With the proliferation of social media, gathering data has become cheaper and easier than before. However, this data can not be used for machine learning without labels. Asking experts to annotate sufficient data for training is both expensive and time-consuming. Current techniques provide two solutions to reducing the cost and providing high quality labels: crowdsourcing and active learning. Crowdsourcing, which outsources tasks to a distributed group of people, can be used to provide a large quantity of labels but controlling the quality of labels is hard. Active learning, which requires experts to annotate a subset of the most informative or uncertain data, is very sensitive to the annotation errors. Though these two techniques can be used independently of one another, by using them in combination they can complement each other’s weakness. In this thesis, I investigate the development of active learning Support Vector Machines (SVMs) and expand this model to sequential data. Then I discuss the weakness of combining active learning and crowdsourcing, since the active learning is very sensitive to low quality annotations which are unavoidable for labels collected from crowdsourcing. The remainder of the thesis focuses on addressing this issue, I propose three possible strategies, incremental relabeling, importance-weighted label prediction and active Bayesian Networks. The incremental relabeling strategy requires workers to devote more annotations to uncertain samples, compared to majority voting which allocates different samples the same number of labels. Importance-weighted label prediction employs an ensemble of classifiers to guide the label requests from a pool of unlabeled training data. An active learning version of Bayesian Networks is used to model the difficulty of samples and the expertise of workers simultaneously to evaluate the relative weight of workers’ labels during the active learning process. All three strategies apply different techniques with the same expectation - identifying the optimal solution for applying an active learning model with mixed label quality to crowdsourced data. However, the active Bayesian Networks model, which is the core element of the proposed work, provides additional benefits by estimating the expertise of workers during the training phase. As an example application, I also demonstrate the utility of crowdsourcing for human activity recognition problems.

Major: Computer Science

Educational Career:
Bachelor’s of Automation, BS, 2006, University of Science and Technology of China
Master’s of Computer Science, MS, 2011, University of Central Florida

Committee in Charge:
Gita Sukthankar, Chair, EECS
Marshall Tappen, EECS
Michael Georgiopoulos, EECS
Rahul Sukthankar, Google Research

Approved for distribution by Gita Sukthankar, Committee Chair, on April 30, 2013.
The public is welcome to attend.