Active Cost-Sensitive Learning

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

Abstract

For many classification tasks a large number of instances available for training are unlabeled and the cost associated with the labeling process varies over the input space. Meanwhile, virtually all these problems require classifiers that minimize a non-uniform loss function associated with the classification decisions (rather than the accuracy or number of errors). For example, to train pattern classification models for a network intrusion detection task, experts need to analyze network events and assign them labels. This can be a very costly procedure if the instances to be labeled are selected at random. In the meantime, the loss associated with mislabeling an intrusion is much higher than the loss associated with the opposite error (i.e., labeling a legal event as being an intrusion).

As a result, to address these types of tasks, practitioners need tools that minimize the total cost computed as a sum of the cost of labeling and the loss associated with the decisions.

This paper describes an approach for addressing this problem.

1 Introduction

A number of applications require learning algorithms that are capable (1) to learn from both labeled and unlabeled examples, (2) to minimize the cost non-uniform associated with the labeling efforts, and (3) to minimize the misclassification loss.

The active learning framework [Cohn *et al.*, 1994] relies on algorithms that select unlabeled instances and provide them to the expert for labeling and learn a classifier based on the labeled data. In most of the cases the data is labeled incrementally and the classification model tries to minimize the misclassification error rate. To our knowledge, virtually all research efforts in active learning have assumed both a uniform loss function (for which the goal was to reduce the *number* of misclassifications) and a uniform labeling cost function (for which the goal was to reduce the *number* of instances that have to be labeled to achieve a certain accuracy).

Meanwhile, research in cost-sensitive learning [Elkan, 2001] has focused mostly on the problem of minimizing the expected loss of the learned models, and has assumed that all (training) instances and their labels were readily available prior to the training phase.

Our approach in addressing the active cost-sensitive learning problem is based on learning algorithms that construct models for conditional density (or class probability) estimation $P(y|x)$, rather than class label predictors. The probability estimates provide an easy means for factoring in the misclassification losses in the classification decision making step. To address the labeling cost problem, we employ the same density estimation techniques over the available unlabeled data. The two methods are then combined into an algorithm that minimizes the combined cost.

Next section presents the general framework for our proposed techniques and discusses the challenges in implementing and applying these algorithms as well as some preliminary experimental results.

2 Active Learning for Minimizing the Combined Costs for Labeling and Decisions

Most learning algorithms can be transformed into class probability estimators that compute estimates $P(y|x)$ for the probabilities that an instance x is in class y ($y \in \{1 \ldots K\}$, where K is the number of classes). We employ these estimates in our algorithm both for minimizing the labeling costs and the misclassification decisions. We make the assumption that the loss function associated with the decisions is represented as a static K-by-K loss matrix L available at learning time. The contents of $L(i,j)$ specify the cost incurred when an example is predicted to be in class i when in fact it belongs to class j.

The pseudo code for our approach, ACTIVE-CSL, is presented in Table 1.

The algorithm trains a base learner to compute the class probability estimates over the unlabeled data (line 2). Then, both the sampling step (line 3) and the decision making step (the hypothesis h) are based on those estimates. The unlabeled instance for which the next label is requested from the expert is selected in line 3. The selection rule chooses the instance that, if labeled provides the most expected gain in terms of the total cost (labeling + decision loss) on the labeled instances (or on a hold-out labeled validation set). It is important to note that iff C and L functions have a shape such that instances that are more expensive to label reduce the

Table 1: Pseudo code for the ACTIVE-CSL active cost-sensitive learning algorithm.

```
Input: a set $S_l$, of $m$ labeled examples:
       $S_l =< (x_i, y_i), i = 1, 2, \ldots, m >$,
       a set $S_u$ of $v$ unlabeled examples:
       $S_u =< (\xi_i), i = 1, 2, \ldots, v >$,
       $L$ (a loss matrix), $C$ (a labeling cost function)
       a stopping criterion

[1] repeat
[2]     for each $j$ learn $P(y|\xi_j)$ using $S_l$ as training data;
[3]     $l = \operatorname*{argmin}_{\xi \in S_u}(C(\xi) +$
        $+ \sum_{k=1}^{K} P(k|\xi) \sum_{i} \sum_{j=1}^{K} P_{(\xi,k)}(j|x_i) L(y_i, j));$
[4]     $\psi_l$ = requested label for $\xi_l$;
[5]     remove $\xi_l$ from $S_u$, add $(\xi_l, \psi_l)$ to $S_l$;
[6] while stopping criterion not met

Output: $h(\mathbf{x}) = \operatorname*{argmin}_{y \in Y} \sum_{j=1}^{K} P(j|\mathbf{x}) L(y, j)$
// the optimal classification with respect to $L$ and $P$
```

misclassification cost more than instances that are less expensive to label - the selection rule will trade the loss associated with misclassifications against examples to be labeled. The procedure selecting for the next instance to be labeled, as employed by ACTIVE-CSL, and based on computing the maximum expected gain upon labeling (the second sum in line 3), is somewhat similar in nature to the querying rule of uncertainty sampling [Lewis and Catlett, 1994].

One important detail that needs to be addressed when applying this algorithm in practice is the choice for the algorithm used for estimating the probabilities (line 3 in Table 1), given that ACTIVE-CSL and its predictions rely on these estimates. In general, the quality of learned density estimates is dependent on the amount and the distribution of the labeled data that is available and on the hypothesis constructed by the base learning algorithm. Several research studies have addressed the problem of learning good probability estimates and calibrating classification scores and ranks into accurate probabilities [Zadrozny and Elkan, 2001].

We have implemented and tested ACTIVE-CSL by using bagged probability estimation trees [Provost and Domingos, 2003] as class probability estimators (in line 2 of the code). We have compared two procedures for computing probability estimates out of the bagged trees: (a) by averaging the probabilities computed by the individual trees and (b) by estimating the confidence for the individual probabilities based on the distribution of the estimates of the individual trees (as suggested by [Margineantu, 2002]). For assessing the quality of the misclassification decisions we have employed active learning curves (showing the gain in terms of cost compared to a random selection of instances for labeling) and BDELTA-COST [Margineantu and Dietterich, 2000] - a paired test for cost-sensitive classification decisions. The loss matrices for our experiments were generated based on some generic loss models, and the labeling cost function mapped instances that were closer to the decision boundary to higher labeling costs (the actual function we employed was a bell-shaped function with a maximum on the decision boundary).

We tested the two implementations of ACTIVE-CSL on five data sets from the UC Irvine Repository [Blake and Merz, 1998] (Breast cancer Wisconsin, Horse colic, King-rook vs. king-pawn, Liver disease, and Sonar) and on the binary version of the KDD Cup 1998 donations data.

The preliminary results show a minor advantage of the implementation employing confidence-based estimates over the implementation using averaged probability estimates. We have also run tests by employing random forests with $\log m$ attributes tested in a node [Breiman, 2001] as the class probability estimator and the results show no significant difference between the classification decisions based on random forests and the decisions based on bagged probability estimation trees.

References

[Blake and Merz, 1998] C. L. Blake and C. J. Merz. UCI Repository of ML databases, 1998. [www.ics.uci.edu/~mlearn/MLRepository.html].

[Breiman, 2001] L. Breiman. Random forests. Technical report, Dept. of Statistics, University of California, Berkeley, 2001.

[Cohn *et al.*, 1994] D. A. Cohn, L. Atlas, and R. Ladner. Improved generalization with active learning. *Machine Learning*, 15:201–221, 1994.

[Elkan, 2001] C. Elkan. The foundations of cost-sensitive learning. In *Proc. of the Seventeenth Intrnl. Joint Conference on Artificial Intelligence*. Morgan Kaufmann, 2001.

[Greiner *et al.*, 2002] R. Greiner, A. J. Grove, and D. Roth. Learning cost-sensitive active classifiers. *Artificial Intelligence*. 139:2, pages 137–174, 2002.

[Hettich and Bay, 1999] S. Hettich and S. D. Bay. The UCI KDD archive, 1999. [http://kdd.ics.uci.edu/].

[Lewis and Catlett, 1994] D. Lewis and J. Catlett. Heterogeneous uncertainty sampling. In *Proc. Eleventh Intrnl. Conference on Machine Learning*, pages 148–156. Morgan Kaufmann, 1994.

[Margineantu and Dietterich, 2000] D. D. Margineantu and T. G. Dietterich. Bootstrap methods for the cost-sensitive evaluation of classifiers. In *Proc. of the Seventeenth Intrnl. Conference on Machine Learning*, pages 583–590. Morgan Kaufmann, 2000.

[Margineantu, 2002] D. D. Margineantu. Class probability estimation and cost-sensitive classification decisions. In *Proc. ECML-2002, 13th European Conference on Machine Learning, Proceedings*, pages 270–281. LNAI 2430, Springer Verlag, 2002.

[Provost and Domingos, 2003] F. Provost and P. Domingos. Tree induction for probability-based rankings. *Machine Learning*, 52:3, 2003.

[Saar-Tsechansky and Provost, 2001] M. Saar-Tsechansky and F. Provost. Active learning for class probability estimation and ranking. In *Proc. Seventeenth Intrnl. Joint Conf. on Artificial Intelligence*, pages 911–917. AAAI Press/MIT Press, 2001.

[Zadrozny and Elkan, 2001] B. Zadrozny and C. Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Proc. of the Eighteenth Intrnl. Conference on Machine Learning*, pages 609–616. Morgan Kaufmann, 2001.