

2007 Student Research Symposium

Language Technologies Institute



Dual Strategy Active Learning

Pinar Donmez

Abstract:

Generally, the research context of the work that I'd like to present is active learning. Active learning deals with small amounts of labeled data and a large pool of unlabeled data. It selectively samples the unlabeled data to achieve high performance with relatively small training data. In most real life applications, obtaining training data is time-consuming and costly, but gathering large amounts of unlabeled data is often cheap. Thus, there is a need to label only the most useful examples from this large pool of unlabeled data to reduce the time and effort of labeling by experts.

The goal of this work was to find an active learning method that works well throughout the entire sampling iterations. Some active learning methods perform best when very few instances have been sampled, and others after substantial sampling. For instance, density estimation-based active sampling methods perform well with minimal labeled data whereas uncertainty-based sampling methods perform well when a larger number of instances have already been sampled. The objective of our work was to combine these two solutions to obtain performance superior to both of the individual methods over the entire operating range.

Our work builds upon the work of Nguyen and Smeulders [ICML04]. They use a weighted uncertainty sampling strategy where the expected error (uncertainty) of the trained model on a data point is weighed by the density of that point. They use K-medoid clustering to cluster the data, and find cluster membership probabilities using EM, and train the labeled data via logistic regression. We realized that their approach reaches a relatively stable state after a certain number of iterations and cannot improve further even though standard uncertainty sampling outperforms it in that range. Hence, in our work (which we call DUAL) we combine the work of Nguyen and Smeulders and uncertainty sampling as follows: DUAL executes the method of Nguyen and Smeulders until it estimates a low derivative of expected error for that method. It indicates that the method has stabilized and it is time to switch to a better strategy. We proposed to use a mixture model of expected error (uncertainty) and density scores for each candidate point. The weight of the uncertainty score is proportional to one minus the expected error of uncertainty sampling so that the weight of the uncertainty score is increased as uncertainty sampling does better. As a result, the algorithm tends to select more uncertain points that are not necessarily highly dense points since towards the end of the active sampling iterations high density regions have already been sampled and the model needs to fine-tune the decision boundary by selecting the most uncertain data.

We used six widely used UCI datasets for evaluation. We start with 0.4% of the entire data as the initial labeled data, and used the rest as the unlabeled data. We ran each algorithm for 100 iterations and recorded the classification error at each iteration. The results are averaged over 4 runs. We tested DUAL against the method of Nguyen and Smeulders, uncertainty sampling, density-based sampling, representative sampling and the COMB method of [Baram, Y. et al, ICML03]. Representative sampling cluster the points within the margin of an SVM classifier and the centroid of the biggest cluster is chosen to be labeled. COMB selects among three alternative active sampling strategies to decide which strategy to use at each iteration. DUAL significantly outperforms all methods on 4 out of 6 datasets. For one dataset, it is outperformed by uncertainty sampling since it fails to switch early enough. For the other, it is outperformed by density-based sampling. As a failure analysis, we simulated a better switching point which makes DUAL the best performer. Also, we tested putting more weight on the density score instead of the uncertainty score for the later case and again observed that DUAL outperforms density-based sampling after the adjustment. It suggests DUAL is a more advantageous hybrid approach that leads to superior performance than individual methods, and hence can be effectively used in areas where active learning can offer substantial reduction in labeling effort; such as text categorization, semantic role labeling, protein structure prediction, etc.

Donmez P., Carbonell J. G., Bennett P.: Dual Strategy Active Learning, to appear in ECML07 (2007), Warsaw, Poland.