Learning from Weak Teachers

Shai Ben-David Phil Long Ohad Shamir Ruth Urner

This work was motivated by the following problem (introduced to us by Russ Greiner). There is considerable interest in the development of automated programs for diagnosis of brain tumors from CT scans of the skull. Machine learning is a natural approach for the development of such programs. However, machine learning tools require input of large labeled samples. In the context of classification of brain images, this means large amounts of images of patients classified according to whether, say, tumors appearing in the images are benign or malignant. One obstacle that arises is that it is very difficult, and expensive, to get such classifications by top human experts. An interesting alternative is to have medical students label the images instead of experienced experts. On top of the advantage of the larger number of students, there is also the cost consideration - they can be hired for the task for a much lower cost than expert physicians. However, there is yet another issue to address with such a solution; the labeling provided by students may be erroneous, especially so in images that are challenge to classify. A possible solution may be to use student's for the vast majority of training images, and refer to an expert only few of those images - those that are most challenging to classify (or most crucial for the design of the classification program).

Similar scenarios of utilizing weak teachers in the process of learning arise in many other practical domains. For example, when one wishes to train a spam detector, it may be difficult to get sufficiently many examples of emails labeled Spam/NotSpam, but one may view emails that are deleted without being opened as a proxy for a Spam classification,and use that more readily available data as a weak teacher for training a spam detector.

Many questions arise in this context: Can such a paradigm be used to generate quality classifiers? Can it save calls to an expert without compromising quality by too much? How should one decide which images to refer to an expert? How should an output classifier be computed from a mixture of expert-labeled and novice-labeled training data?

In this work we make a first step towards a theory to model and support such learning systems. We address the following learning setup: There are two sources of label information ("teachers"); a *strong teacher*, that labels according to the target distribution, and a *weak teacher* whose labelings may deviate from these labels.

Our goal is to learn a good label predictor from random examples that are labeled by these teachers. We assume that the learner can, for any given sample point, choose which teacher to query about its label. We wish to determine conditions, on the nature of the weak teacher and of the target distribution, that will allow utilizing the weakly-labeled examples to save queries to the strong teacher.

We suggest a formal model of what we call *weak teachers*, propose an algorithmic paradigm for deciding on which input sample points should the learner query a *strong teacher* (or an expert),and analyzes conditions under which the proposed paradigm is guaranteed to utilize weak-teacher labeling towards a saving in the required number of correctly-labeled training examples.