Confidence-based Tuning of Nomogram Predictions

Tony Mancill
Washington State University Vancouver
14204 NE Salmon Creek Ave.
Vancouver, WA 98686
tmancill@wsu.edu

Scott A. Wallace
Washington State University Vancouver
14204 NE Salmon Creek Ave.
Vancouver, WA 98686
wallaces@vancouver.wsu.edu

Abstract

Instance classification using machine learning techniques has numerous applications, from automation to medical diagnosis. In many problem domains, such as spam filtering, classification must be performed quickly across large datasets. In this paper we begin with machine learning techniques based on the naïve Bayes classification and attempt to improve classification performance by taking into account attribute confidence intervals. Our prediction functions operate over nominal datasets and retain the asymptotic complexity of one-pass learning and prediction functions. We present preliminary results indicating a modest, albeit inconsistent improvement over the naïve Bayes classifier alone.

Introduction

The promise of machine learning is many-fold, ranging from alleviating repetitive and mundane tasks such as spam filtering to assimilating vast and disparate corpora of knowledge necessary to perform medical diagnosis. The ability to correctly classify an instance based on a (sometimes incomplete) set of attributes is central to many applications, and improvements in classifier performance obviously increase the value of these techniques. Furthermore, many endeavors require methods that are efficient in terms of memory and time complexity, perhaps due to the sheer number of instances to be classified (again, email filtering) or limitations of computational resources. Naïve Bayes classification (henceforth NB) is commonly employed as the basis of classification in such domains both because of its speed and performance (Zhang 2004), and so we use it both as a basis for performance comparison and as the core classification algorithm for our research.

For an instance of a classification problem, NB employs the product of the conditional probabilities of each attribute value pair in an instance times the overall likelihood of a given class to find the maximum-likelihood classification hypothesis (Russell & Norvig 2002, p. 718). A very similar technique is to base the prediction on the sum of the relative frequencies expressed as log odds ratios (LOR) of attribute value pairs with respect to the class hypothesis being tested. This approach is taken in (Možina *et al.* 2004), which is also the source of the confidence interval calculation used in our research. Both techniques allow the requisite conditional probability and log odds ratio and confidence interval calculations to be performed once for the training set. Once computed, we test and compare multiple hypotheses for how to incorporate the attribute confidence interval (CI) into the predicted class for instances in the test dataset.

Datasets

Our experiment utilizes a variety of datasets obtained from the UCI Machine Learning Repository (Asuncion & Newman 2007), all of which are already nominal or have been discretized using the unsupervised discretization module in Weka (Witten & Frank 2005). The datasets range from having between 3 and 34 attributes used for binary and ternary classifications. The number of instances range from 24 to over 8000 in the *mushroom* dataset.

Predictors

In addition to the standard NB and log odds ratio predictors, we explore several strategies to try to improve the performance of the classifier. The first is to either ignore attributes with poor CIs or augment the contribution of attributes with strong CIs. In lieu of an absolute measure of goodness for confidence, strength is determined by the relative ranking of attribute value pairs by inverse confidence (since numerically smaller values indicate better confidence). For predictors using LORs, the CI represents the expected error in the LOR and therefore can be added or subtracted from the LOR points. This can be done on a per-attribute basis, or the variance of the CIs for all attribute values in the instance can be computed and applied to sum of the LOR terms. Another strategy is to scale the CI to fall within the range of LOR, that is, to treat it as an attribute, and then explore weighting the LOR and CI components. Permutations based on these ideas are explored resulting in 14 strategy variations; we introduce those relevant to the evaluation discussion here:

nbc standard naïve Bayes classifier using the maximum a posteriori (or MAP).

cautious LOR-based and similar to **contrarian** - the effect of the CI is added or subtracted such that magnitude of the prediction is minimized.

reckless LOR-based - the effect of the CI is added or subtracted from each attribute such that magnitude of the attribute is maximized.

contrarian LOR-based - the predicted error in the CIs is subtracted from the LOR prediction if it is positive (*i.e.* towards the goal outcome); otherwise it is added. This may change the sign of the prediction.

```
def contrarian(avpl, goal):
    pred = 0
    for avp in avpl:
        pred += getLor(avp, goal)
    predErr = getCi(avpl, goal)
    if pred >= 0:
        return lorToPts(pred - predErr)
    else:
        return lorToPts(pred + predErr)
```

Evaluation

Evaluation of the prediction functions is done using a standard 10-fold cross-fold validation which splits the dataset into training and test sets, or folds, and then the results across all folds are averaged. We ensure baseline operation of our software via a calibration step in which the output of our NB is manually compared to the *NaiveBayesSimple* classifier included in Weka; for our CI calculations we appeal to the nomogram visualization module of the Orange data mining software suite (Zupan & Demšar 2004). All CI calculations are for a confidence level $(1 - \alpha)$ of 95%.

Predictor results are compared by calculating the confusion matrix statistics for the test set: precision, specificity, recall, negative predictive value (precision for non-goal predictions) and accuracy, along with the F-measure (f_1), which is the harmonic mean of precision and recall. All statistics range between 0 and 1. We say that a predictor function *dominates* another if all statistics are greater. A predictor is *acceptable* relative to another predictor if the average of its confusion matrix statistics is equal to or greater than that of the other predictor. These terms are used to compare our hybrid predictors to the *nbc* predictor. An example of predictor performance is depicted in Figure 1, where we can see that the *nbc* is dominated by the *contrarian* predictor, *reckless* proves acceptable despite scoring lower for the recall metric, and *cautious* fails to outperform *nbc* for any metric.

Ideally, one predictor or general predictor strategy, such as omitting the attribute with the lowest confidence, would prove itself preeminent and dominate over NB. However, no single predictor dominated or proved acceptable across all datasets in our experiments, although most datasets had at least one acceptable alternative to *nbc*. A subset of results selected with a bias towards the hybrid predictors appears in Table 1. Datasets with more than 2 classes exhibited less potential in terms of improving prediction using predictors that incorporate CI. In our evaluation, the *cautious* and *contrarian* predictors performed better overall than other hybrid predictors, and as a rule, strategies based on ignoring attributes with poor confidence delivered weaker results. This seems appropriate for rare attribute values which can actually be strong predictors of class.

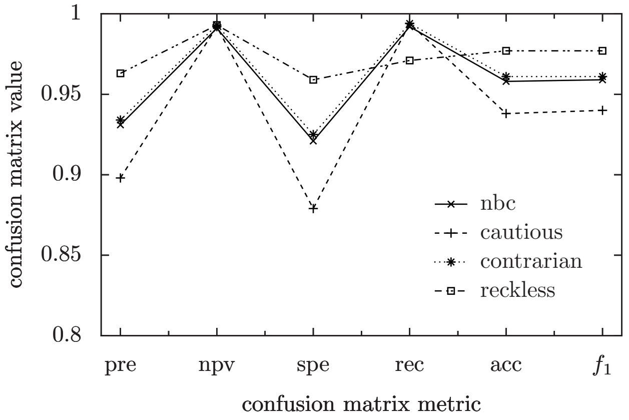

Figure 1: mushroom (edible)

	nbc	**reckless**	**cautious**	**contra**
colic	0.000	−0.022	+0.011*	−0.025
mushroom	0.000	+0.019*	−0.020	+0.002*
sonar	0.000	+0.000	−0.095	−0.035
voting	0.000	+0.010*	+0.007	+0.000

Table 1: Predictor Acceptance and Dominance*

Our goal is to continue to investigate interactions between attribute log odds ratios and attribute confidence in order to develop heuristics based on dataset attributes or other operating conditions that allow for reliable predictor selection. Once these interactions are better understood, such techniques may be applicable to boosting or ensemble learning.

References

Asuncion, A., and Newman, D. 2007. UCI machine learning repository.

Možina, M.; Demšar, J.; Kattan, M.; and Zupan, B. 2004. Nomograms for visualization of naive bayesian classifier. In *PKDD '04: Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 337–348. New York, NY, USA: Springer-Verlag New York, Inc.

Russell, S. J., and Norvig, P. 2002. *Artificial Intelligence: A Modern Approach (2nd Edition)*. Prentice Hall.

Witten, I. H., and Frank, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann.

Zhang, H. 2004. The optimality of naive bayes. In Barr, V.; Markov, Z.; Barr, V.; and Markov, Z., eds., *FLAIRS Conference*. AAAI Press.

Zupan, B., and Demšar, J. 2004. From experimental machine learning to interactive data mining. White paper, Faculty of Computer and Information Science, University of Ljubljana.