Applying Supervised Learning to Real-World Problems

Dragos D. Margineantu
Oregon State University
Department of Computer Science
303 Dearborn Hall
Corvallis, Oregon 97331-3202
margindr@cs.orst.edu

The last years have seen machine learning methods applied to an increasing variety of application problems such as: language, handwriting and speech processing, document classification, knowledge discovery in databases, industrial process control and diagnosis, fraud and intrusion detection, image analysis and many others.

Our work starts from the realization that most of these problems require significant reformulation before learning algorithms can be applied, and in many cases, existing algorithms require modifications before being applied to a problem.

The problems mentioned above differ in many aspects but, if subdivided into smaller problems (and this is the approach commonly taken), the sub-problems can often be formulated and approached by employing similar, unified supervised learning techniques. However, this divide-and-conquer process creates dependencies in the data that violate the assumption that the data are independent and identically distributed (iid) and that all errors are of equal cost, issues that involve making changes to existing learning algorithms.

The purpose of my thesis is to identify some procedural steps that are shared among the learning approaches to complex real-world problems, and to develop robust general purpose techniques and tools to replace ad-hoc decisions that are currently made during an application effort. The main topics my thesis will emphasize on are described as follows.

Learning with misclassification costs. Most classification algorithms assume uniform class distribution and try to minimize the misclassification error. However, many applications require classifiers that minimize an asymmetric loss function rather than the raw misclassification rate. In general, the cost of a wrong prediction depends both on the actual class and on the predicted class. One way to incorporate loss information into classifiers is to alter the priors based on labels of the training examples. For 2-class problems, this can be easily accomplished for any loss matrix, but for $k > 2$ classes, it is not sufficient. We have studied methods for setting the priors to best approximate an arbitrary $k \times k$ loss matrix in decision tree learners. We are currently studying alternative methods to incorporate loss information into decision trees, neural networks and other supervised learning algorithms.

Incorporating prior and common sense knowledge and learning with non-independent data examples. Generally, besides the data available, there is simple, common sense knowledge about the data that is not made explicit in the dataset or database. One question is whether this knowledge is useful, and, if so, how it can be represented and communicated to a learning tool. Another question is what independence assumptions are made by the learning algorithm. The standard learning models and algorithms assume that the data is drawn independently and that it is identically distributed. How is it possible to apply a learning method when this assumption is violated? One result that we have in this direction is that, if the data has a significant number of non-independent examples, ensemble learning methods like AdaBoost and Bagging do not improve performance (as expected) over single learners.

Determination of the granularity of the examples. As described in the previous section, the granularity of the examples is a common problem for supervised learning approaches to different tasks. We believe that there are two distinct granularity problems to be solved for each general learning task: the granularity of the features (e.g. how large is the sliding window in a text-to-speech learning task?) and the granularity of the target labels (what do we want to predict, e.g. phonemes, stresses, or both?). Our work on this topic includes only some experiments on a manufacturing dataset.

Examples with multiple labels. The vast majority of the research in supervised learning assumes that the labels of the examples are single values. However, there are cases when each example belongs to multiple classes. The training examples may be labeled either with single values or with a set of values. Possible approaches to these kind of problems include: using learning algorithms to output multiple values by exploiting the probability distributions, and using neural networks and to change the error function that is propagated back (as implemented in Apple's *Newton* character recognizer).