Learning More with Less*

Jason Rennie
jrennie@ai.mit.edu

July 17, 2003

Abstract

We investigate the problem of classification using a small number of training examples. We assume that we have access to a "reference" classification task (and corresponding training examples) that are similar, but not identical, to the main task. In this paper, we consider the case that the classification problem is similar enough that it is useful to directly incorporate examples from the reference task. We find that by weighting the reference examples appropriately, they provide regularization for the main task and drastically lower classification error. On a newsgroup classification task, using training examples from both the main and reference tasks gives error one-fourth that of using either set of examples individually.

1 Introduction

When we have only a few training examples to solve a specific learning task, it is often beneficial to try to exploit data from another learning task (what we call the "reference" task) that may be available. A number of papers have proposed ways of learning from other tasks or solving multiple tasks jointly. These include using reference tasks to intialize weights in a neural network[1], re-using classifiers trained for other tasks in information filtering [2], demonstrating substantial empirical gains from solving multiple related tasks jointly[3], formalizing the problem of learning internal representations and characterizing the accompanying sample complexity gains[4], or incorporating estimated invariances[5].

A new technique that we explore is using data from the reference data to regularize the classification problem. When there are few training examples

*Joint work with Tommi Jaakkola

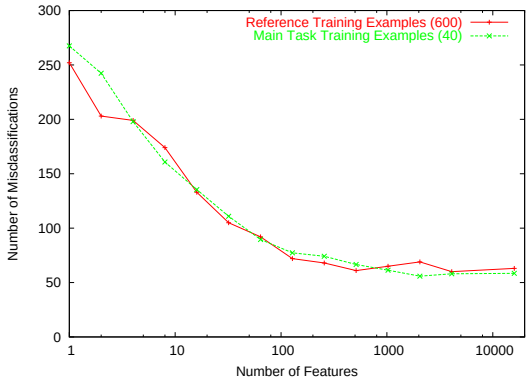

Figure 1: Classification performance on 400/class main task test examples. Performance is shown for (1) a classifier trained with 900/class reference task examples, and (2) a classifier trained with 40/class main task examples. Performance is similar. Reference task examples provide a limited amount of information about the main task.

for a classification task, learned decision boundaries will tend to be of high variance—different sets of examples will yield vastly different boundaries. This is especially true in text classification since the number of features will vastly outnumber the training examples. Examples from another, similar, task can be used to regularized the task and reduce the variability in the learned decision boundary. Since the reference task is used as a regularizer to reduce variance, it is not necessary that the examples be perfectly aligned with the main task. For example, if our main task is to differentiate baseball newsgroup postings from postings about guns, we can successfully utilitize postings on hockey and U.S. politics to regularize the problem, even though they are somewhat tangential to the main task. Next we describe how we use the examples from the reference task to regularize the main task.

2 Using the Reference Examples

Consider the newsgroup posting example. Our main task is to distinguish postings on baseball from postings on guns; we are given a small set of labeled examples. We have many examples for our reference task, distinguishing between hockey and U.S. politics postings. The reference task is close enough to give some performance on the main task. Figure 1 shows this. A classifier built with 900 reference task examples/class achieves about the same rate of error on the main task as one built using 40 main task ex-

amples/class. So, if we have fewer than 40 main task examples/class, one option would be to use the reference task classifier. But, we can do better. The reference task classifier takes no advantage of the information in the main task examples. And, the main task classifier is not regularized with the reference task examples.

We can do better by using both sets of training examples to construct a classifier. But, if we simply train a classifier using 900/class reference examples and 40/class main task examples, we will not do much better than using reference examples alone. This is because the classifier will learn the reference task and for the large part ignore the main task examples (since there are so few of them). We want to focus on the main task, but use the reference examples for regularization. We achieve this by down-weighting the reference examples. For regularized logistic regression, the usual optimization is

$$\min_{w_1,\ldots,w_d} -\sum_{i=1}^{n} g(y_i z_i) + \frac{C}{2}\sum_{j=1}^{d} w_j^2, \tag{1}$$

where $g(x) = \frac{1}{1+e^{-x}}$, $z_i = \sum_j x_{ij} w_j$ and C is the regularization parameter. The y_i are the binary labels and the $x_i = (x_{i1}, \ldots, x_{id})$ are the training examples. To use the reference examples to regularize the main task examples, we use the following optimization,

$$\min_{w_1,\ldots,w_d} -\sum_{i\in\mathcal{M}} g(y_i z_i) - \frac{s}{|\mathcal{R}|}\sum_{i\in\mathcal{R}} g(y_i z_i) + \frac{C}{2}\sum_{j=1}^{d} w_j^2, \tag{2}$$

where $\mathcal{M}$ holds the indices of the main task examples and $\mathcal{R}$ holds the indices of the reference task examples. The constant s is the equivalent size of the reference examples. When s is small, the reference examples serve more as regularization than as the primary focus of the classifier. The effect of combining the two sets of examples and weighting them appropriately can be seen in figure 2. Using 10/class main task examples and an equivalent size $s = 10$, error is much lower than using main task or reference examples individually.

3 Summary

We have discussed the notion of transfer, using knowledge from one supervised learning problem to aid another problem. We achieve transfer by

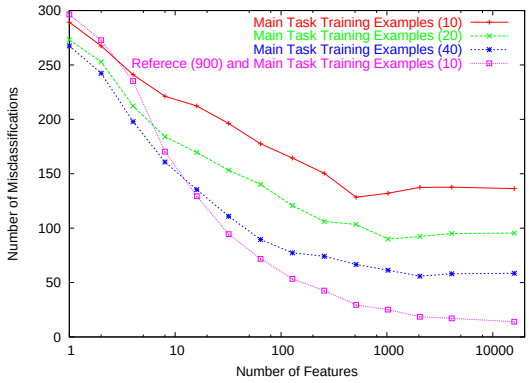

Figure 2: Classification performance on 400/class main task test examples. Performance is shown for classifiers trained with 10/class, 20/class, and 40/class main task training examples. This is compared against performance for a classifier trained with 10/class main task examples and 900/class reference task examples, where the equivalent size of the reference examples is set to $s = 10$. Utilizing both sets of examples greatly reduces error.

learning a classifiers that utilizes both main task and reference task training examples. So that the classifier focses on the main task and merely uses the additional examples for regularization.

References

[1] Lorien Y. Pratt. Discriminability-based transfer between neural networks. In Stephen José Hanson, Jack D. Cowan, and C. Lee Giles, editors, *Advances in Neural Information Processing Systems*, volume 5, pages 204–211. Morgan Kaufmann, San Mateo, CA, 1993.

[2] William W. Cohen and Daniel Kudenko. Transferring and retraining learned information filters. In *Proceedings of AAAI-97*, 1997.

[3] Rich Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997.

[4] Jonathan Baxter. Learning internal representations. In *Proceedings of the Eighth Annual Conference on Computational Learning Theory*, pages 311–320, 1995.

[5] Sebastian Thrun. Is learning the n-th thing any easier than learning the first? In *Neural Information Processing Systems*, volume 8, 1996.