1 Research Activities

Active Class Selection: In collaboration with Carla Brodley. *Active Class Selection (ACS)* addresses the question: if one can collect n additional training instances, how should they be distributed with respect to class? This scenario occurs for applications in which the process of generating data cannot be separated from the process of obtaining labels. If we assume that generating training data is costly, the goal of ACS is to optimize data generation to minimize the number of instances required for optimal classification performance. We recognized this new class of supervised learning problems, and have created several methods of performing ACS to ensure the training set is sufficient for the classification task.

We recognized this new class of supervised learning problems when working with chemists in the Walt Laboratory at Tufts University to train an artificial nose to discriminate vapors. In this domain, creating more data requires the chemist to conduct an experiment where a particular vapor is passed over the sensor (the nose). The nose is a general purpose device that can be trained to discriminate the k vapors of interest, such as dangerous chemicals, for the location in which it is to be placed. For example, a nose could be installed in a subway and sound an alarm if it detects a harmful gas. The accuracy is a function of how well the beads are able differentiate among the vapors of interest.

Additionally, ACS provides an effective framework for domain experts and machine learning experts to collaborate. It allows an iterative process where domain experts can be given recommendations about a small amount of data to generate, pass the data to a machine learning researcher who can then further make recommendations. Including scientists in the iterative process allows the effective integration of machine learning in interdisciplinary research. A workshop paper and a conference paper in computer science have been published from this collaboration. Additionally, both the chemists and the computer scientists have submitted journal papers based on this research.

Active Learning: In collaboration with Carla Brodley. Active learning is an effective technique for classification problems with a set of labeled instances, a pool of unlabeled instances, and an expert willing to label a subset of the unlabeled pool [**?**]. This situation arises in many domains, from information retrieval to image classification. Although several techniques for active learning exist, assessing their relative performance online is an open question. Thus when applying active learning in a real world situation, we have no methods to help us determine which active learning method will work best for the problem at hand.

We introduced an entropy-based measure, *Average Pool Uncertainty*, for assessing the online progress of active learning and illustrated how APU can be applied to assess the performance of active learning methods. In particular, we apple APU to address a neglected problem in active learning – determining when active learning should terminate because additional labeling will not lead to a decrease in empirical error. Additionally, APU is applicable to selecting/combining active learning methods. Given m different active learning methods, we present a method for *combining* their labeling suggestions to best utilize labeling resources.

The motivating problem of this research is the labeling of the Earth's surface to create a land cover classifier. The pool of available sites to label is many times the size of the labeled training set needed to achieve the optimal accuracy. Thus, we must not only use active learning to efficiently choose examples for labeling, but also must determine when adding additional land cover sites would be a waste of resources. The land cover dataset consists of a time series of globally distributed satellite observations of the Earth's surface. ACS is applicable because geographers have knowledge of where on the Earth's surface one can expect to find a given class. The remote sensing observations are measurements of the normalized difference vegetation index [**?**]. The dataset has 36 aggregate statistics of the minimum, maximum, and mean values of the pixels and seventeen classes of land cover.

2 Instrumentation Need

Both Active Class Selection and Active Learning are CPU-intensive. They require working with large datasets. Additionally, experiments are conducted with several methods, each with a large range of parameters. Without the cluster, my research would be so time-consuming to be impractical.

References

[1] D. Angluin. Queries and concept learning. *ML*, 2(4):319–342, 1988.

[2] C. Brodley and M. Friedl. Identifying and eliminating mislabeled training instances. *JAIR*, 11:131–167, 1999.