Cost-sensitive Classifier Evaluation using Cost Curves

Robert C. Holte
Computing Science Department
University of Alberta
Edmonton, Alberta, Canada T6G 2E8
(holte@cs.ualberta.ca)

Chris Drummond
Institute for Information Technology
National Research Council
Ottawa, 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 and introduces a technique for classifier performance visualization – the cost curve – that meets all these requirements. We also briefly describe some application areas in which the usefulness of cost curves for classifier evaluation has been demonstrated.

Introduction

Methods for creating accurate classifiers from data are of central interest to the Artificial Intelligence community. 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.[1]

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 (Provost and Fawcett 2001; 1997), precision-recall curves (Davis and Goadrich 2006), DET curves (Liu and Shriberg 2007), regret graphs (Hilden and Glasziou 1996), loss difference plots (Adams and Hand 1999), skill plots (Briggs and Zaretzki 2007), prevalence-value-accuracy plots (Remaleya et al. 1999), and the method presented in this talk, cost curves (Drummond and Holte 2000; 2006).

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 for a wide variety of different operating conditions (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 conditions.

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:

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

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:

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

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

To draw the cost curve for a classifier we plot 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 (Holte 1993) for the Japanese credit dataset from the UCI repository and the solid line is the cost

[1] We assume the misclassification cost is the same for all instances of a given class; see (Fawcett 2006) for a discussion of performance evaluation when the cost can be different for each instance.

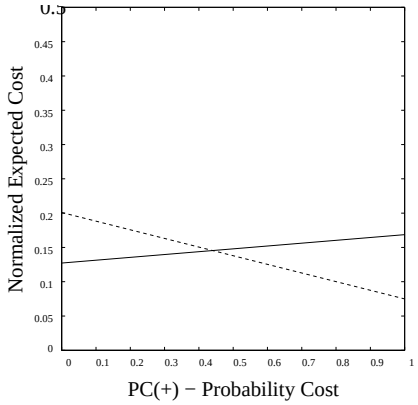

Figure 1: Japanese credit - Cost curves for 1R (dashed line) and C4.5 (solid line)

curve for the decision tree C4.5 (Quinlan 1986) 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 operating conditions. The vertical difference between the two lines is the difference between their normalized expected costs for a specific operating condition. The intersection point of the two lines is the operating condition for which 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 (Drummond and Holte 2006) 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 (Drummond and Holte 2003; 2005a; 2005b) and, as the following examples illustrate, researchers in a wide range of application areas are finding cost curves their analysis method of choice.

Illustrative Applications of Cost Curves

Frattini *et al.* (2010) compare numerous approaches, including cost curves, for evaluating landslide susceptibility models, which are classifiers predicting whether a given locality is, or is not, susceptible to landslides. The costs associated with misclassification in this application are as follows. A false negative is land that is incorrectly classified as being susceptible to landslides. This leads to economic loss, since land that could be developed will not be. A false positive is a much more expensive type of error because land will be developed—buildings built and occupied by people—in an area susceptible to landslides that could cause destruction and death. In this application it is important to compare classifiers' performances across a wide range of operating conditions because it is difficult to determine the class probabilities and misclassification costs for a given locality, and they vary significantly from one locality to another. The difficulty in obtaining these values is not unique to this application and is, in our opinion, a key reason to use cost curves because they allow one to easily see the operating regions for which each classifier is the best choice. It is therefore only necessary to determine which operating region a given locality falls into. Since the operating regions are often broad, selecting the best classifier for a locality does not usually require highly accurate estimates of the misclassification costs or class probabilities. For example, in Frattini *et al.*'s Figure 8, which shows the cost curves for the set of models they are comparing, one model dominates all others when $PC(+) < 0.8$ and a different model dominates when $PC(+) > 0.8$. Clearly, all that needs to be determined to choose the best model is whether $PC(+)$ is less than 0.8. Frattini *et al.* conclude that the use of cost curves, as opposed to ROC curves and the other evaluation methods they considered, "is advisable for evaluation and comparison of susceptibility models when a practical application of the model in land management is expected" (p. 71).

Garcia-Jimenez *et al.* (2010) explore the use of machine learning methods to create a binary classifier to predict if there exists a functional relationship between a given pair of proteins. There is extreme class imbalance in this application domain, with over 96% of the protein pairs being in the negative class (no functional relationship).[2] The performances of eight different machine learning methods were compared. Cost curves were used for this comparison, in preference to ROC curves, because "they are easier to interpret in meaningful units and they facilitate the selection of the best classifier by simple visualization" (p. 8).

Juntu *et al.* (2010) use ROC curves and cost curves to evaluate three machine learning methods to distinguish between benign and malignant soft-tissue tumours. They remark that cost curves are "much better for comparison between classifiers, especially when the ROC curves cross" (p. 685). In analyzing the cost curves of the three methods they observe that one of them is flat across a wide range of operating conditions ($0.2 \leq PC(+) \leq 0.75$). They correctly interpret this as indicating that the performance of that

[2]However, Garcia-Jimenez *et al.* artificially modified their training and the test sets to reduce the imbalance to only 4:1.

classifier is "insensitive to the distribution of the benign and malignant tumors ... or to change of the cost of misclassification" (p. 687) and note that the other two classifiers do not exhibit the same degree of insensitivity to operating conditions. This is a conclusion that is not at all apparent in the ROC curves for the classifiers.

Chen *et al.* (2009) use both ROC curves and cost curves to compare the performances of several machine learning methods for an application called hidden signal detection ('to detect an embedded signal in a data sample"). In this application it is important to have an extremely low false positive rate while still achieving a true positive rate of at least 50%. ROC curves are the appropriate visualization method to assess this criterion since they directly plot these two rates. However, misclassification costs are also important; the cost of a false positive is said to be at least 100 times larger than the cost of a false negative in this application. Cost curves are the appropriate visualization method to assess this criterion. This paper is therefore an example of a situation in which the complementary strengths of the two visualization techniques are both needed.

Hoshino *et al.* (2009) explore the use of a simple classifier to determine whether it is necessary to test a container for fumigants before a customs officer inspects it. The existing policy of the Canadian Border Security Agency is to apply an expensive and time-consuming chemical test to all containers selected for inspection (the trivial "always negative" classifier). For the purpose of visualizing the cost reduction that would be realized by using a specific alternative policy defined by a classifier, Hoshino *et al.* introduce a variant of cost curves called "improvement curves". Improvement curves have the same x-axis as cost curves but have a y-axis defined by $y = 1 - \frac{\texttt{NEC}(x)}{x}$. The authors conclude that this variation on cost curves provides "any easy way for decision-makers to estimate the potential improvement of introducing a predictive model to the overall risk assessment process, without having to determine the exact number of positive and negative containers, or knowing the exact costs of a ventilation or chemical test" (p. 24).

Sacanamboy and Cukic (2009) investigate the use of cost curves to evaluate and compare biometric systems, motivated by the fact that it is essential, in high security applications, to fully take into account the costs of misclassification. Their study looks at both face recognition and fingerprint recognition. They observe that "cost curves have the ability to reveal the differences between the biometric algorithms that are not obvious in the corresponding ROC or DET curves" (p. 6) and suggest that cost curves be included "in the evaluation of recognition performance of biometric algorithms in order to guide the selection of the most adequate algorithm based on the system requirements" (p. 6).

Jiang, Cukic, and Menzies (2008) use cost curves to evaluate methods for predicting whether a software module contains a fault. This seems to be the first time the costs of misclassification have been taken into account in this application area. Software quality professionals did not find it difficult to give approximate values for the misclassification costs. As Jiang *et al.*'s investigation with cost curves unfolded, it became clear to them that "in many cases these [fault prediction] algorithms barely outperform trivial classifiers, especially in extreme high and low risk situations" (p. 202). They proceeded to use the statistical significance test on cost curves to determine operating conditions under which the best fault prediction models performed better than the trivial classifiers. The paper concludes: "we strongly recommend [cost curves'] inclusion in the evaluation of software quality models" (p. 205).

Acknowledgements

This research was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Alberta Ingenuity Centre for Machine Learning (AICML).

References

Adams, N. M., and Hand, D. J. 1999. Comparing classifiers when misclassification costs are uncertain. *Pattern Recognition* 32:1139–1147.

Briggs, W. M., and Zaretzki, R. 2007. The skill plot: a graphical technique for the evaluating the predictive usefulness of continuous diagnostic tests. *Biometrics* OnlineEarly Articles.

Chen, B. Y.; Lemmond, T. D.; and Hanley, W. G. 2009. Building ultra-low false alarm rate support vector classifier ensembles using random subspaces. In *Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, 1–8.

Davis, J., and Goadrich, M. 2006. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning (ICML'06)*, 233–240.

Drummond, C., and Holte, R. C. 2000. Explicitly representing expected cost: An alternative to ROC representation. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 198–207.

Drummond, C., and Holte, R. C. 2003. 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*.

Drummond, C., and Holte, R. C. 2005a. 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*, 21–24.

Drummond, C., and Holte, R. C. 2005b. Severe class imbalance: Why better algorithms aren't the answer. In *Proceedings of the 16th European Conference on Machine Learning (LNAI 3720)*, 539–546. Springer.

Drummond, C., and Holte, R. C. 2006. Cost curves: An improved method for visualizing classifier performance. *Machine Learning* 65(1):95–130.

Fawcett, T. 2006. ROC graphs with instance-varying costs. *Pattern Recognition Letters* 27(8):882–891.

Frattini, P.; Crosta, G.; and Carrara, A. 2010. Techniques for evaluating the performance of landslide susceptibility models. *Engineering Geology* 111(1-4):62 – 72.

García-Jiménez, B.; Juan, D.; Ezkurdia, I.; Andrés-León, E.; and Valencia, A. 2010. Inference of Functional Relations in Predicted Protein Networks with a Machine Learning Approach. *PLoS ONE* 5(4):e9969+.

Hilden, J., and Glasziou, P. 1996. Regret graphs, diagnostic uncertainty, and Youden's index. *Statistics in Medicine* 15:969–986.

Holte, R. C. 1993. Very simple classification rules perform well on most commonly used datasets. *Machine Learning* 11(1):63–91.

Hoshino, R.; Coughtrey, D.; Sivaraja, S.; Volnyansky, I.; Auer, S.; and Trichtchenko, A. 2009. Applications and extensions of cost curves to marine container inspection. *Annals of Operations Research* 1–25.

Jiang, Y.; Cukic, B.; and Menzies, T. 2008. Cost curve evaluation of fault prediction models. In *Proceedings of the 2008 19th International Symposium on Software Reliability Engineering*, 197–206. Washington, DC, USA: IEEE Computer Society.

Juntu, J.; Sijbers, J.; De Backer, S.; Rajan, J.; and Van Dyck, D. 2010. A machine learning study of several classifiers trained with texture analysis features to differentiate benign from malignant soft tissue tumors in T1-MRI images. *Journal of Magnetic Resonance Imaging* 31:680689.

Liu, Y., and Shriberg, E. 2007. Comparing evaluation metrics for sentence boundary detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'07)*, volume 4, IV–185—IV–188.

Provost, F., and Fawcett, T. 1997. 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*, 43–48.

Provost, F., and Fawcett, T. 2001. Robust classification for imprecise environments. *Machine Learning* 42:203–231.

Quinlan, J. R. 1986. Induction of decision trees. *Machine Learning* 1:81–106.

Remaleya, A. T.; Sampson, M. L.; DeLeo, J. M.; Remaley, N. A.; Farsi, B. D.; and Zweig, M. H. 1999. Prevalence-value-accuracy plots: A new method for comparing diagnostic tests based on misclassification costs. *Clinical Chemistry* 45:934–941.

Sacanamboy, M., and Cukic, B. 2009. Cost curve analysis of biometric system performance. In *Proceedings of Biometrics: Theory, Applications, and Systems (BTAS09)*, 1–6.