Bayesian Methods for the Evaluation of Classifiers

Dragos D. Margineantu
The Boeing Company
Mathematics & Computing Technology
Adaptive Systems
P.O. Box 3707, M/S 7L-66
Seattle, WA 98124-2207
dragos.d.margineantu@boeing.com

Roman D. Fresnedo
The Boeing Company
Mathematics & Computing Technology
Applied Statistics
P. O. Box 3707, M/S 7L-22
Seattle, WA 98124-2207
roman.d.fresnedo@boeing.com

1 Introductory Overview

This paper presents a Bayesian approach to estimating the risk (or the expected loss) of classifiers, and discusses some experimental results and the issues that have to be considered when assessing the risk of classifiers. The development of the proposed methodology was motivated by the shortcomings observed in employing the bootstrap tests of Margineantu and Dietterich [10] especially when applied on classifiers that make only a small number of errors, but whose misclassifications are associated with high-risk decisions (small probabilities and large misclassification costs).

2 Extended Abstract

Developing and selecting statistical tests for assessing the expected accuracy, loss, or risk of classifiers, or for comparing classification decisions of two classifiers has proven to be a task as difficult as (or even more difficult than) creating new learning algorithms [4].

Accurate statistical testing of classification decisions becomes even more complicated in the case of domains with *skewed class distributions*, or tasks with *little data available* for testing, but also in the case of *classifiers that tend to make a small number of mistakes* (especially good classifiers that rarely misclassify high risk instances). Indeed, the evaluation of classifiers can be viewed as a statistical inference task on the confusion matrix, and anything that causes small. The z tests based on the normal distribution and other standard statistical tests are even more inaccurate if arbitrary costs are associated with the classification decisions and the goal is to estimate the total risk (or the expected cost) of a classifier, or the difference in risk between classifiers. To address this issue, Margineantu and Dietterich [10] have proposed two new sets of tests for the cost-sensitive evaluation of classifiers (BCOST and BDELTA-COST), based on the *bootstrap* [5]. To correct for small values in the confusion matrix, the proposed bootstrap tests employ a uniform Dirichlet prior (or Laplace correction), λ.

Experimental analyses of the normal distribution based tests have shown that they compute confidence intervals that are too wide and that the two bootstrap tests always compute more accurate, much narrower confidence intervals, for $\alpha \in [0.01, 0.1]$. In the meantime both bootstrap tests have also been shown to be sensitive to the choice of the value of λ, and finding its optimal value is still an open question.

Small counts (or zero counts) are notoriously difficult [7, 6, 8, 9] and no statistical magic will solve the problem in the absence of any knowledge on how the values were generated. But, because in the case of small counts in confusion matrices generated by classifiers, typically the user may be aware of different characteristics of the problem (e.g., class 1 is rare, instances form classes 2 and 4 are always very far apart, etc.) and of the classifier (e.g., the knowledge that a classifier chooses the decision boundaries by balancing the errors) – all probabilities (of the confusion matrix) are clearly dependent.

Therefore, in order to reduce the sensitivity of the tests to the choice of uniform Laplace priors, we explored the possibility of estimating p_{ij} - the probability that an instance from class j is classified into class i, by using Bayesian methods. As mentioned above, we would like the priors should take into consideration both, characteristics of the problem and of the learning algorithms. We were especially interested in improving the quality of the estimates of p_{ij} for the cases where the cell counts are small (i.e., rare errors, and especially costly rare errors) and in computing more accurate estimates for the posterior distribution of risk. For a Bayesian approach, we expect that the estimates may "borrow strength" form each other and help to overcome the arbitrariness of the uniform Laplace correction used by the bootstrap tests.

We employed several structural models for estimating p_{ij} and different priors for their parameters. For the likelihood we used the multinomial, but, as in any real-world application problem, any available information should be used to define each of them. The models described below have been choosen to be not very complicated or detailed, be-

cause our main goal is to present the reader with the general framework of the Bayesian approach such that he or she will be able to construct and apply the appropriate model for a given task. The goal is not to spend a great effort on modeling, but just enough to compute accurate estimates of the risk (if one would attempt to have a detailed model $p_{ij}(x)$, then he could just forget about the learning machine and create a Bayesian model for the original problem thus providing automatically all the answers...).

The first model that we employed is a saturated log-linear model:

$$log(p_{ij}) = \mu + \mu_i^1 + \mu_j^T + \mu_{ij}^{1,T} \quad (1)$$

where $\mu_\cdot^1$ model the row (predicted class) effects, $\mu_\cdot^T$ model column (true class) effects, $mu_{\cdot,\cdot}^{1,T}$ the interactions between rows and columns. and μ is a scaling parameter (T stands for the true/actual class).

Next we employed a simpler direct model, that models separately the diagonal values (the probability of correct classification) and the off-diagonal values:

$$p_{jj} = d_j * \pi_j \quad (2)$$
$$p_{ij} = (1 - d_j) * \pi_i * \pi_j / (1 - \pi_j) \;\; i \neq j \quad (3)$$

This model could be appropriate for the case in which the user has some estimates of the true class proportions π_j, and some estimate of the maximum error (0/1 loss), $1 - d_j$.

If the class variable is ordered and errors are more likely to be made between adjacent classes (or their likelihood decreases with the distance between the classes), then the following model may be employed:

$$p_{jj} \propto d_j * \pi_j \quad (4)$$
$$p_{ij} \propto \pi_j * (1 - d_j) * b_m, i \neq j, m = i - j; b_m \downarrow \quad (5)$$

.

To assess our proposed approach and the different models presented above, we have run validation tests on two synthetic domains. This (using synthetic tasks) was done in order to be able to generate a large number of instances and to have a good estimate of the "true" probabilities (we approximated the true p_{ij} by the estimate $\hat{p}_{ij}$ on a million test instances). For all the models we have run MCMC to compute the posterior distribution of p_{ij} and the distribution of risk, when a (mis)classification loss matrix was available.

First we looked at the probability estimates and assessed their deviation from the truth ($\sum |\hat{p}_{i,j} - p_{i,j}|$). The saturated model has computed the best estimates for the cells corresponding to small probability values, whereas the other models exhibited large errors on those cells (which may result in distorted estimates especially in the case of tasks with very skewed class distributions or in the case of very accurate classifiers). The most overestimated small probabilities were obtained when we used the diagonal model (2,3). When we assessed the computed risk distribution, again, the saturated model was the most accurate, although all models showed a tendency to underestimate the risk.

For the problem of classifier evaluation, Bayesian methods may be judged to be an awkward step back from a purely data driven philosophy. Leo Breiman's "two cultures" paper [1] discusses this issue in detail. Given the complexity of the task of evaluating classifiers and the fact that users and testers of machine learning-based algorithms usually have prior knowledge on some characteristics of the task and/or general behavior of the classifier, we consider, and our experiments show it, that employing the proposed probabilistic models represents a meaningful approach.

References

[1] L. Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231, 2001.

[2] R. Caruana and A. Niculescu. An empirical comparison of supervised learning algorithms. In *Proceedings of ICML-2006*, pages 161–168, 2006.

[3] B. Cestnik. Estimating probabilities: A crucial task in machine learning. In L. C. Aiello, editor, *Proceedings of the Ninth European Conference on Artificial Intelligence*, pages 147–149, London, 1990. Pitman Publishing.

[4] T. G. Dietterich. Statistical tests for comparing supervised classification. *Neural Computation*, 10:1895–1924, 1998.

[5] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, New York, 1993.

[6] S. E. Fienberg and P. W. Holland. Methods for eliminating zero counts in contingency tables. In G. P. Patil, editor, *Random Counts in Scientific Work: Volume 1*. The Pennsylvania State University Press, 1968.

[7] I. J. Good. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. M.I.T. Press, Cambridge, Mass., 1965.

[8] I. J. Good. *Good thinking: The foundations of probability and its applications*. University of Minnesota Press, Minneapolis, MN, 1983.

[9] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[10] D. D. Margineantu and T. G. Dietterich. Bootstrap methods for the cost-sensitive evaluation of classifiers. In *Machine Learning: Proceedings of the Seventeenth International Conference*, pages 583–590, San Francisco, CA, 2000. Morgan Kaufmann.