertical difference between the two lines is the difference between their normalized expected costs at a specific operating point. The intersection point of the two lines is the operating point where 1R's stump and C4.5's tree perform identically. This occurs at $PC(+) = 0.445$. For larger values of $PC(+)$ 1R's performance is better than C4.5's, for smaller values of $PC(+)$ the opposite is true.

Mathematically, cost curves are intimately related to ROC curves: they are "point-line duals" of one another. However, cost curves have the following advantages over ROC curves (see [11] for details):

- Cost curves directly show performance on their y-axis, whereas ROC curves do not explicitly depict performance. This means performance and performance differences can be easily seen in cost curves but not in ROC curves.
- When applied to a set of cost curves the natural way of averaging two-dimensional curves produces a cost curve that represents the average of the performances represented by the given curves. By contrast, there is no agreed upon way to average ROC curves, and none of the proposed averaging methods produces an ROC curve representing average performance.
- Cost curves allow confidence intervals to be estimated for a classifier's performance, and allow the statistical significance of performance differences to be assessed. The confidence interval and statistical significance testing methods for ROC curves do not relate directly to classifier performance.

For these reasons, we have gained insights into classifier performance using cost curves that would likely not have been possible using other methods [8–10] and other data mining researchers are using cost curves in their analyses [3, 5, 22, 23].

Acknowledgments

We thank the Natural Sciences and Engineering Research Council of Canada and the Alberta Ingenuity Centre for Machine Learning for their support of this research.

References

1. N. M. Adams and D. J. Hand. Comparing classifiers when misclassification costs are uncertain. *Pattern Recognition*, 32:1139–1147, 1999.
2. Maria-Luiza Antonie, Osmar R. Zaiane, and Robert C. Holte. Learning to use a learned model: A two-stage approach to classification. In *Proceedings of the 6th IEEE International Conference on Data Mining (ICDM'06)*, pages 33–42, 2006.
3. Andrea Bosin, Nicoletta Dessi, and Barbara Pes. Capturing heuristics and intelligent methods for improving micro-array data classification. In *Intelligent Data Engineering and Automated Learning - IDEAL 2007*, volume 4881 of *Lecture Notes in Computer Science*, pages 790–799. Springer, 2007.
4. William M. Briggs and Russell Zaretzki. The skill plot: a graphical technique for the evaluating the predictive usefulness of continuous diagnostic tests. *Biometrics*, OnlineEarly Articles, 2007.

5. Nitesh V. Chawla, Lawrence O. Hall, and Ajay Joshi. Wrapper-based computation and evaluation of sampling methods for imbalanced datasets. In *Workshop on Utility-Based Data Mining held in conjunction with the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 179–188, 2005.
6. Jesse Davis and Mark Goadrich. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning (ICML'06)*, pages 233–240, 2006.
7. Chris Drummond and Robert C. Holte. Explicitly representing expected cost: An alternative to ROC representation. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 198–207, 2000.
8. Chris Drummond and Robert C. Holte. C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling. In *Workshop on Learning from Imbalanced Datasets II, held in conjunction with ICML'03*, 2003.
9. Chris Drummond and Robert C. Holte. Learning to live with false alarms. In *Workshop on Data Mining Methods for Anomaly Detection held in conjunction with the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 21–24, 2005.
10. Chris Drummond and Robert C. Holte. Severe class imbalance: Why better algorithms aren't the answer. In *Proceedings of the 16th European Conference on Machine Learning (LNAI 3720)*, pages 539–546. Springer, 2005.
11. Chris Drummond and Robert C. Holte. Cost curves: An improved method for visualizing classifier performance. *Machine Learning*, 65(1):95–130, 2006.
12. Tom Fawcett. ROC graphs with instance-varying costs. *Pattern Recognition Letters*, 27(8):882–891, 2006.
13. Jorgen Hilden and Paul Glasziou. Regret graphs, diagnostic uncertainty, and Youden's index. *Statistics in Medicine*, 15:969–986, 1996.
14. Robert C. Holte. Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 11(1):63–91, 1993.
15. Masatoshi Jumi, Einoshin Suzuki, Muneaki Ohshima, Ning Zhong, Hideto Yokoi, and Katsuhiko Takabayashi. Spiral discovery of a separate prediction model from chronic hepatitis data. In *New Frontiers in Artificial Intelligence (LNAI 3609)*, pages 464–473. Springer, 2007.
16. T. Liu and Kai Ming Ting. Variable randomness in decision tree ensembles. In *Advances in Knowledge Discovery and Data Mining (LNAI 3918)*, pages 81–90. Springer, 2006.
17. Yang Liu and Elizabeth Shriberg. Comparing evaluation metrics for sentence boundary detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'07)*, volume 4, pages IV–185—IV–188, 2007.
18. F. Provost and T. Fawcett. Robust classification for imprecise environments. *Machine Learning*, 42:203–231, 2001.
19. Foster Provost and Tom Fawcett. Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, pages 43–48, Menlo Park, CA, 1997.
20. J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.
21. Alan T. Remaleya, Maureen L. Sampson, James M. DeLeo, Nancy A. Remaley, Beriuse D. Farsi, and Mark H. Zweig. Prevalence-value-accuracy plots: A new method for comparing diagnostic tests based on misclassification costs. *Clinical Chemistry*, 45:934–941, 1999.
22. Kai Ming Ting. Issues in classifier evaluation using optimal cost curves. In *Proceedings of The Nineteenth International Conference on Machine Learning*, pages 642–649, 2002.
23. Zhi-Hua Zhou and Xu-Ling Liu. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering*, 18(1):63–77, 2006.