A Machine Learning Researcher's Foray into Recidivism Prediction*

Marcus A. Maloof

Department of Computer Science
Georgetown University
Washington, DC 20057-1232
maloof@cs.georgetown.edu
http://www.cs.georgetown.edu/~maloof

CS-99-02

July 1999

Abstract

We discuss an application of machine learning to recidivism prediction. Our initial results motivate the need for a methodology for technique selection for applications that involve unequal but unknown error costs, a skewed data set, or both. Evaluation methodologies traditionally used in machine learning are inadequate for analyzing performance in these situations, although they arise frequently when addressing real-world problems. After discussing the problem of recidivism prediction and the particulars of our data set, we present experimental results that motivate the need to evaluate learning algorithm over a range of error costs. We then describe Receiver Operating Characteristic (ROC) analysis, which has been used extensively in signal detection theory for decades but has only recently begun to filter into machine learning research. With this new perspective, we revisit the recidivism prediction task and present results that contradict those obtained using a traditional method of evaluation.

Key words: recidivism prediction, evaluation methodology, ROC analysis, error costs, skewed data sets

*This paper is based on a presentation given at the 1998 Annual Meeting of the American Society of Criminology held on November 13 in Washington, DC.

1 Introduction

Recidivism prediction entails predicting if an individual is likely to recommit crime when released from prison (Schmidt & Witte, 1988). We can approach recidivism prediction as a survival analysis task in which we attempt to determine how long an individual will "survive" without committing crime. Alternatively, we can approach recidivism prediction as a detection task in which we attempt to predict, atemporally, if an individual will recommit crime. This is the approach we take here.

Schmidt and Witte (1988) describe an extensive comparison of parametric and nonparametric survival models using two data sets collected in North Carolina in 1978 and 1980. They also report results for individual predictions (p. 138), which involves simply predicting whether an individual is likely to recommit crime once released from prison. For this study, we also focused on individual predictions using the 1978 data set, which we describe in section 2.

When selecting a machine learning technique for an application, like recidivism prediction, we typically conduct a comparative evaluation and select the best performing method using performance measures such as predictive accuracy, learning time, recognition time, concept complexity, and the like. As we will see, employing a traditional evaluation methodology using percent correct as the measure of performance led to an incorrect conclusion: that Schmidt and Witte's proportional hazards model outperformed all of the machine learning techniques we surveyed.

As the machine learning community is now acknowledging (e.g., Bradley, 1997; Provost & Fawcett, 1997; Maloof, Langley, Binford, & Nevatia, 1998; Provost, Fawcett, & Kohavi, 1998), skewed data sets present a challenge because they often interfere with an algorithm's ability to learn to predict the minority class. And, for situations in which mistakes on the minority class are more expensive than mistakes on the majority class, we must compensate for the unequal distribution of training examples by either modifying the learning methods or altering the distribution of the examples in data set. In both cases, the aim is to focus more attention on the minority class, thus minimizing the total cost of mistakes.

In this report, we demonstrate, as others have, that using Receiver Operating Characteristic (ROC) analysis (Swets, 1988) to evaluate learning methods results in more principled comparisons, especially when confronted with skewed data sets, unequal but unknown error costs, or both. As our results show, when we applied ROC analysis, we obtained a very different picture of the performances of the classification methods than that obtained using a traditional evaluation methodology.

In the sections that follow, we describe the 1978 North Carolina data set and evaluate a set of learning techniques in the traditional manner, using percent correct. We then briefly discuss the issues of unequal cost and skewed data sets and present ROC analysis. Finally, we revisit the problem of recidivism prediction with this new perspective by evaluating several cost-sensitive learning techniques and using ROC analysis to evaluate performance.

2 The 1978 North Carolina Data Set

The 1978 data set was derived from a cohort of releasees from the North Carolina prison system from July 1, 1977 through June 30, 1978. The complete data sample ultimately consisted of 4,618 individuals and was randomly split into a training set (i.e., an estimation or an analysis sample) consisting of 1,540 individuals and a testing set (i.e., a validation sample) consisting of 3,078 individuals. In the training set, 570 individuals were recidivists and 970 were not. In the testing set, 1,151 individuals were recidivists and 1,927 were not.

Nine attributes characterized individuals and appear in table 1. There were six other attributes, but these were dropped because, based on a t-test, they were insignificant at the 5% level (Schmidt & Witte, 1988, pp. 86–87). Also, following Schmidt and Witte (1988), we scaled the TSERVED, AGE, and PRIORS attribute values by dividing them by 100, 1,000, and 10, respectively.

3 Evaluating the Methods Traditionally

Our original motivation for undertaking this study was to compare a variety of machine learning methods to Schmidt and Witte's proportional hazards model for individual predictions using the 1978 data set.

Table 1: Attributes used to characterize individuals.

Attribute	Description	Type
TSERVED	Time served in months	Continuous
AGE	Age in months at the time of release	Continuous
PRIORS	Number of previous incarcerations	Continuous
WHITE	Is the individual Caucasian?	Boolean
FELON	Was the sentence for a felony?	Boolean
ALCHY	Does individual's record indicate a serious problem with alcohol?	Boolean
JUNKY	Does individual's record indicate a serious problem with hard drugs?	Boolean
PROPTY	Was individual's sentence for a crime against property?	Boolean
MALE	Is the individual male?	Boolean

Therefore, we used a variety of machine learning techniques to construct classifiers using the training set. We then used these classifiers to predict the class of the individuals in the testing set. Based on the performances of each of the classifiers, we computed the true positive rate (i.e., the number of times the classifiers correctly identified a recidivist) and the true negative rate (i.e., the number of times the classifiers correctly identified a *non*-recidivist).

Using MLC++ (Kohavi & Sommerfield, 1997), we ran k-NN, for $k = 1, 3, \ldots, 9$, naive Bayes, and perceptron. We also ran CN2 (Clark & Niblett, 1989), a rule induction system, and C5.0, a decision tree induction system and the commercial successor of C4.5 (Quinlan, 1993). The true positive and true negatives rates for these methods and for Schmidt and Witte's proportional hazards model appear in table 2. As we can see, the machine learning methods did much worse than the proportional hazards model for this problem. The discussion in the following sections will explain why the machine learning methods fared so poorly.

4 Unequal Error Costs and Skewed Data Sets

In real-world applications, mistakes are seldom equal. For example, sometimes a false negative costs more than a false positive. Furthermore, we often encounter skewed data sets, in which the examples of one class far outnumber the examples in the other. As Breiman, Friedman, Olshen, and Stone (1984) point out, there is a strong relationship between the number of training examples in a class and the class's error cost: We can mitigate the bias against the minority class by duplicating its training examples, implying that the majority class of a skewed data set has an inherently higher cost of error than the minority class. Therefore, a problem arises when the error costs *implied* by the data set run counter to the *true* error costs inherent to an application. If not addressed, many learning methods construct classifiers that are biased toward

Table 2: Performance on a recidivism prediction task.

Classification Method	True Positive	True Negative
Proportional Hazards[a]	0.72	0.53
Nearest Neighbor	0.45	0.70
k-nn ($k = 3$)	0.44	0.72
k-nn ($k = 5$)	0.42	0.74
k-nn ($k = 9$)	0.41	0.77
k-nn ($k = 7$)	0.41	0.76
C5.0 (decision tree)	0.38	0.83
Naive Bayes	0.36	0.85
CN2 (decision rules)	0.20	0.92
Perceptron	0.0	1.0

[a] As reported by Schmidt and Witte (1988, p. 142).

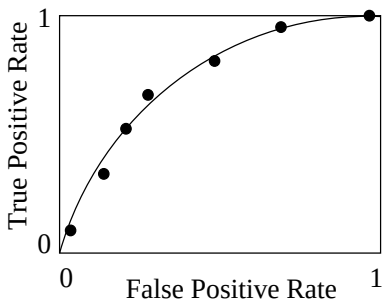

Figure 1: An idealized ROC curve.

predictions on the majority class. As a result, most errors will be on the more expensive minority class.

There are several ways to cope with this problem. For example, we can alter the distribution of the training examples (e.g., Cardie & Howe, 1997; Freund & Schapire, 1996; Kubat & Matwin, 1997) or we can change the way the learning methods treat instances from classes with different error costs (e.g., Lewis & Catlett, 1994; Maloof, Langley, Sage, & Binford, 1997; Bradley, 1997). If we have cost-sensitive learning methods and a cost analysis of our application, then we can construct minimum-cost classifiers and evaluate them by measure cost instead of errors (e.g., Pazzani et al., 1994). Unfortunately, such cost analyses often do not exist. A solution is to evaluate cost-sensitive methods over the range of all possible error costs, but, in this case, standard methodologies in machine learning are insufficient for evaluating performance. Fortunately, as we will see in the next section, a solution exists.

5 ROC Analysis

Receiver Operating Characteristic (ROC) analysis (Swets, 1988), which has existed for decades in the psychophysics literature, provides a means for evaluating cost-sensitive learning methods over a range of costs. Briefly, this approach plots the false positive and true positive rates produced by a classifier operating under a specific set of cost parameters in an *ROC graph* (see figure 1). The lower left corner represents the situation in which false positives are maximally expensive. Similarly, the upper right corner indicates that false negatives are maximally expensive. By varying the setting of the cost parameters from one extrema to the other and conducting learning runs, we sweep out an *ROC curve* that represents a method's accuracy trade-off. The point (0, 1) is where classification is perfect, with a false positive rate of zero and a true positive rate of one, and so we want curves that "push" toward this corner. Traditional ROC analysis uses area under the curve as the measure of performance, which we can approximate by applying the trapezoid rule, and for which Hanley and McNeil (1982) present analytical significance tests. Typically, we prefer curves that cover larger areas, but this measure can be problematic if two curves have the same area but dominate in different regions of the space. In this case, other performance measures are more appropriate (Provost & Fawcett, 1997).

6 The Recidivism Prediction Task Revisited

To demonstrate the value of ROC analysis, we present new results for the recidivism prediction task using the 1978 data set. The distribution of the training examples was 27.5% recidivist and 72.5% non-recidivist, which is not skewed as severely as other reported data sets (e.g., Cardie & Howe, 1997; Maloof et al., 1997), but it was enough to adversely impact the performance of all of the machine learning methods, most notably the perceptron algorithm.

For this problem, we have only an informal notion of error costs: Mistakes on the positive (i.e., the minority) class are more expensive than those on the negative class. Using the same experimental design as before, we ran C5.0 and cost-sensitive versions of naive Bayes and nearest neighbor (Maloof et al., 1997), and generated the ROC curves that appear in figure 2. We accomplished this by running each method over

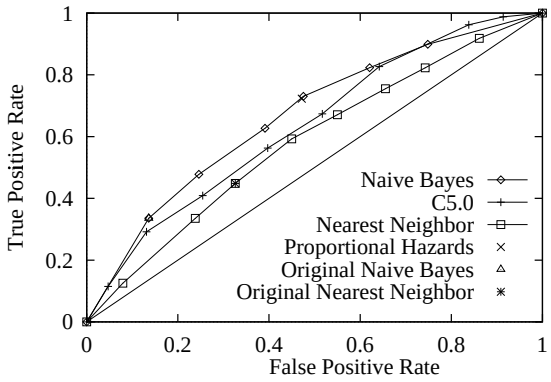

Figure 2: ROC curves for recidivism prediction.

a range of error costs and plotting the true positive and false positive rates for each cost parameter. We also plotted the original points from table 2 for the purposes of comparison.

By finding the appropriate cost parameters, we were able to produce a naive Bayesian classifier with accuracy equal to that of the proportional hazards model. Although we do not have a cost-sensitive version of proportional hazards, we anticipate that its ROC curve would be similar to the one for naive Bayes.

We also concluded the naive Bayes performs better on this task than nearest neighbor and C5.0, since its ROC curve covers a larger area, thus contradicting our original findings: that proportional hazards, nearest neighbor, and C5.0 outperformed naive Bayes. For this experiment, naive Bayes produced an ROC curve with an area of 0.667, C5.0 produced a curve of area 0.635, and nearest neighbor produced one of area 0.584. If we again look at the original points, we see that the one produced by naive Bayes was simply lower on its ROC curve than that produced by nearest neighbor and C5.0, which explains why these performances appeared to be better.

We speculate that each method operates under the influence of an inherent but unknown set of cost parameters, which is associated with its inductive bias. Therefore, naive Bayes, while achieving better performance over a range of costs, performed worse than the other methods in the original condition simply because its inherent cost parameters were not optimal for this problem. The danger is that we might have concluded that proportional hazards was better than the other two methods, or that nearest neighbor was better than naive Bayes. However, when we generated the ROC curves, they revealed a clearer picture of performance.

7 Discussion

There are several questions that arose during this study. The following sections discuss why the proportional hazards model was not affected by the skewed data set, and how we can incorporate a error costs into the proportional hazards model.

7.1 The Affect of Skew on Proportional Hazards

When evaluating their model for individual predictions, Schmidt and Witte computed the false positive and true positive rates using a rule that states that "the proportion of the individuals predicted to fail should be equal to the proportion of the group that we expect to fail" (Schmidt & Witte, 1988, p. 142), which compensated for the skewed distribution of the training examples. They used the failure rate of 0.366 to compute the proportion of recidivists and non-recidivists and then derived the false negative and true negatives rates. It is difficult to apply this same analysis to our machine learning methods because they have nothing equivalent to a failure rate.

With naive Bayes, we might be tempted to interpret the prior probability of the recidivist class as a failure rate, which is 27.5% and slightly below the failure rate of 36.6% for the proportional hazards model. This is not quite correct since the prior probability does take into account information about the characteristics of

the individuals in the cohort, as does the failure rate. Furthermore, under this interpretation, naive Bayes achieves a true positive rate of 0.76 and a true negative rate of 0.37, which is not on its ROC curve, but well within its interior.

7.2 Incorporating Error Costs into Proportional Hazards

As Bradley (1997) states, the misclassification cost of a classifier on a data set is usually defined as

$$\text{Cost} = F_p \cdot C_{F_p} + F_n \cdot C_{F_n}$$

where F_p is the number of false positives, C_{F_p} is the cost of a false positive, F_n, is the number of false negatives, and C_{F_n} is the cost of a false negative. If we are fortunate enough to know the costs involved, C_{F_p} and C_{F_n}, then we can measure the cost of a single classifier, or we can use the cost to select the classifier that minimizes the cost of mistakes.

Because cost analyses often do not exist, we may alternatively specify system performance in terms of the desired false positive rate, $\Pr(F_p)$, and minimize the false negative rate, $\Pr(F_n)$, using the Neyman-Pearson method (Fukunaga, 1990) by varying the decision threshold.

With proportional hazards, we compute the probability that an individual will fail at some time t. As described above, Schmidt and Witte use the failure rate of 0.366 as the decision threshold. Therefore, we should be able to vary this threshold and produce an ROC curve for proportional hazards. Indeed, this would be a interesting direction for future work.

8 Concluding Remarks

By generating an ROC curve, we effectively find classifiers for all possible error costs, which gives us a much clearer picture of a method's performance, than do traditional evaluation methods. We have given one compelling example of how ROC analysis proved useful in revealing previously hidden aspects of performance.

When building an application, unless the system will know the error costs for its decisions, we will have to select a fixed-cost classifier. ROC analysis provides several ways in which to pick such a classifier. First, we should select the method that produces the ROC curve with the largest area, provided that area is an appropriate measure of performance (Provost & Fawcett, 1997). To choose a fixed-cost classifier, we could then pick the classifier that achieves a certain true positive rate. Alternatively, we could choose the classifier that operates under the influence of a specific set of costs. Finally, if we know nothing about the error costs or the desired hit rate, then we can argue that the classifier nearest the apex of the curve is, in some sense, the "best," since this classifier is closest to the point of perfect classification (0, 1).

The utility of ROC analysis for situations in which we have unequal but unknown error costs or skewed data sets is evident. Although we have not conducted extensive studies using data sets that are not skewed, we wonder if compiling ROC curves will reveal behavior similar to that in our recidivism prediction study. In these situations, if a method operating under its inherent cost parameters consistently produces suboptimal points along its ROC curve, then this has profound implications for the evaluation and selection of machine learning techniques. Indeed, it brings into question the validity of past studies that used predictive accuracy as the sole measure of performance.

Acknowledgements. The author thanks Pam Lattimore, Pat Langley, Eric Bloedorn, and Stephanie Sage for valuable comments and advice. This research was conducted at the Institute for the Study of Learning and Expertise and in the Center for the Study of Language and Information at Stanford University. This work was partially supported by ISLE, 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. The author also thanks the Department of Computer Science at Georgetown University for its support of this work.

References

Bradley, A. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*(7), 1145–1159.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees.* Belmont, CA: Wadsworth.

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

Clark, P., & Niblett, T. (1989). The CN2 induction algorithm. *Machine Learning, 3*, 261–284.

Freund, Y., & Schapire, R. (1996). Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning* (pp. 148–156). San Francisco, CA: Morgan Kaufmann.

Fukunaga, K. (1990). *Introduction to statistical pattern recognition.* Boston, MA: Academic Press.

Hanley, J., & McNeil, B. (1982). The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve. *Radiology, 143*, 29–36.

Kohavi, R., & Sommerfield, D. (1997). *MLC++: machine learning library in C++.* Mountain View, CA. (http://www.sgi.com/Technology/mlc)

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

Lewis, D., & Catlett, J. (1994). Heterogeneous uncertainty sampling for supervised learning. In *Proceedings of the Eleventh International Conference on Machine Learning* (pp. 148–156). San Francisco, CA: Morgan Kaufmann.

Maloof, M., Langley, P., Binford, T., & Nevatia, R. (1998). Generalizing over aspect and location for rooftop detection. In *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., Sage, S., & Binford, T. (1997). Learning to detect rooftops in aerial images. In *Proceedings of the Image Understanding Workshop* (pp. 835–845). San Francisco, CA: Morgan Kaufmann.

Pazzani, M., Merz, C., Murphy, P., Ali, K., Hume, T., & Brunk, C. (1994). Reducing misclassification costs. In *Proceedings of the Eleventh International Conference on Machine Learning* (pp. 217–225). San Francisco, CA: Morgan Kaufmann.

Provost, F., & 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* (pp. 43–48). Menlo Park, CA: AAAI Press.

Provost, F., Fawcett, T., & Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms. In *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. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*, 1285–1293.