Slide 1

A Machine Learning Researcher's Foray into Recidivism Prediction

Mark Maloof
Department of Computer Science
Georgetown University

maloof@cs.georgetown.edu
http://www.cs.georgetown.edu/~maloof

Annual Meeting of the American Society of Criminology
Washington, DC
13 November 1998

Slide 2

Overview of Presentation

- Empirical study
  - examined machine learning methods
  - used Schmidt and Witte's 1978 NC data set
- Issue of unequal error costs and skewed data sets
- Recidivism prediction revisited
- Summary and future directions

Copyright ©1998 Marcus A. Maloof

Slide 3

Summary of Activities

- Replicated Schmidt and Witte's (1984) experiments using a variety of machine learning methods
- Concentrated on individual predictions (p. 138)
- Ran algorithms using the analysis and validation sets

Schmidt, P. and Witte, A.D. 1984. *Predicting recidivism using survival models.* New York, NY: Springer.

Slide 4

Results for Individual Predictions

Classification Method	True Positive	True Negative
Proportional Hazards*	72	53
Nearest Neighbor	45	70
k-nn ($k = 3$)	44	72
k-nn ($k = 5$)	42	74
k-nn ($k = 9$)	41	77
k-nn ($k = 7$)	41	76
C5.0 (decision tree)	38	83
Naive Bayes	36	85
CN2 (decision rules)	20	92
Perceptron	0	100

*As reported by Schmidt and Witte (1988, p. 142).

Copyright ©1998 Marcus A. Maloof

Slide 5

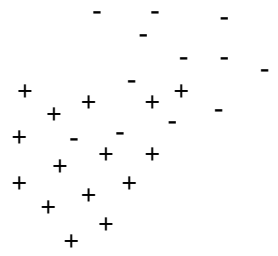
Unequal Error Costs and Skewed Data Sets

- For many applications, such as vision, fraud detection, computer security, medical diagnosis, mistakes are not equal
- Example: it is better tell someone they have cancer when they don't than to tell them they don't have cancer when they do
- Why? Additional tests will show if someone has cancer
- Most learning algorithms assume equal error costs among classes
- Skewed data sets complicate this issue:
  - By adding examples to a class, we can increase its prior probability *and* its error cost (Breiman et al. 1984)
  - As a result, many methods learn to predict the majority class
  - But it is often the *minority class* that is more important!
  - NC analysis set: 27.5% recidivist; 72.5% non-recidivist

Slide 6

Using a Cost Heuristic to Change the Decision Boundary

- Incorporate a cost heuristic into a method to change the decision boundary at which it selects one class over the other
- This lets us bias the algorithm toward the less costly class
- If we know the actual costs of errors, we can find the minimum-cost classifier (e.g., Pazzani et al. 1994)
- Problem: We often do not know the costs of errors



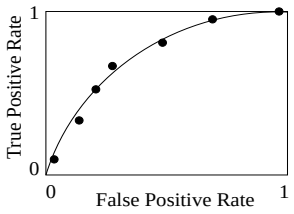
Copyright ©1998 Marcus A. Maloof

Slide 7

ROC Analysis

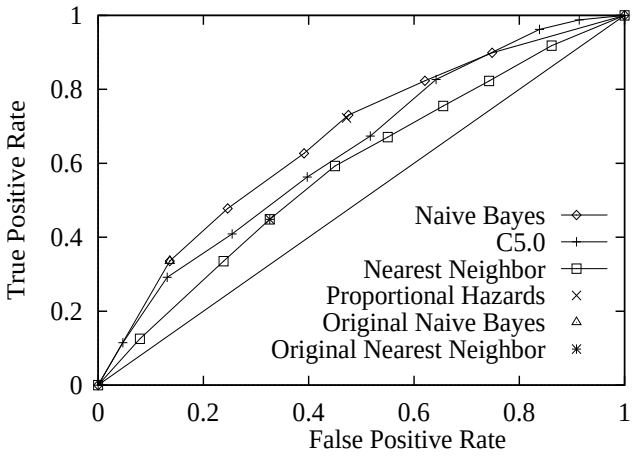
(Receiver Operating Characteristic)

- Lets us evaluate performance for a variety of error costs
- ROC curve plots the true positive and false positive rates over a range of misclassification costs for a given method
- The point (0, 1) is where classification is perfect, so we want curves that "push" toward this corner
- Traditional ROC analysis uses area under the curve as the measure of performance



Slide 8

ROC Curve for Individual Predictions



Copyright ©1998 Marcus A. Maloof

Slide 9

Summary and Future Directions

- A naive Bayesian classifier was able to match the performance of the proportional hazards model
- Why was the proportional hazards model able to find a better accuracy trade-off than the machine learning methods?
  - Probably due to maximum likelihood estimation
- Conjecture: The ROC curves for naive Bayes and the proportional hazards models are identical
- We see two future directions:
  1. empirical studies that take into account error costs, and
  2. alternative methods for survival analysis

Slide 10

Empirical Studies Taking into Account Error Costs

- We've established the importance of taking into account error costs for this and other problems
- Is there another cost-sensitive method that will perform better than proportional hazards and naive Bayes?
- How could we introduce costs into proportional hazards?
  Has this been done?

Copyright ©1998 Marcus A. Maloof

Slide 11

Alternative Methods for Survival Analysis

- Ohno-Machado (1996) showed that a connectionist model of survival outperformed a proportional hazards model for an AIDS data set
- Used "sequential neural networks," but her idea generalizes to other methods, like naive Bayes
- Basic idea:
  - For a time interval t, train a neural network to predict survival
  - Do the same thing for time $t + 1$, but use the prediction of the time t network as an input to the time $t + 1$ network

Ohno-Machado, L. 1996. Medical applications of artificial neural networks: connectionist models of survival. Ph.D. Dissertation, Stanford University, Stanford, CA.

Copyright ©1998 Marcus A. Maloof