

Improving Data Mining via Noisy Micro-outsourcing

Victor S. Sheng, Foster Provost, Panagiotis G. Ipeirotis
New York University

This is based on a paper that appeared as: Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers. Sheng, S., F. Provost and P. Ipeirotis. ACM SIGKDD 2008. Best Paper Award Runner-up. In the winter we'd expect to have more results.

The KDD paper is available: <http://pages.stern.nyu.edu/~panos/publications/kdd2008.pdf>

This work focuses on problems where it is possible to obtain certain (noisy) data values (“labels”) relatively cheaply, from on-line micro outsourcing sources (“non-expert labelers”). A main focus is the strategy of outsourcing to obtain these values as training labels for supervised modeling. (This setting is in direct contrast to the setting motivating active learning and semi-supervised learning, where unlabeled points are relatively inexpensive, but labeling is expensive.) Our ability to perform non-expert labeling cheaply and easily is facilitated by on-line outsourcing systems such as Rent-A-Coder (<http://www.rentacoder.com>) and Amazon’s Mechanical Turk (<http://www.mturk.com>), which match workers with arbitrary (well-defined) tasks, as well as by creative labeling solutions like the ESP game (<http://www.espgame.org>).

These cheap labels may be noisy due to lack of expertise, dedication, interest, or other factors. In the face of noisy labeling, it is natural to consider repeated labeling: obtaining multiple labels for some or all data points. We explore whether, when, and for which data points one should obtain multiple, noisy training labels, as well as what to do with them once they have been obtained.

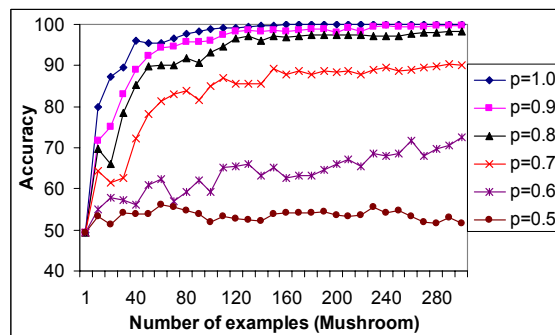


Figure 1. Learning curves under different quality levels of training data (p is the probability of a label being correct).

Figure 1 shows learning curves under different labeling qualities for the mushroom data set, specifically, for the different quality levels of the training data. The test set has perfect quality with zero noise. The figure shows learning curves relating the classification accuracy of a Weka J48 model to the number of training data. This data set is illustrative because with zero-noise labels one can achieve perfect classification after some training, as demonstrated by the $p=1.0$ curve.

Figure 1 illustrates that the performance of a learned model depends both on the quality of the training labels and on the number of training examples. Of course if the training labels are

uninformative ($p=0.5$), no amount of training data helps. As one would expect, under the same labeling quality, more training examples lead to better performance, and the higher the quality of the training data, the better the performance of the learned model. However, the relationship between the two factors is complex: the marginal increase in performance for a given change along each dimension is quite different for different combinations of values for both dimensions. To this complexity one must overlay the different costs of acquiring only new labels versus whole new examples, as well as the expected improvement in quality when acquiring multiple new labels.

Our work makes several contributions. First, under gradually weakening assumptions, we assess the impact of repeated-labeling on the quality of the resultant labels, as a function of the number and the individual qualities of the labelers. We derive analytically the conditions under which repeated-labeling will be more or less effective in improving resultant label quality. We then consider the effect of repeated-labeling on the accuracy of supervised modeling. As demonstrated in Figure 1, the relative advantage of increasing the quality of labeling, as compared to acquiring new data points, depends on the position on the learning curves. We show that even if we ignore the cost of obtaining the unlabeled part of a data point, there are times when repeated-labeling is preferable compared to getting labels for unlabeled examples. Furthermore, when we do consider the cost of obtaining the unlabeled portion, repeated-labeling can give considerable advantage.

We present a comprehensive experimental analysis of the relationships between quality, cost, and technique for repeated-labeling. The results show that even a straightforward, round-robin technique for repeated-labeling (*GRR*) can give substantial benefit over single-labeling. We then show that selectively choosing the examples to label repeatedly yields substantial extra benefit, showing in Figure 2. A key question is: How should we select data points for repeated-labeling? We present two techniques (*LU* and *MU*) based on two different types of uncertainty: label multiset uncertainty and model uncertainty, each of which improves over round-robin repeated labeling (*GRR*). Then we show that a technique (called *LMU*) that combines the two types of uncertainty is even better.

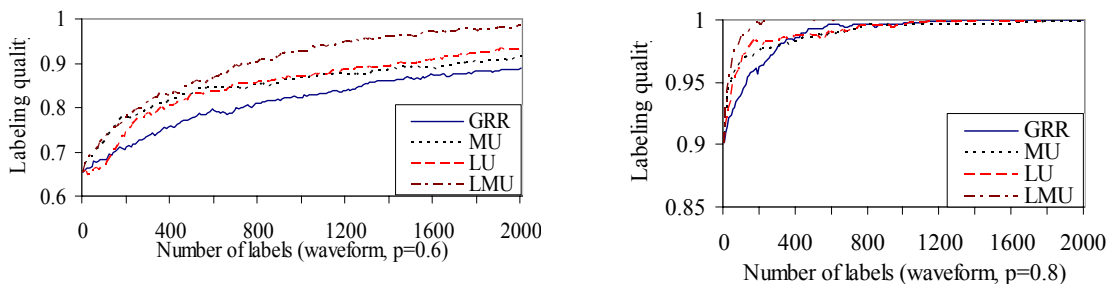


Figure 2. The data quality improvement of the four strategies (*GRR*, *LU*, *MU* and *LMU*) for the waveform dataset.

This research opens many new directions (some of them we are working on). One of the questions we are working on is to estimate the labeler's qualities. For most of the work we assumed that all the labelers have the same quality p and that we do not know p . In real applications, it is natural that different labelers have different labeling qualities. How can we estimate the quality of each labeler? Good estimates of individual labelers' qualities inferred by observing the assigned labels could allow more sophisticated selective repeated-labeling strategies.

In our analyses we also assumed that the difficulty of labeling an example is constant across examples. In reality, some examples are more difficult to label than others and building a selective repeated-labeling framework that explicitly acknowledges this, and directs resources to more difficult examples, is an important direction. This is another topic we are working on.

Of course, there are many other questions left that we will work on. In our repeated-labeling strategy we compared repeated-labeling vs. single labeling, and did not consider any hybrid scheme that can combine the two strategies. A promising direction for future research is to build an approach that decides dynamically which action will give the highest marginal accuracy benefit for the cost. Such an algorithm would compare on-the-fly the expected benefit of acquiring new examples versus selectively repeated-labeling existing, noisy examples and/or features.

Intuitively, we might also expect that labelers would exhibit higher quality in exchange for a higher payment. It would be interesting to observe empirically how individual labeler quality varies as we vary the payment. The labeling process is assumed a noisy process over a true label. An alternative, practically relevant setting is where the label assignment to a case is inherently uncertain. This is a separate setting where repeated-labeling could provide benefits.