Learning When Data Sets are Imbalanced and When Costs are Unequal and Unknown

Marcus A. Maloof MALOOF@CS.GEORGETOWN.EDU
Department of Computer Science, Georgetown University, Washington, DC 20057-1232, USA

Abstract

The problem of learning from imbalanced data sets, while not the same problem as learning when misclassification costs are unequal and unknown, can be handled in a similar manner. That is, in both contexts, we can use techniques from ROC analysis to help with classifier design. We present results from two studies in which we dealt with skewed data sets and unequal, but unknown costs of error. We also compare for one domain these results to those obtained by over-sampling and under-sampling the data set. The operations of sampling, moving the decision threshold, and adjusting the cost matrix produced sets of classifiers that fell on the same ROC curve.

1. Introduction

We are interested in the connection between learning from imbalanced or skewed data sets and learning when error costs are unequal, but unknown. In this paper, we argue that while these problem are not exactly the same, they can be handled in the same manner. To illustrate, we present results from two previous studies in which we used techniques from ROC analysis to cope with data sets with different amounts of skew. We also over-sampled (or up-sampled) and under-sampled (or down-sampled) one of the data sets, showing that the ROC curves produced by this procedure are similar to those produced by varying the decision threshold or the cost matrix. ROC analysis is most often associated with cost-sensitive learning, but it is equally applicable to the problem of learning from imbalanced data sets, which we discuss further in the next section.

2. The Problem of Imbalanced Data Sets

The problem of learning from imbalanced or skewed data sets occurs when the number of examples in one class is significantly greater than that of the other.[1] Breiman et al. (1984) discussed the connection between the prior probability of a class and its error cost. Classes with fewer examples in the training set have a lower prior probability and a lower error cost. This is problematic when true error cost of the minority class is higher than is implied by the distribution of examples in the training set.

When applying learning methods to skewed data sets, some algorithms will find an acceptable trade-off between the true-positive and false-positive rates. However, others learn simply to predict the majority class. Indeed, classifiers that always predict the majority class can obtain higher predictive accuracies than those that predict both classes equally well. Skewed data sets arise frequently in many real-world applications, such as fraud detection (Fawcett & Provost, 1997), vision (Maloof et al., to appear), medicine (Mac Namee et al., 2002), and language (Cardie & Howe, 1997).

There have been several proposals for coping with skewed data sets (Japkowicz, 2000). For instance, there are sampling approaches in which we over-sample (i.e., duplicate) examples of the minority class (Ling & Li, 1998), under-sample (i.e., remove) examples of the majority class (Kubat & Matwin, 1997), or both (Chawla et al., 2002). We can also learn to predict the minority class with the majority class as the default prediction (Kubat et al., 1998). Schemes also exist to weight examples in an effort to bias the performance element toward the minority class (Cardie & Howe, 1997) and to weight the rules themselves (Grzymala-Busse et al., 2000). There have also been proposals to boost the examples of the minority class (Joshi et al., 2001).

The analysis of Breiman et al. (1984) establishes the connection among the distribution of examples in the training set, the prior probability of each class, the

[1] For simplicity, we will restrict discussion to the two-class case.

costs of mistakes on each class, and the placement of the decision threshold. Varying one of these elements is equivalent to varying any other. For example, learning from a set of under-sampled data sets is equivalent to evaluating a classifier at different decision thresholds. However, the precise relationship among these things is complex and task- and method-specific. In the next section we discuss some basic concepts of ROC analysis, which is useful for analyzing performance when varying the decision threshold, the cost of misclassification, or the distribution of training examples.

3. Basic Concepts of ROC Analysis

Receiver Operating Characteristic (ROC) analysis (Swets & Pickett, 1982) has its origin in signal detection theory, but most of the current work occurs in the medical decision making community. Researchers in the machine learning community have just recently become interested in ROC analysis as a method for evaluating classifiers. Indeed, it is a method of analysis unconfounded by inductive bias, by unknown, but unequal error costs, and, as we describe in this paper, by the class distribution of examples (Metz, 1978; Provost et al., 1998; Maloof et al., to appear).

Parametric ROC analysis is based on a *binormal assumption*, meaning that the actually positive cases are normally distributed and the actually negative cases are normally distributed (Metz, 1978). Naturally, it is the overlap between these two distributions that results in the Bayes error rate (Duda et al., 2000). Once we have characterized in some way the training examples drawn from these two distributions, then we are free to set a decision threshold most anywhere. It is typically best to select the decision threshold that minimizes the Bayes error rate. Alternatively, if error costs are unequal and known, then we can adjust the decision threshold to minimize the overall cost of errors.

As stated previously, there is a strong connection between the prior probability of a class and its error cost. If the class distribution of examples is consistent with the cost of errors, then building a classifier consistent with those costs should pose little problem. However, when data sets are skewed in a manner that runs counter to the true cost of errors, then even if we know the cost of errors, it may be difficult to build a classifier that is consistent with those costs. To make matters worse, we often have only anecdotal evidence about the relationship between the class distribution and the cost of errors. For instance, on a rooftop detection task, which we discuss further in Section 5,

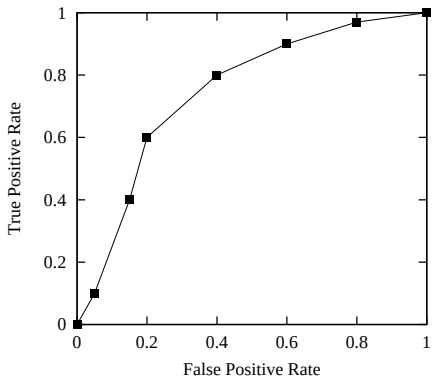

Figure 1. A hypothetical Receiver Operating Characteristic (ROC) curve.

we had a highly skewed data set (i.e., 781 rooftops versus 17,048 non-rooftops), we knew mistakes on the rooftop class were much more expensive than those on the other class, but we had no way of conducting a cost analysis (Maloof et al., 1997).

In these situations, one way to proceed is to move the decision threshold for a given classifier from the point at which mistakes on the positive class are maximally expensive to the point at which mistakes on the negative class are maximally expensive. Doing so will produce a set of true-positive and false-positive rates. Graphing these points yields an ROC curve, similar to the one pictured in Figure 1. There are also parametric methods for fitting to these points the curve of maximum likelihood (Dorfman & Alf, 1969; Metz et al., 1998).

It is often convenient to characterize ROC curves using a single measure. Many exist (Swets & Pickett, 1982), but ubiquitous is the area under the ROC curve. We can use the trapezoid rule to approximate the area, and it is also a simple matter to compute the area under the curve of maximum likelihood (Bamber, 1975; Thompson & Zucchini, 1986).

It is also possible to produce ROC curves from *case ratings*, whereby we modify the performance element to produce a rating for each test case. For example, we modified naive Bayes to output the posterior probability of the negative class for each test example (Maloof et al., 2002). Given m ratings of negative cases, $\mathbf{r}^-$, and n ratings of positive cases, $\mathbf{r}^+$,

$$\hat{A} = \frac{1}{m\,n} \sum_{i=1}^{m} \sum_{j=1}^{n} I(r_i^-, r_j^+),$$

where

$$I(r^-, r^+) = \begin{cases} 1 & \text{if } r^- > r^+; \\ \frac{1}{2} & \text{if } r^- = r^+; \\ 0 & \text{if } r^- < r^+. \end{cases}$$

Table 1. Performance on a recidivism prediction task.

Classification Method	TP Rate	TN Rate
Proportional Hazards[a]	0.72	0.53
Nearest Neighbor	0.45	0.70
C5.0	0.36	0.83
Naive Bayes	0.36	0.85

[a]As reported by Schmidt & Witte, 1988.

This is the Mann-Whitney two-sample statistic, and researchers have shown it to be equivalent to computing the area under the ROC curve using the trapezoid rule (DeLong et al., 1988). We can map the sorted case ratings into, say, 10–12 bins (Wagner et al., 2001) and use the number of true-positive and true-negative cases to determine points on an ROC curve (Metz et al., 1998).

Area under the curve is most appropriate when each curve dominates another. However, researchers have proposed analyses for when curves cross (Provost & Fawcett, 2001). There are also analyses for when only a portion of the ROC curve is of interest (McClish, 1989; Woods et al., 1997) and when analyzing more than two decisions (Swets & Pickett, 1982; Mossman, 1999; Hand & Till, 2001). Cost curves are equivalent to ROC curves, but plot expected cost explicitly, which can make for easier comparisons (Drummond & Holte, 2000).

To conduct a statistical analysis of ROC curves and their areas, one can use traditional tests, such as the t-test or analysis of variance (ANOVA) (Bradley, 1997; Maloof et al., to appear), but these procedures do not take into account the case-sample variance (Metz, 1989). Indeed, since ANOVA does not take into account all sources of variance, it may have higher Type I error than will such tests designed expressly for ROC curves (Maloof, 2002; Metz & Kronman, 1980). LABMRMC takes into account case-sample variance using the jackknife method, and then uses ANOVA to determine if treatment means are equal (Dorfman et al., 1992). Naturally, ANOVA carries with it an assumption of normality and is robust when this assumption is violated, but researchers have recently proposed nonparametric methods of analysis (Beiden et al., 2000).

4. Recidivism Prediction

To illustrate the value of ROC analysis for learning from imbalanced data sets, we first present results on a recidivism prediction task (Maloof, 1999). We must predict if an individual will re-commit a crime after release from prison based on characteristics such as age, type of crime, history of alcohol and drug abuse, and similar indicators. The distribution of training examples was 27.5% recidivist (i.e., positive) and 72.5% non-recidivist (i.e., negative), which is not skewed as severely as other reported data sets (e.g., Cardie & Howe, 1997; Kubat et al., 1998; Maloof et al., to appear). Schmidt and Witte (1988) give further details about this problem, including their results using a proportional hazards model.[2]

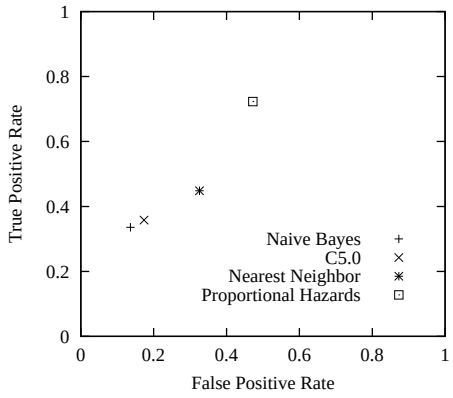

Figure 2. Results plotted in an ROC graph for recidivism prediction.

Using the hold-out method, the same experimental design used by Schmidt and Witte (1988), we ran naive Bayes (e.g., Langley et al., 1992), nearest neighbor (e.g., Aha et al., 1991), and C5.0, the commercial successor of C4.5 (Quinlan, 1993), which produced the results appearing in Table 1. We also plotted these results in an ROC graph, and these appear in Figure 2.

As one can see, the proportional hazards model performed better than did the other learners, mostly in terms of the true-positive rate. We have only an informal notion of error costs for this problem: mistakes on the positive class are more expensive than those on the negative class. When taking this into account, we would prefer methods achieving higher true-positive rates to those achieving higher true-negative rates. Consequently, the results for the learning methods are actually worse than they appear.

The problem with this analysis is that we have not accounted for differences in inductive bias, in error cost, and in how each method copes with the imbalance of the data set. Indeed, each method will yield a classifier subject to these factors, but there is no guarantee

[2]This is a nonparametric technique that predicts the time until recidivism using an individual's characteristics. We estimate the model by maximizing a *partial likelihood function* that indicates the probability of failure (i.e., recidivism) of individuals as a function of time.

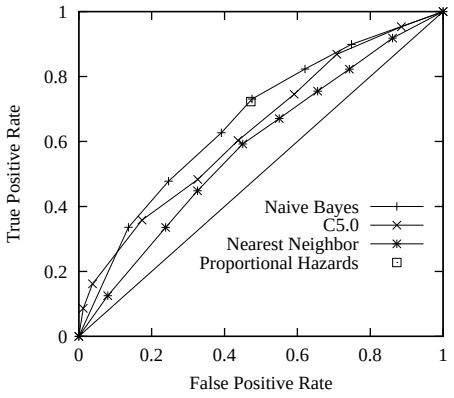

Figure 3. ROC curves for recidivism prediction.

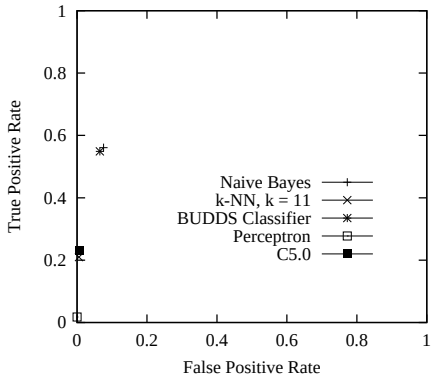

Figure 4. Results for the rooftop detection task plotted in an ROC graph. (Maloof et al., to appear). © 2003 Kluwer Academic Publishers.

that the method will have found the right trade-off among them. However, we can account for all three by using cost-sensitive learning algorithms, evaluating classifiers at different decision thresholds or with different cost matrices, and plotting performance as ROC curves. Therefore, we repeated the previous experiment, but varied the cost matrix of C5.0 and the decision thresholds of naive Bayes and nearest neighbor. The results for this experiment appear in Figure 3. We also plotted the original point for proportional hazard model for the sake of comparison.

By finding the appropriate decision threshold, we were able to compensate for the imbalance in the training set and produce a naive Bayesian classifier with performance equal to that of the proportional hazards model. We did not have a cost-sensitive version of proportional hazards, but we anticipate that its ROC curve would be similar to that of naive Bayes. We also concluded the naive Bayes performed better on this task than did nearest neighbor, since the former's ROC curve covers a larger area. For this experiment, naive Bayes produced an ROC curve with an area of 0.667, while C5.0 produced one of area 0.635 and nearest neighbor produced one of area 0.584. Note that the diagonal line in Figure 3 represents discrimination at the chance level; thus, none of the methods performed much better than this.

5. Rooftop Detection

We applied a similar methodology to the problem of learning to detect rooftops in overhead imagery, a problem with a more severely imbalanced data set (Maloof et al., 1998; Maloof et al., to appear). We used the Building Detection and Description System (Lin & Nevatia, 1998), or BUDDS, to extract candidate rooftops (i.e., parallelograms) from six large-area images. Such processing resulted in 17,829 such candidates, which an expert labeled as 781 positive examples and 17,048 negative examples of the concept "rooftop." Nine continuous attributes characterized each example, taking into account the strength of edges and corners, the degree to which opposing sides are parallel, and other similar evidence.

Using a variety of learning methods, we conducted a traditional evaluation using ten iterations of the 60/40% hold-out method. For the sake of comparison, we also included the heuristic present in the BUDDS system, which we call the BUDDS classifier. It is a linear classifier with handcrafted weights. Table 2 shows results from the evaluation, and as before, we plotted the true-positive and false-positive rates in an ROC graph, which appear in Figure 4.

C5.0 achieved the highest overall accuracy, but naive Bayes was best at detecting rooftops. Unfortunately, naive Bayes performed only slightly better than the BUDDS classifier, upon which we were trying to improve. The perceptron algorithm performed well overall, but by learning to always predict the negative (i.e., majority) class.

We repeated this experiment using cost-sensitive learning algorithms and plotted the results as ROC curves, which appear in Figure 5. The areas under these curves and their 95% confidence intervals appear in Table 3. As with the previous domain, cost-sensitive learning algorithms and ROC analysis not only let us cope with a skewed data set, but also let us better visualize the performance of the learning methods.

Note that since we evaluated each method at the same decision thresholds, we produced an average ROC curve by *pooling* (Swets & Pickett, 1982) the ROC curves from the ten runs; that is, we averaged the true-positive and

Table 2. Results for rooftop detection task. Measures are accuracy, true-positive (TP) rate, false-positive (FP) rate with 95% confidence intervals. Italics type shows the best measure in each column. (Maloof et al., to appear). © 2003 Kluwer Academic Publishers.

Method	Accuracy	TP Rate	FP Rate
C5.0	*0.963±0.003*	0.23±0.022	0.0034±0.0011
k-NN ($k = 17$)	0.961±0.001	0.19±0.015	0.0037±0.0003
k-NN ($k = 11$)	0.960±0.001	0.21±0.017	0.0056±0.0006
k-NN ($k = 5$)	0.957±0.001	0.23±0.010	0.0097±0.0009
Perceptron	0.957±0.001	0.02±0.011	*0.0001±0.0001*
BUDDS classifier	0.917±0.001	0.54±0.018	0.0657±0.0008
Naive Bayes	0.908±0.003	*0.56±0.008*	0.0761±0.0036

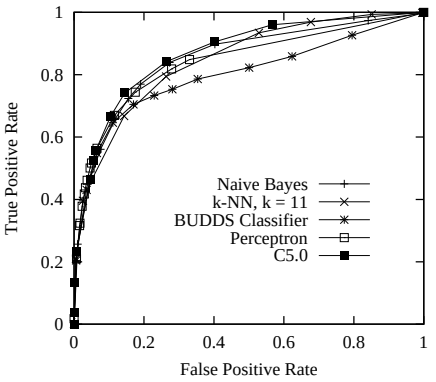

Figure 5. ROC curves for the rooftop detection task (Maloof et al., to appear). © 2003 Kluwer Academic Publishers.

Table 3. Areas under the ROC curves and 95% confidence intervals for the rooftop detection task (Maloof et al., to appear). © 2003 Kluwer Academic Publishers.

Classifier	Area under ROC Curve
C5.0	0.867±0.006
Naive Bayes	0.854±0.009
Perceptron	0.853±0.010
k-NN ($k = 11$)	0.847±0.006
BUDDS classifier	0.802±0.014

false-positive rates over the ten runs and then plotted the ROC curve (cf. Provost et al. 1998). In other work, to produce an average ROC curve, we fit the curve of maximum likelihood to case ratings under a binormal assumption, averaged the ROC-curve parameters a and b (or a and Δm), and produced an ROC curve using these averaged parameters.

6. Discussion

If we again look at the points in the ROC graphs in Figures 2 and 4, we see that each method performed quite differently when presented with the same skewed data set, a phenomenon due to each method's inductive bias. However, simply because a given classifier produced a point in a better part of the ROC graph, this did not mean that the classifier's ROC curve would dominate all other curves. This was true for naive Bayes on the recidivism prediction task and true for C5.0 on the rooftop detection task. Therefore, by varying the decision threshold or the cost matrix, we can compensate for skewed data sets. In the next section, we examine the connection between these operations and sampling.

6.1. Why Sample?

In previous sections, we examined the use of cost-sensitive learning algorithms and ROC analysis to cope with imbalanced data sets. As we have mentioned, several researchers have investigated sampling approaches for coping with skewed data sets. We anticipate that sampling will produce the same effect as moving the decision threshold or adjusting the cost matrix. To investigate this notion, we used C5.0 and naive Bayes on the rooftops data to conduct an experiment in which we under-sampled the negative, majority class and then over-sampled the positive, minority class.

To execute this experiment, we randomly divided the rooftops data into training (60%) and testing (40%) portions. For the under-sampling condition, we created ten training sets using all of the positive examples and decreasing amounts of negative examples. We then built classifiers and evaluated them on the examples in the test set. We repeated this procedure ten times and plotted the average true-positive and false-positives rates for these runs as an ROC curve.

For the over-sampling condition, we proceeded similarly, but created training sets using increasing amounts of positive examples. Specifically, we created a total of ten training sets, and for each, we duplicated the positive examples by ten fold. That is, for the first run, we included ten copies of the positive data, for the

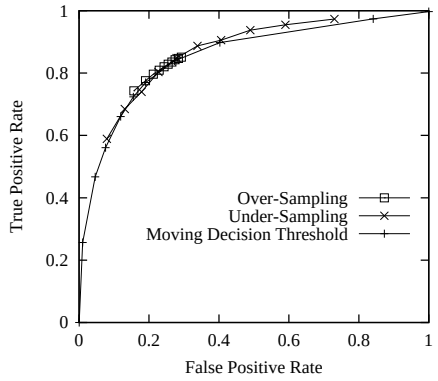

Figure 6. ROC curves for over-sampling and under-sampling using naive Bayes on the rooftop detection task.

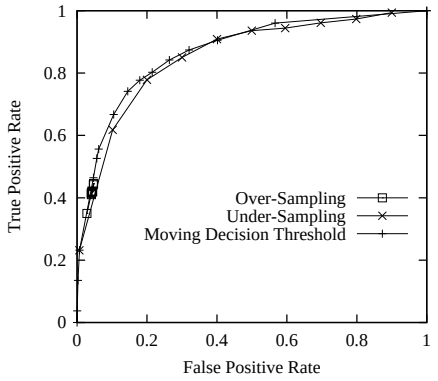

Figure 7. ROC curves for over-sampling and under-sampling using C5.0 on the rooftop detection task.

second, we included twenty copies, and so on. For each of these runs, we constructed classifiers and evaluated them using the test set. As before, we repeated this procedure ten times, plotting the average true-positive and false-positive rates as an ROC curve. The results for naive Bayes appear in Figure 6, and the results for C5.0 appear in Figure 7.

As we can see in Figure 6, the over- and under-sampling procedures produced ROC curves almost identical to that produced by varying the decision threshold of naive Bayes. Because the three curves are so similar, they are difficult to discern, but the over-sampling curve ranges between $(0.16, 0.74)$ and $(0.29, 0.85)$, while the under-sampled curve ranges between $(0.08, 0.59)$ and $(0.73, 0.97)$. In Figure 7, we see similar curves for C5.0, although these curves are not as tightly grouped as the ones for naive Bayes.

These results suggest that sampling produces classifiers similar to those produced by directly varying the decision threshold or cost matrix. A disadvantage of under-sampling is that in order to produce a desired point on an ROC, we may need to under-sample below the amount of available training data. Moreover, the under-sampled training data may not be sufficient for learning adequate concept descriptions.

Similarly, when over-sampling, to produce a desired point on an ROC curve, we may have to over-sample a data set so much that the time to learn becomes impractical. Indeed, with our rooftops domain, to produce a data set with a prior probability of 0.9 for the rooftop class, we would have to duplicate all examples of the positive class about 196 times, resulting in a data set with more than 153,000 positive examples. Depending on the algorithm, this will lead to unacceptable learning times, especially if we need to learn from several over-sampled data sets. We may be able to find some balance between over-sampling the positive class and under-sampling the negative class, but because we are under-sampling, we are using fewer examples for learning.

6.2. Selecting the Best Classifier

Regardless of how we produce ROC curves—by sampling, by moving the decision threshold, or by varying the cost matrix—the problem still remains of selecting the single best method and the single best classifier for deployment in an intelligent system. If the binormal assumption holds, the variances of the two distributions are equal, and error costs are the same, then the classifier at the apex of the dominant curve is the best choice.

When applying machine learning to real-world problems, rarely would one or more of these assumptions hold, but to select a classifier, certain conditions must exist, and we may need more information. If one ROC curve dominates all others, then the best method is the one that produced the dominant curve, which is also the curve with the largest area. This was generally true of our domains, but it is not true of others (Bradley, 1997; Provost & Fawcett, 2001). To select a classifier from the dominant curve, we need additional information, such as a target false-positive rate. On the other hand, if multiple curves dominate in different parts of the ROC space, then we can use the ROC Convex Hull method to select the optimal classifier (Provost & Fawcett, 2001).

7. Concluding Remarks

In this paper, we have examined how varying the decision threshold and ROC analysis helped with the problem of imbalanced data sets. We also presented evidence suggesting that over-sampling and under-sampling produces nearly the same classifiers as does

moving the decision threshold and varying the cost matrix. We reported these results for only one data set and for only two classification methods, but the analysis of Breiman et al. (1984) implies that sampling and adjusting the cost matrix have the same effect. Adjusting the cost matrix, in turn, has the same effect as moving the decision threshold. ROC analysis let us evaluate performance when varying any of these aspects of the learning method or its training.

For future work, we hope to explore further the connections between sampling and cost-sensitive learning for imbalanced data sets. We are also interested whether weighting examples or concept descriptions produces classifiers on the same ROC curve produced by moving the decision threshold or varying error costs. For instance, when boosting, are successive iterations producing classifiers on the same ROC curve, or generating a series of curves of increasing area? Indeed, ROC analysis may be a tool for developing a unified framework for understanding sampling, adjusting costs, moving decision thresholds, and weighting examples from underrepresented classes.

Acknowledgements

The author thanks the anonymous reviewers for their helpful comments, and Pat Langley and Bob Wagner for many discussions about the ideas expressed in this paper. Part of this research was conducted at the Institute for the Study of Learning and Expertise and in the Computational Learning Laboratory, Center for the Study of Language and Information, at Stanford University. This portion was supported by the Defense Advanced Research Projects Agency, under grant N00014-94-1-0746, administered by the Office of Naval Research, and by Sun Microsystems through a generous equipment grant. Part of this research was conducted in the Department of Computer Science at Georgetown University. This portion was partially supported by the National Institute of Standards and Technology under grant 60NANB2D0013.

References

Aha, D., Kibler, D., & Albert, M. (1991). Instance-based learning algorithms. *Machine Learning*, *6*, 37–66.

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, *12*, 387–415.

Beiden, S., Wagner, R., & Campbell, G. (2000). Components-of-variance models and multiple-bootstrap experiments: An alternative method for random-effects Receiver Operating Characteristic analysis. *Academic Radiology*, *7*, 341–349.

Bradley, A. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*, 1145–1159.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Boca Raton, FL: Chapman & Hall/CRC Press.

Cardie, C., & Howe, N. (1997). Improving minority class prediction using case-specific feature weights. *Proceedings of the Fourteenth International Conference on Machine Learning* (pp. 57–65). San Francisco, CA: Morgan Kaufmann.

Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

DeLong, E., DeLong, D., & Clarke-Peterson, D. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, *44*, 837–845.

Dorfman, D., & Alf, Jr., E. (1969). Maximum likelihood estimation of parameters of signal-detection theory and determination of confidence intervals—rating method data. *Journal of Mathematical Psychology*, *6*, 487–496.

Dorfman, D., Berbaum, K., & Metz, C. (1992). Receiver Operating Characteristic rating analysis: Generalization to the population of readers and patients with the Jackknife method. *Investigative Radiology*, *27*, 723–731.

Drummond, C., & Holte, R. (2000). Explicitly representing expected cost: An alternative to ROC representation. *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 198–207). New York, NY: ACM Press.

Duda, R., Hart, P., & Stork, D. (2000). *Pattern classification*. New York, NY: John Wiley & Sons. 2nd edition.

Fawcett, T., & Provost, F. (1997). Adaptive fraud detection. *Data Mining and Knowledge Discovery*, *1*, 291–316.

Grzymala-Busse, J., Zheng, X., Goodwin, L., & Grzymala-Busse, W. (2000). An approach to imbalanced data sets based on changing rule strength. *Learning from imbalanced data sets: Papers from the AAAI Workshop* (pp. 69–74). Menlo Park, CA: AAAI Press. Technical Report WS-00-05.

Hand, D., & Till, R. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, *45*, 171–186.

Japkowicz, N. (2000). Learning from imbalanced data sets: A comparison of various strategies. *Learning from imbalanced data sets: Papers from the AAAI Workshop* (pp. 10–15). Menlo Park, CA: AAAI Press. Technical Report WS-00-05.

Joshi, M., Kumar, V., & Agarwal, R. (2001). Evaluating boosting algorithms to classify rare classes: Comparison and improvements. *Proceedings of the First IEEE International Conference on Data Mining* (pp. 257–264). Los Alamitos, CA: IEEE Press.

Kubat, M., Holte, R., & Matwin, S. (1998). Machine learning for the detection of oil spills in satellite images. *Machine Learning*, *30*, 195–215.

Kubat, M., & Matwin, S. (1997). Addressing the curse of imbalanced training sets: One-sided selection. *Proceedings of the Fourteenth International Conference on Machine Learning* (pp. 179–186). San Francisco, CA: Morgan Kaufmann.

Langley, P., Iba, W., & Thompson, K. (1992). An analysis of Bayesian classifiers. *Proceedings of the Tenth National Conference on Artificial Intelligence* (pp. 223–228). Menlo Park, CA: AAAI Press.

Lin, C., & Nevatia, R. (1998). Building detection and description from a single intensity image. *Computer Vision and Image Understanding*, *72*, 101–121.

Ling, C., & Li, C. (1998). Data mining for direct marketing: Problems and solutions. *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD '98)* (pp. 73–79). Menlo Park, CA: AAAI Press.

Mac Namee, B., Cunningham, P., Byrne, S., & Corrigan, O. (2002). The problem of bias in training data in regression problems in medical decision support. *Artificial Intelligence in Medicine*, *24*, 51–70.

Maloof, M. (1999). *A machine learning researcher's foray into recidivism prediction* (Technical Report CS-99-02). Department of Computer Science, Georgetown University, Wasington, DC. http://www.cs.georgetown.edu/∼maloof/pubs/cstr-99-02.pdf.

Maloof, M. (2002). On machine learning, ROC analysis, and statistical tests of significance. *Proceedings of the Sixteenth International Conference on Pattern Recognition* (pp. 204–207). Los Alamitos, CA: IEEE Press.

Maloof, M., Beiden, S., & Wagner, R. (2002). *Analysis of competing classifiers in terms of components of variance of ROC accuracy measures* (Technical Report CS-02-01). Department of Computer Science, Georgetown University, Washington, DC. http://www.cs.georgetown.edu/∼maloof/pubs/cstr-02-01.pdf.

Maloof, M., Langley, P., Binford, T., & Nevatia, R. (1998). Generalizing over aspect and location for rooftop detection. *Proceedings of the Fourth IEEE Workshop on Applications of Computer Vision (WACV '98)* (pp. 194–199). Los Alamitos, CA: IEEE Press.

Maloof, M., Langley, P., Binford, T., Nevatia, R., & Sage, S. (to appear). Improved rooftop detection in aerial images with machine learning. *Machine Learning*. http://www.kluweronline.com/issn/0885-6125.

Maloof, M., Langley, P., Sage, S., & Binford, T. (1997). Learning to detect rooftops in aerial images. *Proceedings of the Image Understanding Workshop* (pp. 835–845). San Francisco, CA: Morgan Kaufmann.

McClish, D. (1989). Analyzing a portion of the ROC curve. *Medical Decision Making*, *9*, 190–195.

Metz, C. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, *VIII*, 283–298.

Metz, C. (1989). Some practical issues of experimental design and data analysis in radiological ROC studies. *Investigative Radiology*, *24*, 234–245.

Metz, C., Herman, B., & Shen, J.-H. (1998). Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Statistics in Medicine*, *17*, 1033–1053.

Metz, C., & Kronman, H. (1980). Statistical significance tests for binormal ROC curves. *Journal of Mathematical Psychology*, *22*, 218–243.

Mossman, D. (1999). Three-way ROCs. *Medical Decision Making*, *19*, 78–89.

Provost, F., & Fawcett, T. (2001). Robust classification for imprecise environments. *Machine Learning*, *42*, 203–231.

Provost, F., Fawcett, T., & Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms. *Proceedings of the Fifteenth International Conference on Machine Learning* (pp. 445–453). San Francisco, CA: Morgan Kaufmann.

Quinlan, J. (1993). *C4.5: Programs for machine learning*. San Francisco, CA: Morgan Kaufmann.

Schmidt, P., & Witte, A. (1988). *Predicting recidivism using survival models*. New York, NY: Springer-Verlag.

Swets, J., & Pickett, R. (1982). *Evaluation of diagnostic systems: Methods from signal detection theory*. New York, NY: Academic Press.

Thompson, M., & Zucchini, W. (1986). On the statistical analysis of ROC curves. *Statistics in Medicine*, *18*, 452–462.

Wagner, R., Beiden, S., & Metz, C. (2001). Continuous versus categorical data for ROC analysis: Some quantitative considerations. *Academic Radiology*, *8*, 328–334.

Woods, K., Kegelmeyer, W., & Bowyer, K. (1997). Combination of multiple classifiers using local accuracy estimates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*, 405–410.