# Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers

Victor S. Sheng
ssheng@stern.nyu.edu

Foster Provost
fprovost@stern.nyu.edu

Panagiotis G. Ipeirotis
panos@stern.nyu.edu

Department of Information, Operations, and Management Sciences
Leonard N. Stern School of Business, New York University

## ABSTRACT

This paper addresses the repeated acquisition of labels for data items when the labeling is imperfect. We examine the improvement (or lack thereof) in data quality via repeated labeling, and focus especially on the improvement of training labels for supervised induction. With the outsourcing of small tasks becoming easier, for example via Rent-A-Coder or Amazon's Mechanical Turk, it often is possible to obtain less-than-expert labeling at low cost. With low-cost labeling, preparing the unlabeled part of the data can become considerably more expensive than labeling. We present repeated-labeling strategies of increasing complexity, and show several main results. (i) Repeated-labeling can improve label quality and model quality, but not always. (ii) When labels are noisy, repeated labeling can be preferable to single labeling even in the traditional setting where labels are not particularly cheap. (iii) As soon as the cost of processing the unlabeled data is not free, even the simple strategy of labeling everything multiple times can give considerable advantage. (iv) Repeatedly labeling a carefully chosen set of points is generally preferable, and we present a robust technique that combines different notions of uncertainty to select data points for which quality should be improved. The bottom line: the results show clearly that when labeling is not perfect, selective acquisition of multiple labels is a strategy that data miners should have in their repertoire; for certain label-quality/cost regimes, the benefit is substantial.

## Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data mining; I.5.2 [**Design Methodology**]: Classifier design and evaluation

## General Terms

Algorithms, Design, Experimentation, Management, Measurement, Performance

## Keywords

data selection, data preprocessing

*KDD'08,* August 24–27, 2008, Las Vegas, Nevada, USA.

## 1. INTRODUCTION

There are various costs associated with the *preprocessing* stage of the KDD process, including costs of acquiring features, formulating data, cleaning data, obtaining expert labeling of data, and so on [31, 32]. For example, in order to build a model to recognize whether two products described on two web pages are the same, one must extract the product information from the pages, formulate features for comparing the two along relevant dimensions, and label product pairs as identical or not; this process involves costly manual intervention at several points. To build a model that recognizes whether an image contains an object of interest, one first needs to take pictures in appropriate contexts, sometimes at substantial cost.

This paper focuses on problems where it is possible to obtain certain (noisy) data values ("labels") relatively cheaply, from multiple sources ("labelers"). A main focus of this paper is the use of these values as training labels for supervised modeling.[1] For our two examples above, once we have constructed the unlabeled example, for relatively low cost one can obtain non-expert opinions on whether two products are the same or whether an image contains a person or a storefront or a building. These cheap labels may be noisy due to lack of expertise, dedication, interest, or other factors. Our ability to perform non-expert labeling cheaply and easily is facilitated by on-line outsourcing systems such as Rent-A-Coder[2] and Amazon's Mechanical Turk,[3] which match workers with arbitrary (well-defined) tasks, as well as by creative labeling solutions like the ESP game.[4]

In the face of noisy labeling, as the ratio increases between the cost of preprocessing a data point and the cost of labeling it, it is natural to consider *repeated labeling*: obtaining multiple labels for some or all data points. This paper explores 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 shows learning curves under different labeling qualities for the *mushroom* data set (see Section 4.1). Specifically, for the different quality levels of the *training* data,[5] the figure shows learning curves relating the classification accuracy of a Weka J48 model [34] 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 $q = 1.0$ curve.

Figure 1 illustrates that the performance of a learned model

[1] 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.

[2] http://www.rentacoder.com

[3] http://www.mturk.com

[4] http://www.espgame.org

[5] The test set has perfect quality with zero noise.

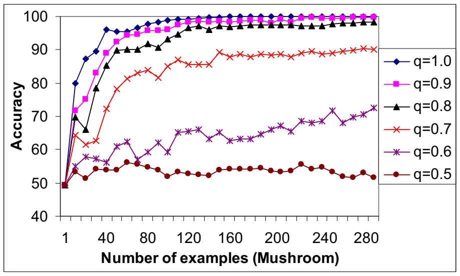


**Figure 1: Learning curves under different quality levels of training data ($q$ is the probability of a label being correct).**

depends both on the quality of the training labels and on the number of training examples. Of course if the training labels are uninformative ($q = 0.5$), no amount of training data helps. As expected, 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, 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.

This paper 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 can give substantial benefit over single-labeling. We then show that selectively choosing the examples to label repeatedly yields substantial extra benefit. A key question is: How should we select data points for repeated-labeling? We present two techniques based on different types of information, each of which improves over round-robin repeated labeling. Then we show that a technique that combines the two types of information is even better.

Although this paper covers a good deal of ground, there is much left to be done to understand how best to label using multiple, noisy labelers; so, the paper closes with a summary of the key limitations, and some suggestions for future work.

## 2. RELATED WORK

Repeatedly labeling the same data point is practiced in applications where labeling is not perfect (e.g., [27, 28]). We are not aware of a systematic assessment of the relationship between the resultant quality of supervised modeling and the number of, quality of, and method of selection of data points for repeated-labeling. To our knowledge, the typical strategy used in practice is what we call “round-robin” repeated-labeling, where cases are given a fixed number of labels—so we focus considerable attention in the paper to this strategy. A related important problem is how in practice to assess the generalization performance of a learned model with uncertain labels [28], which we do not consider in this paper. Prior research has addressed important problems necessary for a full labeling solution that uses multiple noisy labelers, such as estimating the quality of labelers [6, 26, 28], and learning with uncertain labels [13, 24, 25]. So we treat these topics quickly when they arise, and lean on the prior work.

Repeated-labeling using multiple noisy labelers is different from multiple label classification [3, 15], where one example could have multiple *correct* class labels. As we discuss in Section 5, repeated-labeling can apply regardless of the number of true class labels. The key difference is whether the labels are noisy. A closely related problem setting is described by Jin and Ghahramani [10]. In their variant of the multiple label classification problem, each example presents itself with a set mutually exclusive labels, one of which is correct. The setting for repeated-labeling has important differences: labels are acquired (at a cost); the same label may appear many times, and the true label may not appear at all. Again, the level of error in labeling is a key factor.

The consideration of data acquisition costs has seen increasing research attention, both explicitly (e.g., cost-sensitive learning [31], utility-based data mining [19]) and implicitly, as in the case of active learning [5]. Turney [31] provides a short but comprehensive survey of the different sorts of costs that should be considered, including data acquisition costs and labeling costs. Most previous work on cost-sensitive learning does not consider labeling cost, assuming that a fixed set of labeled training examples is given, and that the learner cannot acquire additional information during learning (e.g., [7, 8, 30]).

Active learning [5] focuses on the problem of costly label acquisition, although often the cost is not made explicit. Active learning (cf., optimal experimental design [33]) uses the existing model to help select additional data for which to acquire labels [1, 14, 23]. The usual problem setting for active learning is in direct contrast to the setting we consider for repeated-labeling. For active learning, the assumption is that the cost of labeling is considerably higher than the cost of obtaining unlabeled examples (essentially zero for “pool-based” active learning).

Some previous work studies data acquisition cost explicitly. For example, several authors [11, 12, 16, 17, 22, 32, 37] study the costly acquisition of feature information, assuming that the labels are known in advance. Saar-Tsechansky et al. [22] consider acquiring both costly feature and label information.

None of this prior work considers selectively obtaining multiple labels for data points to improve labeling quality, and the relative advantages and disadvantages for improving model performance. An important difference from the setting for traditional active learning is that labeling strategies that use multiple noisy labelers have access to potentially relevant additional information. The multisets of existing labels intuitively should play a role in determining the examples for which to acquire additional labels. For example, presumably one would be less interested in getting another label for an example that already has a dozen identical labels, than for one with just two, conflicting labels.

## 3. REPEATED LABELING: THE BASICS

Figure 1 illustrates that the quality of the labels can have a marked effect on classification accuracy. Intuitively, using repeated-labeling to shift from a lower-$q$ curve to a higher-$q$ curve can, under some settings, improve learning considerably. In order to treat this more formally, we first introduce some terminology and simplifying assumptions.

## 3.1 Notation and Assumptions

We consider a problem of supervised induction of a (binary) classification model. The setting is the typical one, with some important exceptions. For each training example $\langle y_i, x_i \rangle$, procuring the *unlabeled* "feature" portion, $x_i$, incurs cost $C_U$. The action of *labeling* the training example with a label $y_i$ incurs cost $C_L$. For simplicity, we assume that each cost is constant across all examples. Each example $\langle y_i, x_i \rangle$ has a true label $y_i$, but labeling is error-prone. Specifically, each label $y_{ij}$ comes from a labeler $j$ exhibiting an individual labeling quality $p_j$, which is $Pr(y_{ij} = y_i)$; since we consider the case of binary classification, the label assigned by labeler $j$ will be incorrect with probability $1 - p_j$.

In the current paper, we work under a set of assumptions that allow us to focus on a certain set of problems that arise when labeling using multiple noisy labelers. First, we assume that $Pr(y_{ij} = y_i | x_i) = Pr(y_{ij} = y_i) = p_j$, that is, individual labeling quality is independent of the specific data point being labeled. We sidestep the issue of knowing $p_j$: the techniques we present do not rely on this knowledge. Inferring $p_j$ accurately should lead to improved techniques; Dawid and Skene [6] and Smyth et al. [26, 28] have shown how to use an expectation-maximization framework for estimating the quality of labelers when all labelers label all available examples. It seems likely that this work can be adapted to work in a more general setting, and applied to repeated-labeling. We also assume for simplicity that each labeler $j$ only gives one label, but that is not a restrictive assumption in what follows. We further discuss limitations and directions for future research in Section 5.

## 3.2 Majority Voting and Label Quality

To investigate the relationship between labeler quality, number of labels, and the overall quality of labeling using multiple labelers, we start by considering the case where for induction each repeatedly-labeled example is assigned a *single* "integrated" label $\hat{y}_i$, inferred from the individual $y_{ij}$'s by majority voting. For simplicity, and to avoid having to break ties, we assume that we always obtain an odd number of labels. The quality $q_i = Pr(\hat{y}_i = y_i)$ of the integrated label $\hat{y}_i$ will be called the *integrated quality*. Where no confusion will arise, we will omit the subscript $i$ for brevity and clarity.

### 3.2.1 Uniform Labeler Quality

We first consider the case where all labelers exhibit the same quality, that is, $p_j = p$ for all $j$ (we will relax this assumption later). Using $2N + 1$ labelers with uniform quality $p$, the integrated labeling quality $q$ is:

$$q = Pr(\hat{y} = y) = \sum_{i=0}^{N} \binom{2N+1}{i} \cdot p^{2N+1-i} \cdot (1-p)^i \qquad (1)$$

which is the sum of the probabilities that we have more correct labels than incorrect (the index $i$ corresponds to the number of incorrect labels).

Not surprisingly, from the formula above, we can infer that the integrated quality $q$ is greater than $p$ only when $p > 0.5$. When $p < 0.5$, we have an adversarial setting where $q < p$, and, not surprisingly, the quality decreases as we increase the number of labelers.

Figure 2 demonstrates the analytical relationship between the integrated quality and the number of labelers, for different individual labeler qualities. As expected, the integrated quality improves with larger numbers of labelers, when the individual labeling quality $p > 0.5$; however, the marginal improvement decreases as the number of labelers increases. Moreover, the benefit of getting more labelers also depends on the underlying value of $p$. Figure 3 shows how integrated quality $q$ increases compared to the case of single-labeling, for different values of $p$ and for different numbers of labelers. For example, when $p = 0.9$, there is little benefit when the number of labelers increase from 3 to 11. However, when $p = 0.7$, going just from single labeling to three labelers increases integrated quality by about 0.1, which in Figure 1 would yield a substantial upward shift in the learning curve (from the $q = 0.7$ to the $q = 0.8$ curve); in short, a small amount of repeated-labeling can have a noticeable effect for moderate levels of noise.

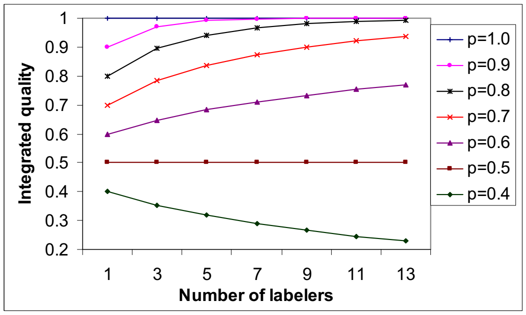


**Figure 2: The relationship between integrated labeling quality, individual quality, and the number of labelers.**

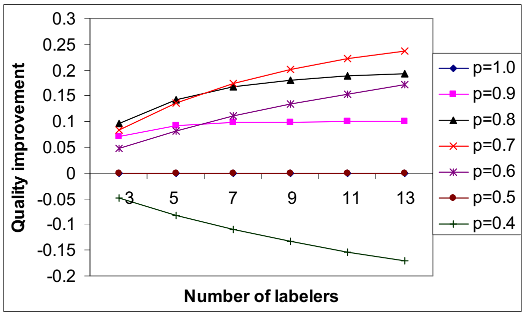


**Figure 3: Improvement in integrated quality compared to single-labeling, as a function of the number of labelers, for different labeler qualities.**

Therefore, for cost-effective labeling using multiple noisy labelers we need to consider: (a) the effect of the integrated quality $q$ on learning, and (b) the number of labelers required to increase $q$ under different levels of labeler quality $p$; we will return to this later, in Section 4.

### 3.2.2 Different Labeler Quality

If we relax the assumption that $p_j = p$ for all $j$, and allow labelers to have different qualities, a new question arises: what is preferable: using multiple labelers or using the best individual labeler? A full analysis is beyond the scope (and space limit) of this paper, but let us consider the special case that we have a group of three labelers, where the middle labeling quality is $p$, the lowest one is $p - d$, and the highest one is $p + d$. In this case, the integrated quality $q$ is:

$$\begin{aligned} (p-d) \cdot p \cdot (p+d) + (p-d) \cdot p \cdot (1-(p+d)) + \\ (p-d) \cdot (1-p) \cdot (p+d) + (1-(p-d)) \cdot p \cdot (p+d) = \\ -2p^3 + 2pd^2 + 3p^2 - d^2 \end{aligned}$$

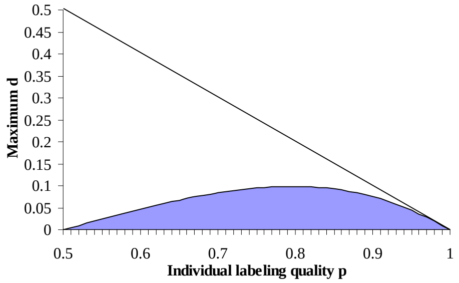


**Figure 4: Repeated-labeling improves quality when $d$ is below the curve (see text).**

When is this quantity greater than that of the best labeler $p + d$? We omit the derivation for brevity, but Figure 4 plots the values of $d$ that satisfy this relationship. If $d$ is below the curve, using multiple labelers improves quality; otherwise, it is preferable to use the single highest-quality labeler.

## 3.3 Uncertainty-preserving Labeling

Majority voting is a simple and straightforward method for integrating the information from multiple labels, but clearly with its simplicity comes a potentially serious drawback: information is lost about label uncertainty. In principle, an alternative is to move to some form of *"soft" labeling*, with the multiset of labels resulting in a probabilistic label for an example [25]. One concern with soft labeling is that even in cases where, in principle, modeling techniques should be able to incorporate soft labeling directly (which would be true for techniques such as naive Bayes, logistic regression, tree induction, and beyond), existing software packages do not accommodate soft labels. Fortunately, we can finesse this.

Consider the following straightforward method for integrating labels. For each unlabeled example $x_i$, the *multiplied examples (ME)* procedure considers the multiset of existing labels $L_i = \{y_{ij}\}$. *ME* creates one replica of $x_i$ labeled by each unique label appearing in $L_i$. Then, for each replica, *ME* assigns a weight $1/|L_i|$, where $|L_i|$ is the number of occurrences of this label in $L_i$. These weighted replicas can be used in different ways by different learning algorithms, e.g., in algorithms that take weights directly (such as cost-sensitive tree [29]), or in techniques like naive Bayes that naturally incorporate uncertain labels. Moreover, any importance-weighted classification problem can be reduced to a uniform-weighted classification problem [35], often performing better than handcrafted weighted-classification algorithms.

# 4. REPEATED-LABELING AND MODELING

The previous section examined when repeated-labeling can improve quality. We now consider when repeated-labeling should be chosen for *modeling*. What is the relationship to label quality? (Since we see that for $p = 1.0$ and $p = 0.5$, repeated-labeling adds no value.) How cheap (relatively speaking) does labeling have to be? For a given cost setting, is repeated-labeling much better or only marginally better? Can selectively choosing data points to label improve performance?

## 4.1 Experimental Setup

Practically speaking, the answers to these questions rely on the conditional distributions being modeled, and so we shift to an empirical analysis based on experiments with benchmark data sets.

To investigate the questions above, we present experiments on 12 real-world datasets from [2] and [36]. These datasets were chosen because they are classification problems with a moderate number of examples, allowing the development of learning curves based on a large numbers of individual experiments. The datasets are described in Table 1. If necessary, we convert the target to binary (for *thyroid* we keep the negative class and integrate the other three classes into positive; for *splice*, we integrate classes IE and EI; for *waveform*, we integrate class 1 and 2.)

| Data Set | #Attributes | #Examples | Pos | Neg |
|---|---|---|---|---|
| bmg | 41 | 2417 | 547 | 1840 |
| expedia | 41 | 3125 | 417 | 2708 |
| kr-vs-kp | 37 | 3196 | 1669 | 1527 |
| mushroom | 22 | 8124 | 4208 | 3916 |
| qvc | 41 | 2152 | 386 | 1766 |
| sick | 30 | 3772 | 231 | 3541 |
| spambase | 58 | 4601 | 1813 | 2788 |
| splice | 61 | 3190 | 1535 | 1655 |
| thyroid | 30 | 3772 | 291 | 3481 |
| tic-tac-toe | 10 | 958 | 332 | 626 |
| travelocity | 42 | 8598 | 1842 | 6756 |
| waveform | 41 | 5000 | 1692 | 3308 |

**Table 1: The 12 datasets used in the experiments: the numbers of attributes and examples in each, and the split into positive and negative examples.**

For each dataset, 30% of the examples are held out, in every run, as the test set from which we calculate generalization performance. The rest is the "pool" from which we acquire unlabeled and labeled examples. To simulate noisy label acquisition, we first hide the labels of all examples for each dataset. At the point in an experiment when a label is acquired, we generate a label according to the labeler quality $p$: we assign the example's original label with probability $p$ and the opposite value with probability $1 - p$.

After obtaining the labels, we add them to the training set to induce a classifier. For the results presented, models are induced with J48, the implementation of C4.5 [21] in WEKA [34]. The classifier is evaluated on the test set (with the true labels). Each experiment is repeated 10 times with a different random data partition, and average results are reported.

## 4.2 Generalized Round-robin Strategies

We first study the setting where we have the choice of either:

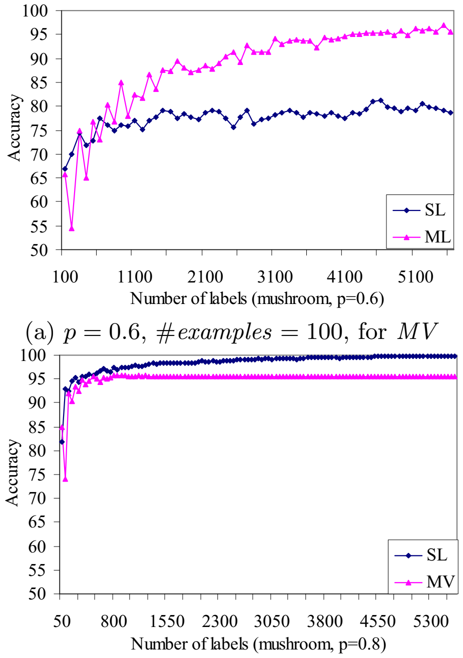
- acquiring a new training example for cost $C_U + C_L$, ($C_U$ for the *unlabeled* portion, and $C_L$ for the label), or
- get another label for an existing example for cost $C_L$.

We assume for this section that examples are selected from the unlabeled pool at random and that repeated-labeling selects examples to re-label in a *generalized round-robin* fashion: specifically, given a set $L$ of to-be-labeled examples (a subset of the entire set of examples) the next label goes to the example in $L$ with the fewest labels, with ties broken according to some rule (in our case, by cycling through a fixed order).

### 4.2.1 Round-robin Strategies, $C_U \ll C_L$

When $C_U \ll C_L$, then $C_U + C_L \approxeq C_L$ and intuitively it may seem that the additional information on the conditional label distribution brought by an additional whole training example, even with a noisy label, would outweigh the cost-equivalent benefit of a single new label. However, Figure 1 suggests otherwise, especially when considered together with the quality improvements illustrated in Figure 3.

Figure 5 shows the generalization performance of repeated-labeling with majority vote ($MV$) compared to that of single labeling ($SL$), as a function of the *number of labels* acquired



(a) $p = 0.6$, $\#examples = 100$, for $MV$

(b) $p = 0.8$, $\#examples = 50$, for $MV$

**Figure 5: Comparing the increase in accuracy for the *mushroom* data set as a function of the number of labels acquired, when the cost of an unlabeled example is negligible, i.e., $C_U = 0$. Repeated-labeling with majority vote ($MV$) starts with an existing set of examples and only acquires additional labels for them, and single labeling ($SL$) acquires additional examples. Other data sets show similar results.**

for a fixed labeler quality. Both $MV$ and $SL$ start with the same number of single-labeled examples. Then, $MV$ starts acquiring additional labels *only* for the *existing* examples, while $SL$ acquires new examples and labels them.

Generally, whether to invest in another whole training example or another label depends on the gradient of generalization performance as a function of obtaining another label or a new example. We will return to this when we discuss future work, but for illustration, Figure 5 shows scenarios for our example problem, where each strategy is preferable to the other. From Figure 1 we see that for $p = 0.6$, and with 100 examples, there is a lot of headroom for repeated-labeling to improve generalization performance by improving the overall labeling quality. Figure 5(a) indeed shows that for $p = 0.6$, repeated-labeling does improve generalization performance (per label) as compared to single-labeling new examples. On the other hand, for high initial quality or steep sections of the learning curve, repeated-labeling may not compete with single labeling. Figure 5(b) shows that single labeling performs better than repeated-labeling when we have a fixed set of 50 training examples with labeling quality $p = 0.8$. Particularly, repeated-labeling could not further improve its performance after acquiring a certain amount of labels (cf., the $q = 1$ curve in Figure 1).

The results for other datasets are similar to Figure 5: under noisy labels, and with $C_U \ll C_L$, round-robin repeated-labeling can perform better than single-labeling when there are enough training examples, i.e., after the learning curves are not so steep (cf., Figure 1).

### 4.2.2 Round-robin Strategies, General Costs

We illustrated above that repeated-labeling is a viable alternative to single-labeling, even when the cost of acquiring the "feature" part of an example is negligible compared to the cost of label acquisition. However, as described in the introduction, often the cost of (noisy) label acquisition $C_L$ is low compared to the cost $C_U$ of acquiring an unlabeled example. In this case, clearly repeated-labeling should be considered: using multiple labels can shift the learning curve up significantly. To compare any two strategies on equal footing, we calculate generalization performance "per unit cost" of acquired data; we then compare the different strategies for combining multiple labels, under different individual labeling qualities.

We start by defining the data acquisition cost $C_D$:

$$C_D = C_U \cdot T_r + C_L \cdot N_L \qquad (2)$$

to be the sum of the cost of acquiring $T_r$ unlabeled examples ($C_U \cdot T_r$), plus the cost of acquiring the associated $N_L$ labels ($C_L \cdot N_L$). For single labeling we have $N_L = T_r$, but for repeated-labeling $N_L > T_r$.

We extend the setting of Section 4.2.1 slightly: repeated-labeling now acquires and labels new examples; single labeling $SL$ is unchanged. Repeated-labeling again is generalized round-robin: for each new example acquired, repeated-labeling acquires a fixed number of labels $k$, and in this case $N_L = k \cdot T_r$. (In our experiments, $k = 5$.) Thus, for round-robin repeated-labeling, in these experiments the cost setting can be described compactly by the cost ratio $\rho = \frac{C_U}{C_L}$, and in this case $C_D = \rho \cdot C_L \cdot T_r + k \cdot C_L \cdot T_r$, i.e.,

$$C_D \propto \rho + k \qquad (3)$$

We examine two versions of repeated-labeling, repeated-labeling with majority voting ($MV$) and uncertainty-preserving repeated-labeling ($ME$), where we generate *multiple examples* with different weights to preserve the uncertainty of the label multiset as described in Section 3.3.

**Performance of different labeling strategies**: Figure 6 plots the generalization accuracy of the models as a function of data acquisition cost. Here $\rho = 3$, and we see very clearly that for $p = 0.6$ both versions of repeated-labeling are preferable to single labeling. $MV$ and $ME$ outperform $SL$ consistently (on all but waveform, where $MV$ ties with $SL$) and, interestingly, the comparative performance of repeated-labeling tends to increase as one spends more on labeling.

Figure 7 shows the effect of the cost ratio $\rho$, plotting the average improvement per unit cost of $MV$ over $SL$ as a function of $\rho$. Specifically, for each data set the vertical differences between the curves are averaged across all costs, and then these are averaged across all data sets. The figure shows that the general phenomenon illustrated in Figure 6 is not tied closely to the specific choice of $\rho = 3$.

Furthermore, from the results in Figure 6, we can see that the uncertainty-preserving repeated-labeling $ME$ always performs at least as well as $MV$ and in the majority of the cases $ME$ outperforms $MV$. This is not apparent in all graphs, since Figure 6 only shows the beginning part of the learning curves for $MV$ and $ME$ (because for a given cost, $SL$ uses up training examples quicker than $MV$ and $ME$). However, as the number of training examples increases further, then (for $p = 0.6$) $ME$ outperforms $MV$. For example, Figure 8 illustrates for the *splice* dataset, comparing the two techniques for a larger range of costs.

In other results (not shown) we see that when labeling quality is substantially higher (e.g., $p = 0.8$), repeated-labeling still is increasingly preferable to single labeling as $\rho$ increases; however, we no longer see an advantage for $ME$ over $MV$. These results suggest that when labeler quality is low, inductive modeling often can benefit from the explicit representation

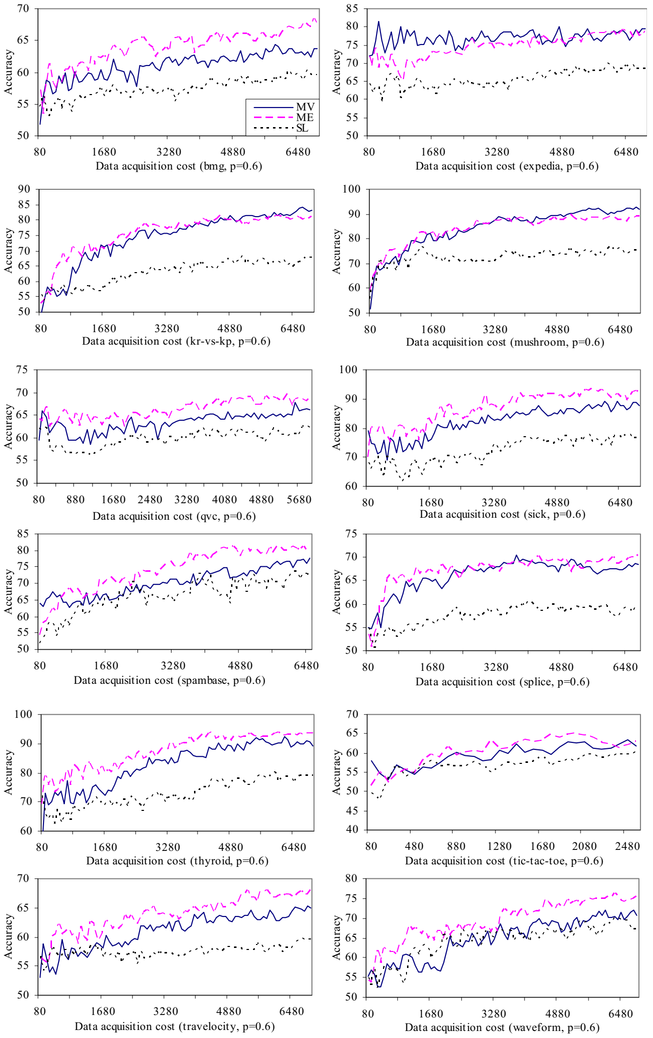


**Figure 6: Increase in model accuracy as a function of data acquisition cost for the 12 datasets; ($p = 0.6$, $\rho = 3$, $k = 5$). *SL* is single labeling; *MV* is repeated-labeling with majority voting, and *ME* is uncertainty-preserving repeated-labeling.**

of the uncertainty incorporated in the multiset of labels for each example. When labeler quality is relatively higher, this additional information apparently is superfluous, and straight majority voting is sufficient.

## 4.3 Selective Repeated-Labeling

The final questions this paper examines are (1) whether selective allocation of labeling resources can further improve performance, and (2) if so, how should the examples be selected. For example, intuitively it would seem better to augment the label multiset $\{+, -, +\}$ than to augment $\{+, +, +, +, +\}$.

### 4.3.1 What Not To Do

The example above suggests a straightforward procedure for selective repeated-labeling: acquire additional labels for those examples where the current multiset of labels is impure. Two natural measures of purity are (i) the entropy of the multiset of labels, and (ii) how close the frequency of the majority label is to the decision threshold (here, 0.5). These two measures rank the examples the same. Unfortunately, there is a clear problem: under noise these measures do not really measure the uncertainty in the estimation of the class label. For example, $\{+, +, +\}$ is perfectly pure, but the true class is not certain (e.g., with $p = 0.6$ one is not 95% confident of the true label). Applying a small-sample shrinkage correction (e.g., Laplace) to the probabilities is not sufficient. Figure 9 demonstrates how labeling quality increases as a function of assigned labels, using the (Laplace-corrected) entropy-based estimation of uncertainty (ENTROPY). For small amounts of repeated-labeling the technique does indeed select useful examples to label, but the fact that the estimates are not true estimates of uncertainty hurts the procedure in the long run—generalized round-robin repeated-labeling (GRR) from Section 4.2 outperforms the entropy-based approach. This happens because most of the labeling resources are wasted, with the procedure labeling a small set of examples very many times. Note that with a high noise level, the long-run label mixture will be quite impure, even though the true class of the example may be quite certain (e.g., consider the case of 600 positive labels and 400 negative labels with $p = 0.6$). More-pure, but incorrect, label multisets are never revisited.

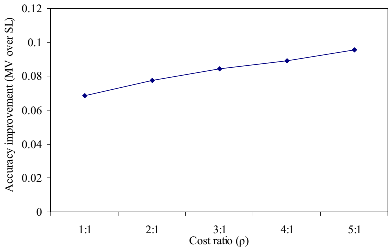


**Figure 7: The average improvement per unit cost of repeated-labeling with majority voting (*MV*) over single labeling (*SL*).**

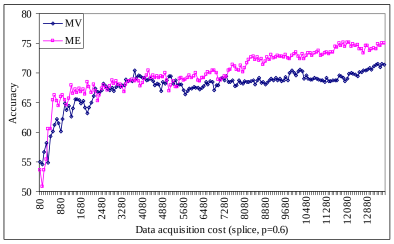


**Figure 8: The learning curves of *MV* and *ME* with $p = 0.6$, $\rho = 3$, $k = 5$, using the *splice* dataset.**

### 4.3.2 Estimating Label Uncertainty

For a given multiset of labels, we compute a Bayesian estimate of the uncertainty in the class of the example. Specifically, we would like to estimate our uncertainty that the true class $y$ of the example is the majority class $y_m$ of the multiset. Consider a Bayesian estimation of the probability that $y_m$ is incorrect. Here we do not assume that we know (or have estimated well) the labeler quality,[6] and so we presume the prior distribution over the true label (quality) $p(y)$ to be uniform in the $[0, 1]$ interval. Thus, after observing $L_{pos}$ positive labels and $L_{neg}$ negative labels, the posterior probability $p(y)$ follows a Beta distribution $B(L_{pos} + 1, L_{neg} + 1)$ [9].

We compute the level of uncertainty as the tail probability below the labeling decision threshold. Formally, the uncertainty is equal to the CDF at the decision threshold of the Beta distribution, which is given by the regularized incomplete beta function $I_x(\alpha, \beta) = \sum_{j=a}^{\alpha+\beta-1} \frac{(\alpha+\beta-1)!}{j!(\alpha+\beta-1-j)!} x^j (1-x)^{\alpha+\beta-1-j}$. In our case, the decision threshold is $x = 0.5$, and $\alpha =$

[6] Doing so may improve the results presented below.

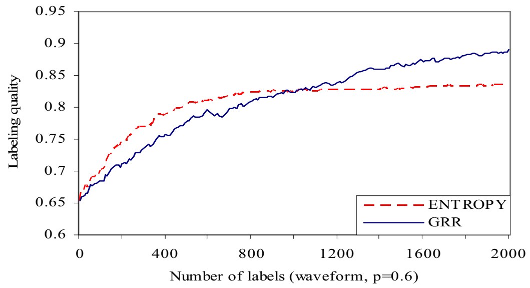
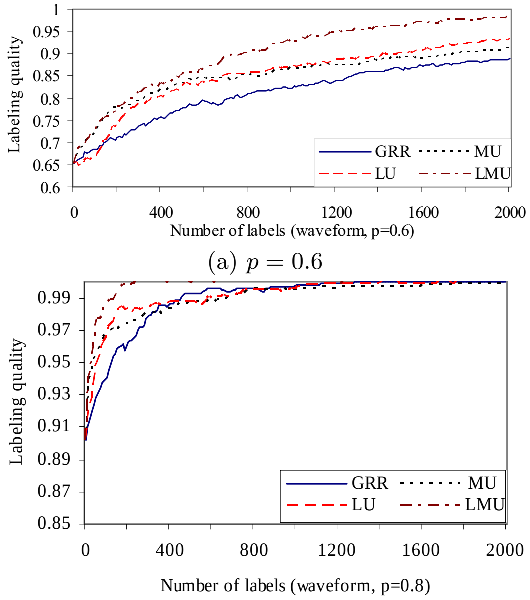


**Figure 9: What not to do: data quality improvement for an entropy-based selective repeated-labeling strategy vs. round-robin repeated-labeling.**



(a) $p = 0.6$

(b) $p = 0.8$

**Figure 10: The data quality improvement of the four strategies (*GRR*, *LU*, *MU*, and *LMU*) for the *waveform* dataset.**

$L_{pos} + 1, \beta = L_{neg} + 1$. Thus, we set:

$$S_{LU} = \min\{I_{0.5}(L_{pos}, L_{neg}), 1 - I_{0.5}(L_{pos}, L_{neg})\} \tag{4}$$

We compare selective repeated-labeling based on $S_{LU}$ to round-robin repeated-labeling (GRR), which we showed to perform well in Section 4.2. To compare repeated-labeling strategies, we followed the experimental procedure of Section 4.2, with the following modification. Since we are asking whether label uncertainty can help with the selection of examples for which to obtain additional labels, each training example starts with three initial labels. Then, each repeated-labeling strategy iteratively selects examples for which it acquires additional labels (two at a time in these experiments).

Comparing selective repeated-labeling using $S_{LU}$ (call that *LU*) to *GRR*, we observed similar patterns across all twelve data sets; therefore we only show the results for the *waveform* dataset (Figure 10; ignore the *MU* and *LMU* lines for now, we discuss these techniques in the next section), which are representative. The results indicate that *LU* performs substantially better than *GRR*, identifying the examples for which repeated-labeling is more likely to improve quality.

### 4.3.3 Using Model Uncertainty

A different perspective on the certainty of an example's label can be borrowed from active learning. If a predictive model has high confidence in the label of an example, perhaps we should expend our repeated-labeling resources elsewhere.

| Data Set | GRR | MU | LU | LMU |
|---|---|---|---|---|
| bmg | *62.97* | **71.90** | 64.82 | 68.93 |
| expedia | *80.61* | 84.72 | 81.72 | **85.01** |
| kr-vs-kp | 76.75 | *76.71* | 81.25 | **82.55** |
| mushroom | *89.07* | 94.17 | 92.56 | **95.52** |
| qvc | *64.67* | **76.12** | 66.88 | 74.54 |
| sick | *88.50* | 93.72 | 91.06 | **93.75** |
| spambase | *72.79* | 79.52 | 77.04 | **80.69** |
| splice | 69.76 | *68.16* | **73.23** | 73.06 |
| thyroid | *89.54* | 93.59 | 92.12 | **93.97** |
| tic-tac-toe | *59.59* | 62.87 | 61.96 | **62.91** |
| travelocity | *64.29* | **73.94** | 67.18 | 72.31 |
| waveform | *65.34* | 69.88 | 66.36 | **70.24** |
| **average** | *73.65* | 78.77 | 76.35 | **79.46** |

**Table 2: Average accuracies of the four strategies over the 12 datasets, for $p = 0.6$. For each dataset, the best performance is in boldface and the worst in italics.**

- **Model Uncertainty (MU)** applies traditional *active learning* scoring, ignoring the current multiset of labels. Specifically, for the experiments below the model-uncertainty score is based on learning a set of models, each of which predicts the probability of class membership, yielding the uncertainty score:

$$S_{MU} = 0.5 - \left| \frac{1}{m} \sum_{i=1}^{m} Pr(+|x, H_i) - 0.5 \right| \tag{5}$$

  where $Pr(+|x, H_i)$ is the probability of classifying the example $x$ into $+$ by the learned model $H_i$, and $m$ is the number of learned models. In our experiments, $m = 10$, and the model set is a *random forest* [4] (generated by WEKA).

Of course, by ignoring the label set, *MU* has the complementary problem to *LU*: even if the model is uncertain about a case, should we acquire more labels if the existing label multiset is very certain about the example's class? The investment in these labels would be wasted, since they would have a small effect on either the integrated labels or the learning.

- **Label and Model Uncertainty (LMU)** combines the two uncertainty scores to avoid examples where either model is certain. This is done by computing the score $S_{LMU}$ as the geometric average of $S_{LU}$ and $S_{MU}$. That is:

$$S_{LMU} = \sqrt{S_{MU} \cdot S_{LU}} \tag{6}$$

Figure 10 demonstrates the improvement in data quality when using model information. We can observe that the *LMU* model strongly dominates all other strategies. In high-noise settings ($p = 0.6$) *MU* also performs well compared to *GRR* and *LU*, indicating that when noise is high, using learned models helps to focus the investment in improving quality. In settings with low noise ($p = 0.8$), *LMU* continues to dominate, but *MU* no longer outperforms *LU* and *GRR*.

### 4.3.4 Model Performance with Selective ML

So, finally, let us assess whether selective repeated-labeling accelerates learning (i.e., improves model generalization performance, in addition to data quality). Again, experiments are conducted as described above, except here we compute

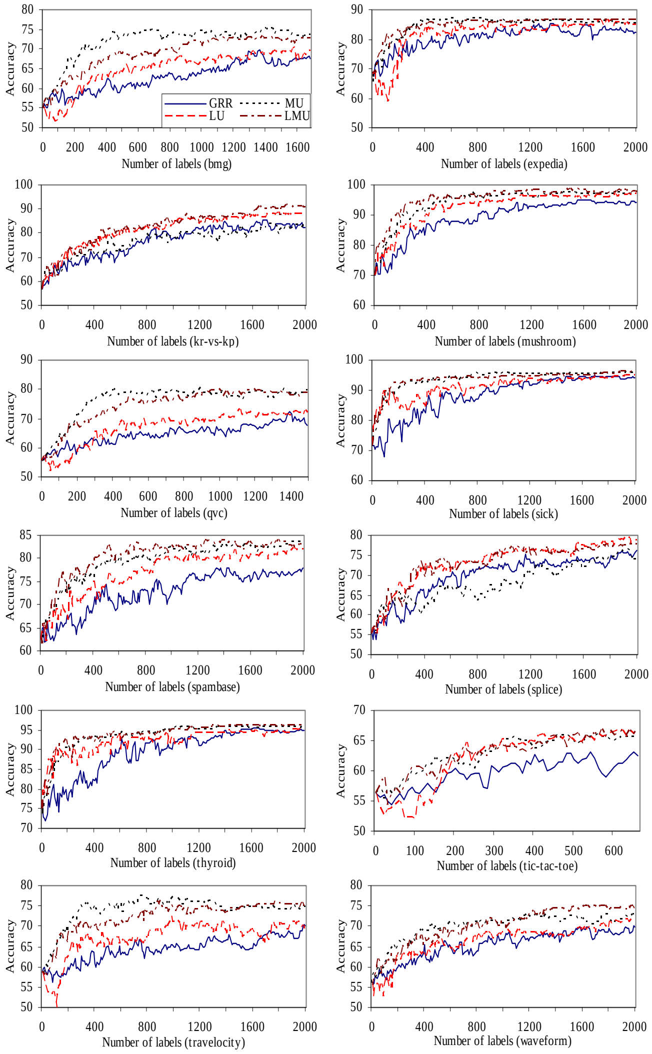


**Figure 11: Accuracy as a function of the number of labels acquired for the four selective repeated-labeling strategies for the 12 datasets ($p = 0.6$).**

generalization accuracy averaged over the held-out test sets (as described in Section 4.1). The results (Figure 11) show that the improvements in data quality indeed do accelerate learning. (We report values for $p = 0.6$, a high-noise setting that can occur in real-life training data.[7]) Table 2 summarizes the results of the experiments, reporting accuracies averaged across the acquisition iterations for each data set, with the maximum accuracy across all the strategies highlighted in bold, the minimum accuracy italicized, and the grand averages reported at the bottom of the columns.

The results are satisfying. The two methods that incorporate label uncertainty ($LU$ and $LMU$) are consistently better than round-robin repeated-labeling, achieving higher accuracy for every data set. (Recall that in the previous section, round-robin repeated-labeling was shown to be substantially better than the baseline single labeling in this setting.) The performance of model uncertainty alone ($MU$), which can be viewed as the active learning baseline, is more variable: in three cases giving the best accuracy, but in other cases not even reaching the accuracy of round-robin repeated-labeling. Overall, combining label and model uncertainty ($LMU$) is the best approach: in these experiments $LMU$ always outperforms round-robin repeated-labeling, and as hypothesized, generally it is better than the strategies based on only one type of uncertainty (in each case, statistically significant by a one-tailed sign test at $p < 0.1$ or better).

## 5. CONCLUSIONS, LIMITATIONS, AND FUTURE WORK

Repeated-labeling is a tool that should be considered whenever labeling might be noisy, but can be repeated. We showed that under a wide range of conditions, it can improve both the quality of the labeled data directly, and the quality of the models learned from the data. In particular, selective repeated-labeling seems to be preferable, taking into account both labeling uncertainty and model uncertainty. Also, when quality is low, preserving the uncertainty in the label multisets for learning [25] can give considerable added value.

Our focus in this paper has been on improving data quality for supervised learning; however, the results have implications for data mining generally. We showed that selective repeated-labeling improves the data quality directly and substantially. Presumably, this could be helpful for many data mining applications.

This paper makes important assumptions that should be visited in future work, in order for us to understand practical repeated-labeling and realize its full benefits.

- For most of the work we assumed that all the labelers have the same quality $p$ and that we do not know $p$. As we showed briefly in Section 3.2.2, differing qualities complicates the picture. On the other hand, good estimates of individual labelers' qualities inferred by observing the assigned labels [6, 26, 28] could allow more sophisticated selective repeated-labeling strategies.
- 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 $C_U$ and $C_L$, and to build models that dynamically increase or decrease the amounts paid to the labelers, depending on the quality requirements of the task. Morrison and Cohen [18] determine the optimal amount to pay for noisy information in a decision-making context, where the amount paid affects the level of noise.
- In our experiments, we introduced noise to existing, benchmark datasets. Future experiments, that use real labelers (e.g., using Mechanical Turk) should give a better understanding on how to better use repeated-labeling strategies in a practical setting. For example, in practice we expect labelers to exhibit different levels of noise and to have correlated errors; moreover, there may not be sufficiently many labelers to achieve very high confidence for any particular example.
- 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 for future work. We have not yet explored to what extent techniques like LMU (which are agnostic to the difficulty of

[7] From [20]: *"No two experts, of the 5 experts surveyed, agreed upon diagnoses more than 65% of the time. This might be evidence for the differences that exist between sites, as the experts surveyed had gained their expertise at different locations. If not, however, it raises questions about the correctness of the expert data."*

labeling) would deal naturally with example-conditional qualities.

- We also assumed that $C_L$ and $C_U$ are fixed and indivisible. Clearly there are domains where $C_L$ and $C_U$ would differ for different examples, and could even be broken down into different acquisition costs for different features. Thus, repeated-labeling may have to be considered in tandem with costly feature-value acquisition. Indeed, feature-value acquisition may be noisy as well, so one could envision a generalized repeated-labeling problem that includes both costly, noisy feature acquisition and label acquisition.
- In this paper, we consider the labeling process to be 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, but we leave it for future analysis.
- 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 a *"learning curve gradient"-based* 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.

Despite these limitations, we hope that this study provides a solid foundation on which future work can build. Furthermore, we believe that both the analyses and the techniques introduced can have immediate, beneficial practical application.

## Acknowledgements

Thanks to Carla Brodley, John Langford, and Sanjoy Dasgupta for enlightening discussions and comments. This work was supported in part by the National Science Foundation under Grant No. IIS-0643846, by an NSERC Postdoctoral Fellowship, and by an NEC Faculty Fellowship

## References

[1] Baram, Y., El-Yaniv, R., and Luz, K. Online choice of active learning algorithms. *Journal of Machine Learning Research 5* (Mar. 2004), 255–291.

[2] Blake, C. L., and Merz, C. J. UCI repository of machine learning databases. http://www.ics.uci.edu/~mlearn/MLRepository.html, 1998.

[3] Boutell, M. R., Luo, J., Shen, X., and Brown, C. M. Learning multi-label scene classification. *Pattern Recognition 37*, 9 (Sept. 2004), 1757–1771.

[4] Breiman, L. Random forests. *Machine Learning 45*, 1 (Oct. 2001), 5–32.

[5] Cohn, D. A., Atlas, L. E., and Ladner, R. E. Improving generalization with active learning. *Machine Learning 15*, 2 (May 1994), 201–221.

[6] Dawid, A. P., and Skene, A. M. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics 28*, 1 (Sept. 1979), 20–28.

[7] Domingos, P. MetaCost: A general method for making classifiers cost-sensitive. In *KDD* (1999), pp. 155–164.

[8] Elkan, C. The foundations of cost-sensitive learning. In *IJCAI* (2001), pp. 973–978.

[9] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. *Bayesian Data Analysis*, 2nd ed. Chapman and Hall/CRC, 2003.

[10] Jin, R., and Ghahramani, Z. Learning with multiple labels. In *NIPS* (2002), pp. 897–904.

[11] Kapoor, A., and Greiner, R. Learning and classifying under hard budgets. In *ECML* (2005), pp. 170–181.

[12] Lizotte, D. J., Madani, O., and Greiner, R. Budgeted learning of naive-bayes classifiers. In *UAI)* (2003), pp. 378–385.

[13] Lugosi, G. Learning with an unreliable teacher. *Pattern Recognition 25*, 1 (Jan. 1992), 79–87.

[14] Margineantu, D. D. Active cost-sensitive learning. In *IJCAI)* (2005), pp. 1622–1613.

[15] McCallum, A. Multi-label text classification with a mixture model trained by EM. In *AAAI'99 Workshop on Text Learning* (1999).

[16] Melville, P., Provost, F. J., and Mooney, R. J. An expected utility approach to active feature-value acquisition. In *ICDM* (2005), pp. 745–748.

[17] Melville, P., Saar-Tsechansky, M., Provost, F. J., and Mooney, R. J. Active feature-value acquisition for classifier induction. In *ICDM* (2004), pp. 483–486.

[18] Morrison, C. T., and Cohen, P. R. Noisy information value in utility-based decision making. In *UBDM'05: Proceedings of the First International Workshop on Utility-based Data Mining* (2005), pp. 34–38.

[19] Provost, F. Toward economic machine learning and utility-based data mining. In *UBDM '05: Proceedings of the 1st International Workshop on Utility-based Data Mining* (2005), pp. 1–1.

[20] Provost, F., and Danyluk, A. Learning from Bad Data. In *Proceedings of the ML-95 Workshop on Applying Machine Learning in Practice* (1995).

[21] Quinlan, J. R. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, Inc., 1992.

[22] Saar-Tsechansky, M., Melville, P., and Provost, F. J. Active feature-value acquisition. Tech. Rep. IROM-08-06, University of Texas at Austin, McCombs Research Paper Series, Sept. 2007.

[23] Saar-Tsechansky, M., and Provost, F. Active sampling for class probability estimation and ranking. *Journal of Artificial Intelligence Research 54*, 2 (2004), 153–178.

[24] Silverman, B. W. Some asymptotic properties of the probabilistic teacher. *IEEE Transactions on Information Theory 26*, 2 (Mar. 1980), 246–249.

[25] Smyth, P. Learning with probabilistic supervision. In *Computational Learning Theory and Natural Learning Systems, Vol. III: Selecting Good Models*, T. Petsche, Ed. MIT Press, Apr. 1995.

[26] Smyth, P. Bounds on the mean classification error rate of multiple experts. *Pattern Recognition Letters 17*, 12 (May 1996).

[27] Smyth, P., Burl, M. C., Fayyad, U. M., and Perona, P. Knowledge discovery in large image databases: Dealing with uncertainties in ground truth. In *Knowledge Discovery in Databases: Papers from the 1994 AAAI Workshop (KDD-94)* (1994), pp. 109–120.

[28] Smyth, P., Fayyad, U. M., Burl, M. C., Perona, P., and Baldi, P. Inferring ground truth from subjective labelling of Venus images. In *NIPS* (1994), pp. 1085–1092.

[29] Ting, K. M. An instance-weighting method to induce cost-sensitive trees. *IEEE Transactions on Knowledge and Data Engineering 14*, 3 (Mar. 2002), 659–665.

[30] Turney, P. D. Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm. *Journal of Artificial Intelligence Research 2* (1995), 369–409.

[31] Turney, P. D. Types of cost in inductive concept learning. In *Proceedings of the ICML-2000 Workshop on Cost-Sensitive Learning* (2000), pp. 15–21.

[32] Weiss, G. M., and Provost, F. J. Learning when training data are costly: The effect of class distribution on tree induction. *Journal of Artificial Intelligence Research 19* (2003), 315–354.

[33] Whittle, P. Some general points in the theory of optimal experimental design. *Journal of the Royal Statistical Society, Series B (Methodological) 35*, 1 (1973), 123–130.

[34] Witten, I. H., and Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. Morgan Kaufmann Publishing, June 2005.

[35] Zadrozny, B., Langford, J., and Abe, N. Cost-sensitive learning by cost-proportionate example weighting. In *ICDM* (2003), pp. 435–442.

[36] Zheng, Z., and Padmanabhan, B. Selectively acquiring customer information: A new data acquisition problem and an active learning-based solution. *Management Science 52*, 5 (May 2006), 697–712.

[37] Zhu, X., and Wu, X. Cost-constrained data acquisition for intelligent data preparation. *IEEE TKDE 17*, 11 (Nov. 2005), 1542–1556.