International Journal of Engineering Research
Volume No.8, Issue No.4, pp : 99-103
ISSN:2319-6890 (online),2347-5013(print)
1 Oct. 2019

Data Mining using Repeated Labeling Technique for Predicting Unknown Class Label

[1]A. Ramesh Babu, [2]Raja Ashok Kumar, [3]Shaik Mahammad Rafi
[1,2,3] Assistant Professors, Department of Computer Science and Engineering
[1,2,3]Annamacharya Institute of Technology and Sciences::Rajampet, AP, India

Abstract : The current topic is discussing about predicting the class label attribute for the number of unknown samples is not correct. So we are going to describe the Repeated Labeling Technique to increase the efficiency, robustness and quality of the data for supervised learning methodology. With the outsourcing of small tasks becoming easier, for example via Rent-A-Coder or Amazon's Mechanical Turk, it of ten is possible too btainless-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 strate- gies 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 tradi-tional 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 issubstantial.

Categories and Subject Descriptors

H.2.8[**DatabaseApplications**]:Datamining;I.5.2[**Design Methodology**]: Classifier design andevaluation

GeneralTerms

Algorithms, Design, Experimentation, Management, Measurement, Performance

Keywords

data selection, data preprocessing.

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 soon[31,32]. For example, in order to build a model to recognize whether two products describe don't webpages 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 points. To build a model that recognizes whether an image contains an object of interest, one first needs to take pictures inappropriate contexts, sometimes at substantial cost. This paper focuses on problems where it is possible too- certain (noisy) data values ("labels") relatively cheaply, from multiple sources("labelers"). Again focus of this paper is the use of these values as training labels for unsupervised mod- eling.[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 ESPgame.[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. Figure1 shows learning curves under different labeling qualities for the *mushroom* data set (see Section4.1). 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. Figure1illustrates that the performance of a learned model

ISSN:2319-6890 (online),2347-5013(print)
1 Oct. 2019

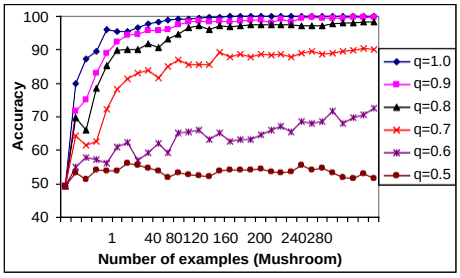

Figure 1: Learning curves under different quality lev- els of training data (*q* is the probability of a label being correct).

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 newlabels.

This paper makes several contributions. 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 Figure1, 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 considerableadvantage.

We present a comprehensive experimental analysis of the relationships between quality, cost, and technique for repeated- labeling. The results show that evens 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 n 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. RELATEDWORK

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, 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 Section5, 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 datamining[19]) and implicitly, a sin 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] studythe costly acquisition of feature information, assuming thatthe 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 therelative advantages and disadvantages for improving modelperformance. An important difference from the setting fortraditional active learning is that labeling strategies that usemultiple noisy labelers have access to potentially relevant additional information. The multisets of existing labels intuitivelyshould play a role in determining the examples for which toacquire additional labels. For example, presumably one wouldbe less interested in getting another label for an example thatalready has a dozen identical labels, than for one with just two, conflicting labels.

3. REPEATED-LABELINGANDMODELING

The previous section examined when repeated-labeling can improvequality.Wenowconsiderwhenrepeated-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?

3.1 ExperimentalSetup

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 datasets.

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 ofTable1: The 12 datasets used in the experiments: the numbers of attributes and examples in each, and thesplitintopositiveandnegativeexamples.

Learning curves based on a large numbers of individual experiments. The datasets are described in Table1.If necessary, we convert the target to binary (for *thyroid* we keep the nega- tive class and integrate the other three classes into positive; for *splice*, we integrate classes IE and EI; for *waveform*, we integrate class 1 and2.)

For each dataset, 30% of the examples are held out, in every run, as the test set from which we calculate general-ization 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 probability1 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 evaluate the test set(with the true labels). Each experiment is repeated 10 times with a different random data partition, and average results are reported.

3.2 Generalized Round-robin Strategies

We first study the setting where we have the choice of either: acquiring a new training example for cost $CU+CL$,(CU for the *unlabeled* portion, and CL for the label),or get another label for an existing example for cost CL.

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) then extra 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).

3.2.1 Round-robin Strategies, $CUCL$

When CU CL, then $CU + CL \div CL$ and intuitively it mayseemthattheadditionalinformationontheconditional label distribution brought by an additional whole training example, even with a noisy label, would outweigh thecost- equivalent benefit of a single new label. However, Figure1 suggests otherwise, especially when considered together with the quality improvements illustrated inFigure3.

Figure5shows 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 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 in correct, label multiset are never revisited

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 in correct, label multisets are never revisited.

3.2.2 Estimating LabelUncertainty

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 of the example is the majority class ym of the multiset. Consider a Bayesian estimation of the probability that ym is incorrect. Here we do not assume that we know (or have estimated well) the labeler quality,6and so we presume the prior distribution over the true label(quality)p(y) to be uni- form in the[0,1] interval. Thus, after observing Lpos positive labels and Lneg negative labels, the posterior probability p(y) follows a Beta distribution B(Lpos+ 1, Lneg+ 1)[9]. We compute the level of uncertainty as the tail probability below the labeling decision threshold. Formally, the uncer- tainty is equal to the CDF at the decision threshold of the Beta distribution, which is given by the regularized in complete beta function $Ix(\alpha, \beta) = \alpha+\beta-1 \quad (\alpha+\beta-1)! \quad xj(1 - x)\alpha+\beta-1-j$. In our case, the decision threshold is x=0.5,6 Doing so may improve the results presented below.

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 CU and CL, 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 particularexample.

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 difficultyoflabeling)wouldealnaturallywithexample-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 downintodifferentacquisitioncostsfordifferentfeatures.
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 labelacquisition.

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 comparedrepeated-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 compareon-the-flytheexpectedbenefitofacquiringnew 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 practicalapplication.

References

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

ii. *Blake, C. L., and Merz, C. J. http://www.ics.uci.edu/~mlearn/MLRepository.html, 1998.*

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

iv. *Breiman, L.Randomforests. Machine Learning45,1(Oct.2001), 5–32.*

v. *Cohn, D. A., Atlas, L. E., and Ladner, R. E. Improving generalization with active learning. Machine Learning 15,2(May1994), 201–221.*

vi. *Dawid, A. P., and Skene, A. M. Maximum likelihood estimation ofobservererror-ratesusingtheEMalgorithm.AppliedStatistics 28, 1 (Sept. 1979),20–28.*

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

viii. *Elkan, C. The foundations of cost-sensitive learning. InIJCAI*

ix. *(2001), pp. 973–978.*

x. *Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D.B.*

xi. Bayesian Data Analysis, 2nd ed. Chapman and Hall/CRC, 2003.
xii. Jin,R.,and Ghahramani, Z.Learning with multiple labels. In NIPS (2002), pp. 897–904.
xiii. Kapoor, A., and Greiner, R. Learning and classifying under hard budgets. In ECML (2005), pp.170–181.

Authors Proposed by:

1. A. Ramesh Babu,
Assistant Professor,
Department of C.S.E,
Annamacharya Institute of Technology and Sciences.
2. Raja Ashok Kumar,
Assistant Professor,
Department of C.S.E,
Annamacharya Institute of Technology and Sciences
3. Shaik Mahammad Rafi,
Assistant Professor,
Department of C.S.E,
Annamacharya Institute of Technology and Sciences.