Interpretable Boosted Naïve Bayes Classification

Greg Ridgeway, David Madigan, Thomas Richardson
Department of Statistics
Box 354322, University of Washington
Seattle, WA 98195-4322
{greg, madigan, tsr}@stat.washington.edu

John O'Kane
Department of Orthopedics
Sports Medicine Clinic, Box 354060
University of Washington
Seattle, WA 98195-4060
jokane@u.washington.edu

Abstract

Voting methods such as boosting and bagging provide substantial improvements in classification performance in many problem domains. However, the resulting predictions can prove inscrutable to end-users. This is especially problematic in domains such as medicine, where end-user acceptance often depends on the ability of a classifier to explain its reasoning. Here we propose a variant of the boosted naïve Bayes classifier that facilitates explanations while retaining predictive performance.

Introduction

Efforts to develop classifiers with strong discrimination power using voting methods have marginalized the importance of comprehensibility. Bauer and Kohavi [1998] state that "for learning tasks where comprehensibility is not crucial, voting methods are extremely useful." However, as many authors have pointed out, problem domains, such as credit approval and medical diagnosis, do require interpretable as well as accurate classification methods. For instance, Swartout [1983] commented that "trust in a system is developed not only by the quality of the results but also by clear description of how they were derived. … In addition to providing diagnoses or prescriptions, a consultant program must be able to explain what it is doing and why it is doing it." In this note we present a boosted naïve Bayes classifier with both competitive discrimination ability and transparent reasoning.

The next section provides a very brief introduction to boosting. Then we describe our proposed boosted, interpretable naïve Bayes classifier while the last section examines its performance empirically.

Boosting

Boosting describes a general voting method for learning from a sequence of models. Observations poorly modeled by H_t receive greater weight for learning H_{t+1}. The final boosted model is a combination of the predictions from each H_t where each H_t is weighted according to the quality of its classification of the training data. Freund and Schapire [1995] presented a boosting algorithm that empirically has yielded reduction in bias, variance, and misclassification rates with a variety of base classifiers and problem settings. The AdaBoost (adaptive boosting) algorithm of Freund and Schapire involves the following steps.

The data for this problem take on the form $(X,Y)_i$ where $Y_i \in \{0,1\}$. Initialize the weight of each observation to $w_i^{(1)} = \frac{1}{N}$. For t in 1 to T do the following…

1. Using the weights, learn model $H_t(x_i) : X \rightarrow [0,1]$.
2. Compute $\varepsilon_t = \sum_{i=1}^{N} w_i^{(t)} \left| y_i - H_t(x_i) \right|$ as the error for H_t.
3. Let $\beta_t = \frac{\varepsilon_t}{1-\varepsilon_t}$ and update the weights of the observations as $w_i^{(t+1)} = w_i^{(t)} \beta_t^{1-\left|y_i - H_t(x_i)\right|}$. This scheme increases the weights of observations poorly predicted by H_t.
4. Normalize $w^{(t+1)}$ so that they sum to one.

To classify a new observation Freund and Schapire suggest combining the classifiers as:

$$H(x) = \frac{1}{1 + \prod_{t=1}^{T} \beta_t^{2r(x)-1}} \quad \text{where} \quad r(x) = \frac{\sum_{t=1}^{T} (\log \frac{1}{\beta_t}) H_t(x)}{\sum_{t=1}^{T} (\log \frac{1}{\beta_t})}.$$

Boosting the naïve Bayes classifier involves

substituting in the above algorithm $H_t = P_t(Y=1|X)$ where estimation of $P_t(Y=1|X)$ assumes that

$$P_t(Y=1|X) \propto P_t(Y=1)P_t(X_1|Y=1)\cdots P_t(X_d|Y=1).$$

The boosted naïve Bayes classifier has proven effective – see, for example, Bauer and Kohavi [1998]. Elkan's application of boosted naïve Bayes won first place out of 45 entries in the data mining competition KDD'97 (Elkan [1997]). However, while the regular naïve Bayes approach leads to elegant and effective explanations (see, for example, Madigan, *et al* [1996] and Becker, *et al* [1997]), the AdaBoost-ed version destroys this feature.

Boosting and Weights of Evidence

Under the naïve Bayes assumption, writing the log-odds in favor of $Y=1$ we obtain the following:

$$\begin{aligned}\log\frac{P(Y=1|X)}{P(Y=0|X)} &= \log\frac{P(Y=1)}{P(Y=0)} + \sum_{j=1}^{d}\log\frac{P(X_j|Y=1)}{P(X_j|Y=0)} \\ &= w_0 + \sum_{j=1}^{d} w_j(X_j)\end{aligned}$$

The w_j are the weights of evidence described by Good [1965]. A positive $w_j(X_j)$ indicates that the state of X_j is evidence in favor of the hypothesis that $Y=1$. A negative weight is evidence for $Y=0$. Spiegelhalter and Knill-Jones [1984] advocate the use of weights of evidence extensively in medical diagnosis and propose evidence balance sheets as a means of viewing the reasoning process of the naïve Bayes classifier. Madigan, *et al* [1996] and Becker, *et al* [1997] develop this idea further.

In what follows, we derive a "boosted weight of evidence" that the explanation facilities of Madigan, *et al* [1996] and Becker, *et al* [1997] can use directly.

Writing the AdaBoost combined classifiers, $H(x)$, in the form of a log-odds simplifies the combination procedure.

$$\begin{aligned}\log\frac{H(x)}{1-H(x)} &= -\log\prod_{t=1}^{T}\beta_t^{2r(x)-1} \\ &= (1-2r(x))\sum_{t=1}^{T}\log\beta_t \\ &= \sum_{t=1}^{T}(\log\beta_t)(1-2P_t(Y=1|X))\end{aligned}$$

Using the fact that

$$P_t(Y=1|X) = \frac{1}{1+\frac{P(Y=0|X)}{P(Y=1|X)}} = \left(1+e^{-\log\frac{P_t(Y=1|X)}{P_t(Y=0|X)}}\right)^{-1}$$

we can rewrite the log-odds of the combined classifiers as a function of the log-odds of each $P_t(\bullet)$:

$$= \sum_{t=1}^{T}(\log\beta_t)\left(1-2\left(1+e^{-\log\frac{P_t(Y=1|X)}{P_t(Y=0|X)}}\right)^{-1}\right).$$

Substituting the Taylor approximation to the sigmoid function $\left(\frac{1}{1+e^{-x}} = \frac{1}{2}+\frac{1}{4}x-\frac{1}{48}x^3+O\left(x^5\right)\right)$ up to the linear term produces a linear combination of the log-odds from each boosted naïve Bayes classifier:

$$\approx \sum_{t=1}^{T}(\tfrac{1}{2}\log\tfrac{1}{\beta_t})\log\frac{P_t(Y=1|X)}{P_t(Y=0|X)}.$$

The linear approximation to the sigmoid function is exact at 0 and loses precision as the argument moves away from 0. On the probability scale, the two functions agree at $P(Y=1|X) = ½$. This is precisely the point where classification is the most in doubt. On the probability interval [0.25, 0.75] the absolute error of the approximation is less than 0.025. In the region near the extremes the approximation preserves the sign of the AdaBoost combined hypothesis. Although the probability estimates will differ in this case, both methods will make the same classification.

Finally, to help maintain some probabilistic interpretation of the final model and to remove some of the effect of the number of boosting iterations, we normalize the classifier weights $\left(\frac{1}{2}\log\frac{1}{\beta_t}\right)$. Intuitively, each probability model casts a log-odds vote for or against classification into $Y=1$ where each models' vote is weighted according to its quality. So letting $\alpha_t = \frac{\log\frac{1}{\beta_t}}{\sum\log\frac{1}{\beta_t}}$ and assuming a naïve Bayes model for $P_t(Y=1|X)$, a boosted estimate for the log-odds in favor of $Y=1$ is:

$$\begin{aligned}&\sum_{t=1}^{T}\alpha_t\log\frac{P_t(Y=1)}{P_t(Y=0)} + \sum_{j=1}^{d}\sum_{t=1}^{T}\alpha_t\log\frac{P_t(X_j|Y=1)}{P_t(X_j|Y=0)} \\ &\quad = \text{boosted prior weight of evidence} + \\ &\qquad \sum_{j=1}^{d}\text{boosted weight of evidence from } X_j\end{aligned}$$

This again is just a naïve Bayes classifier where the estimates of the weights of evidence have been biased by boosting. The sigmoid function transforms the boosted log-odds to a boosted predicted probability as

$$P(Y=1|X) = \frac{1}{1+\exp(-\log\text{odds})}.$$

Performance

The substantial changes to the classifier combination step proposed in the previous section may have produced a voting method with different discrimination properties. In this section we show empirically on five datasets that the misclassification rates for this new method are comparable to those generated by the AdaBoost algorithm. We note that each one of these datasets concerns a problem domain where human understanding of the machine reasoning process is likely to be important.

	N	Predictors	Positive cases
Knee diagnosis	99	26	44%
Diabetes	768	8	35%
Credit approval	690	15	44%
Coronary artery disease	303	13	44%
Breast tumors	699	9	34%

Table 1 : Datasets used for comparison

In a retrospective study, O'Kane, *et al* [1998] use the weight of evidence boosted naïve Bayes classifier to diagnose candidates for knee surgery at a sports medicine clinic. They present boosted parameter estimates and example evidence balance sheets that physicians can easily interpret and utilize in diagnosis. A prospective study is in progress. The remaining datasets are from the UCI repository (Merz and Murphy [1998]).

	Naïve Bayes	AdaBoost	Weight of evidence
Knee diagnosis	14.0% (5.0%)	13.8% (5.5%)	13.4% (5.7%)
Diabetes	25.0% (2.0%)	24.4% (2.5%)	24.4% (2.6%)
Credit approval	16.8% (2.0%)	15.5% (2.1%)	15.5% (2.1%)
Coronary artery disease	18.4% (3.0%)	18.3% (3.2%)	18.3% (3.3%)
Breast tumors	3.9% (1.0%)	3.8% (1.0%)	3.8% (1.0%)

Table 2: Misclassification rates (standard deviation)

For each dataset we created 100 training sets randomly composed of two-thirds of the data. Then we compared the misclassification rate of naïve Bayes, AdaBoost naïve Bayes, and the proposed boosted weight of evidence naïve Bayes on the remaining one-third of the cases comprising the test set. Cross-validation on the training data yielded an estimate of the optimal number of boosting iterations. Table 2 compares the misclassification rates of the three methods. Both boosting methods demonstrate marginal absolute improvements in misclassification rate for most of the datasets. However, the most important result from Table 2 is that boosted weight of evidence naïve Bayes almost perfectly matches the performance of the AdaBoost method.

We have also examined aspects of the performance of the boosted weight of evidence naïve Bayes classifier that go beyond just misclassification rate. The so-called "Brier Score" is a proper scoring rule that considers both discrimination ability and calibration:

$$\overline{B} = \tfrac{1}{N} \sum (y_i - \hat{P}(Y_i = 1))^2$$

Yates [1982] and Spiegelhalter [1986] provide an extensive discussion of this scoring rule that has its roots in meteorology. Table 3 indicates that boosting generally decreases the mean Brier score and again the boosted weight of evidence naïve Bayes classifier is competitive.

	Naïve Bayes	AdaBoost	Weight of evidence
Knee diagnosis	0.107 (0.040)	0.108 (0.038)	0.103 (0.036)
Diabetes	0.170 (0.010)	0.164 (0.011)	0.164 (0.011)
Credit approval	0.127 (0.020)	0.112 (0.012)	0.113 (0.012)
Coronary artery disease	0.135 (0.020)	0.132 (0.016)	0.130 (0.021)
Breast tumors	0.035 (0.009)	0.035 (0.009)	0.034 (0.009)

Table 3: Mean Brier score (standard deviation)

When a probabilistic prediction model assigns, e.g., a 30% probability to a positive outcome, one would like 30% of such cases to actually have a positive outcome. Various authors have proposed *calibration* scores and plots that attempt to represent this property. Calibration plots (Copas [1983]) show whether probabilistic predictions are calibrated. Figure 1 displays an example calibration plot for a random split of the data. The perfectly calibrated predictor will have a straight line. Across the five datasets, the plots generally show that the boosting methods provide improved calibration, but the two boosting methods are indistinguishable.

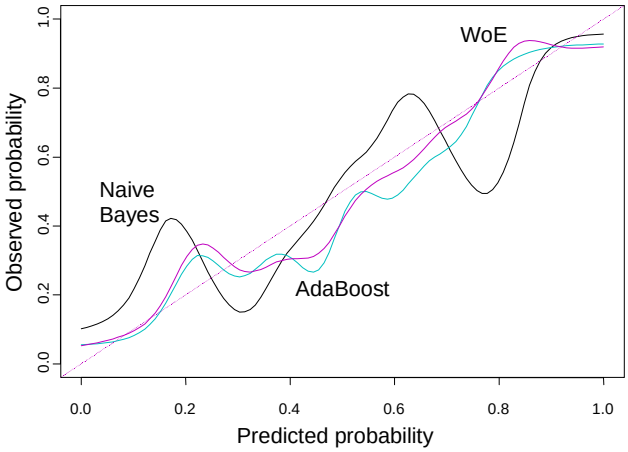

Figure 1: Example calibration plots (credit data)

Further research may show that boosting smoothes predictions in such a way that improves their calibration.

Summary

The proposed boosting by weight of evidence method offers an easily interpretable model with discrimination power equivalent to that of AdaBoost. More importantly, this shows that voting methods do not necessarily have to dispense with interpretability in order to obtain classification strength.

Acknowledgements

A grant from the National Science Foundation supported this work (DMS 9704573).

References

Bauer, E. and R. Kohavi [1998] An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. *Machine Learning*, vv, 1-33.

Becker, B., R. Kohavi, and D. Sommerfield. [1997] Visualizing the Simple Bayesian Classifier. *KDD 1997 Workshop on Issues in the Integration of Data Mining and Data Visualization*.

Copas, J.B. [1983]. Plotting p against x. *Applied Statistics*, **32**, 25-31.

Elkan, C. [1997]. Boosting and Naïve Bayes Learning. Technical Report No. CS97-557, September 1997, UCSD.

Freund, Y. and R. Schapire [1997] A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences*, 55(1):119-139.

Good, I. J. [1965] *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*, MIT Press.

Madigan, David, K. Mosurski, and R.G. Almond [1996] Explanation in Belief Networks. *Journal of Computational and Graphical Statistics*, **6**, 160-181.

Merz, C.J., and P.M. Murphy [1998]. UCI Repository of machine learning databases. Irvine, CA: University of California, Department of Information and Computer Science.
http://www.ics.uci.edu/~mlearn/MLRepository.html

O'Kane, John, G. Ridgeway, and D. Madigan [1998]. *In submission*.

Spiegelhalter, D.J. and R.P. Knill-Jones [1984] Statistical and Knowledge-based Approaches to Clinical Decision-support Systems, with an Application in Gastroenterology (with discussion). *Journal of the Royal Statistical Society (Series A)*, **147**, 35-77.

Spiegelhalter, David J. [1986] Probabilistic Prediction in Patient Management and Clinical Trials. *Statistics in Medicine*, **5**, 421-433.

Swartout, W. [1983] XPLAIN: A system for creating and explaining expert consulting programs. *Artificial Intelligence*, **21**, 285-325.

Yates, J.F. [1982] External correspondence: decomposition of the mean probability score. *Organisational Behaviour and Human Performance*, **30**, 132-156.