Optical Character Recognition Domain Expert Approximation Through Oracle Learning

by Joshua Menke

One of the challenges in using optical character recognition to automatically scan in information from family history documents is the varying levels of quality in the appearance of the documents. A single document can contain both high quality or ``clean'' images and low quality or ``noisy'' images. A common solution to learning in a noisy environment is to train a single classifier on a mixed data set of both clean and noisy data. Often, the resulting classifier performs worse on clean data than a classifier trained only on clean data, and likewise for a classifier trained only on noisy data. It would be preferable to use the two domain specific classifiers instead, but this requires knowing if a given sample is clean or noisy---an often difficult problem to solve. Here, oracle learning is used to approximate the two domain specific classifiers with a single oracle-trained model. The classifier trained only on noisy data and the classifier trained only on clean data are used together as an oracle to label those parts of the training set that correspond to each model's ``expertise.'' On a set of both noisy and clean optical character recognition data, using oracle learning to approximate the domain experts resulted in a statistically significant improvement ($p < 0.0001$) over standard training on the mixed data.