Active Learning with Committees

Ray Liere
lierer@research.cs.orst.edu

Prasad Tadepalli
tadepall@research.cs.orst.edu

Department of Computer Science, Oregon State University,
Dearborn Hall 303, Corvallis, OR 97331-3202, USA

In many real-world domains like text categorization, supervised learning requires a large number of training examples. In our research, we are using *active learning with committees* methods to reduce the number of training examples required for learning. Disagreement among the committee members on the predicted label for the input part of each example is used to determine the need for knowing the actual value of the label. Our experiments in text categorization using this approach demonstrate a 1-2 orders of magnitude reduction in the number of labeled training examples required.

We present here a summary of our research. Please see the full paper *Active Learning with Committees for Text Categorization* in this Proceedings for additional details and for the references.

The goal of *text categorization* is to assign each document to the appropriate categories, based on the semantic content of the document. Our goal is to develop automatic methods for text categorization through the application of machine learning techniques. The text categorization domain has several characteristics that make it a difficult domain for the use of machine learning, including a very large number of input features (10,000+), high levels of attribute and class noise, and a large percentage of features that are irrelevant. As a result, the use of supervised learning requires a relatively large number of labeled examples. We have been developing methods that will dramatically reduce the number of labeled examples needed in order to train the system, without incurring unacceptable decreases in prediction accuracy.

One approach to reducing the number of labeled examples needed, called *active learning*, allows the learning program to exert some control over the examples on which it learns [Cohn94]. *Query by Committee* (QBC) is one specific type of active learning. It starts with a committee of all possible hypotheses. Each feature vector is presented to the committee. A high degree of disagreement among the hypotheses as to the predicted value of the label indicates that the example will be very informative, and so the actual label is requested. The label is then used to remove all hypotheses from the committee that do not predict the actual label. A major advantage of QBC is that the number of examples required is logarithmic in the number of examples required for random example selection learning. See [Freund92, Seung92, Freund95].

Here is an intuitive explanation for why QBC requires far fewer examples. Initially, assuming that the hypotheses in the committee are sufficiently diverse, two randomly chosen hypotheses disagree on an example with a significantly high probability. Labels are thus used for a significant fraction of the examples. As the learning progresses, each hypothesis approaches the optimal target hypothesis, and so the diversity between the different hypotheses decreases. Therefore the informativeness of an example as measured by the probability of disagreement between two randomly chosen hypotheses decreases, and so the distance between two successive label requests increases.

Our learning methods are similar to QBC, in that they use disagreement among the committee members as to the value of the predicted label to determine the need for requesting the actual value of that example's label from the teacher. Unlike QBC, our committee consists of a small finite number of hypotheses, which are updated during learning.

We use Winnow as the learning algorithm for each committee member since it is especially suited to large attribute spaces and to situations in which there is a large percentage of irrelevant features [Littlestone88].

We use majority voting to combine the predictions of the committee members into a prediction of the committee as a whole.

Our experiments were conducted using titles from the Reuters-22173 corpus [Reuters], which contains 22,173 documents. Our results indicate that active learning with committees can, as compared to supervised learning with a single learner, result in learning methods that use only 2.9% as many labeled examples but still achieve the same accuracy. In addition, this method gives better accuracy *during* learning than the other systems examined, and thus is a good choice in situations where one has either a limited number of training examples or a limited amount of time in which to learn.

This research was partially supported by the National Science Foundation under grant number IRI-9520243.