Expanding annotated data with informed labels for weak supervision

Eura Shin, Sam Berglin, Jacob Furst and Daniela Raicu

October 3, 2019

Expanding annotated data with informed labels for weak supervision

Eura Shin[1], Sam Berglin[2], Jacob Furst[3], and Daniela Raicu[3]

[1] University of Kentucky, Lexington KY 40351, USA
[2] University of Wisconsin, Madison WI 53706
[3] DePaul University, Chicago IL 60604

Abstract. In this paper we present an instance of the weak supervision paradigm, the multi-uncertain learning scenario. Our multi-uncertain scenario has three facets, all which are related to instance labels and their corresponding human labelers: there are multiple labels per instance, the gold standard label may not be included in this label set, and the identity of the labelers is unknown. In order to avoid disposing of expensive and potentially useful labels, we outline a method of adding informed labels to a label set by using label propagation. Under the smoothness assumption, we are able to introduce new, informative labels into an existing training label set to improve performance under highly uncertain constraints. For complex classification tasks with three or more classes, we report that this method of adding informed labels is capable of producing classifiers with high accuracy and low complexity, despite being trained on these multi-uncertain datasets.

1 Introduction

In a typical supervised learning scenario, a chosen classifier is trained on a set of instances for which there is feature information and an associated gold standard or ground truth label. Realistically speaking, these gold standard labels are expensive, and sometimes impossible, to obtain. This is often the case in medical settings, such as in diagnosing lung cancer, where a radiologist may perform a diagnosis from a Computed Tomography (CT) scan of a suspicious lung nodule. In cases for which follow up information is not available, multiple experts may be asked to annotate (i.e. provide labels for) the same scan to formulate a stronger *reference truth*, which may or may not reflect the true gold standard.

In this paper we analyze a realistic learning problem, which we term a *multi uncertain* scenario, for which each instance is associated with a set of multiple labels, these labels are reference truths, and information on annotator identity is not available. The Lung Image Database Consortium (LIDC) is a real life example of this sort of learning scenario [1].

Multi-uncertain data sources such as the LIDC are becoming increasingly common with the onset of resources like Amazon's Mechanical Turk. These large data sets are characterized by their uncertain label sets, which when utilized, may result in inaccurate classifier performance. Research in the weak supervised learning paradigm has attempted to address this problem. Weak supervision attempts to unify and de-noise noisy data sets that may consist of disagreeing label sets, obtained from sources of varying reliability [12].

The weak learning paradigm may be broken into three standard modeling tasks: learning the accuracies, modeling correlations between weaker supervision sources, and modeling expertise (information on the accuracy of each labeler). The task of learning accuracies assumes no labeled data and a model structure and learns the weights of the model. This is inappropriate for our scenario, which includes labeled data from radiologists. Furthermore, it is not possible to monitor expertise in our scenario, in which we assume the source of each label in the data set is unknown (radiologists anonymously label CT scans). Thus, our task is reduced to that of modeling correlations. Specifically, our goal is to impose a structure that draws correlations between the behavior of the data in the feature space and the corresponding labels.

We outline a novel approach to adding what we term *informed labels* to a multi-uncertain label set for training classifiers. Within the weak supervision paradigm, we impose the smoothness assumption, which states that observations with similar features will have similar labels. We propagate informed labels within clusters to unify and de-noise existing label sets. We show that this simple strategy allows simple classifiers to match the performance of more complex classifiers by introducing informative, and potentially gold, labels to an uncertain label set. We provide an analysis of our method on five different classification methods and six University of California at Irvine (UCI) machine learning datasets. We consider accuracy, computational cost, interpretability, and confidence in assessing the quality of these classifiers for this weak learning problem. Using these evaluation criteria, we demonstrate that adding informed labels has the ability to produce the best classifier for datasets with a low saturation of gold labels, and in some cases, across all levels of uncertainty.

2 Related Work

2.1 Partial learning scenarios

We begin by reviewing current approaches to weakly supervised learning problems associated with multiple annotators. Otherwise known as a partial learning scenario, this is a case in which each training example is associated with a set of candidate labels, only one of which is the ground truth. This is similar to the semi-supervised multi-view learning problem discussed by Ceci et al. that combines multiple outputs of classifiers for better classification [4]. However, their work was tailored towards gene regulatory network reconstruction which was solely a binary classification problem with only positive and unlabeled data, while we define a multi-class problem. Partial learning problems make the strong assumption that the ground truth is included in the candidate set, one relaxed by our own multi-uncertain problem.

Modeling expertise In the case of inferring the ground truth from a set of novice annoated labels, Smyth et al. first introduced an implementation of the Expectation-Maximization (EM) algorithm [5]. Jin and Ghahramani also applied EM to instead learn a probabilistic classifier whose predicted probabilities match the probability distributions of the reference truths [7]. Other variations of this EM algorithm [16, 13, 19, 14, 17] attempt to learn better classifiers from noisy label sets by *modeling the expertise of the annotators.* Raykar et al. first incorporate annotator accuracies by modeling ground truth labels as latent variables in this model [13]. Yan et al. include active learning in this framework by attempting to identify the most useful annotator to label a given instance[7]. Rodrigues et al. explore the advantages of including annotator accuracies as latent variables [14]. All of these works take advantage of annotator identities to build stronger classifiers. However, these strategies are not applicable to our multi-uncertain learning scenario because they require labeler identities and assume the presence of a gold standard label in the candidate set.

Correlation Modeling Recent work has focused on taking advantage of information available in the feature space to model their correlations with the corresponding label sets, thus resulting in disambiguated candidate label sets [20, 22]. These papers make the smoothness assumption, which assumes instances closely packed in the feature space are likely to share the same label. We maintain this assumption in our attempt to reduce uncertainty using label propagation. However, disambiguation strategies are easily misled by the false-positive labels that occur with the gold standard in a candidate set. In other words, for a high number of annotators per instance, these strategies are distracted from the gold standard. [21]

2.2 Improving label quality and label propagation

Instead of identifying an algorithm that will distinguish high quality labels in an otherwise uncertain set, we consider strategies to improve the quality of the label set itself [15]. Brodley and Friedl create classifiers that serve as noisy filters on training data with potentially mislabeled instances [3]. In a more recent study on removing noise prior to training, Northcutt et al. use rank pruning to estimate error rates within the data set and remove mislabeled instances based on these error rates [11].

In practice, it does not make sense to disregard expensive, expert provided labels. We approach this challenge of improving label quality by *adding* informed labels rather than removing noisy ones.

Propagation is the process of extending information from a well known instance to a lesser known instance. It is often applied in a semi-supervised learning scenario, where known labels are extended unlabeled points that are nearby in the feature space. Zhu et al. proposed an algorithm that uses the labeled points to “push” labels to unlabeled instances [23] . Wang et al. presented a similar idea based on a linear neighborhood model [18]. Kang et al. expanded this idea to the *multi-label* problem, where each instance can have multiple correct labels [8]. Note the difference between the multi-label problem and uncertain labels. In multi-labels, each instance may have more than one *correct* label (such as with categorizing an image); in uncertain labels, the gold standard label may not exist in an instances given label set.

Rather than propagating labels exclusively to unlabeled points, we propagate labels to all points and add them within their uncertain label sets. This way we may leverage label information within the existing label set to extrapolate new, informed labels to train classifiers. Our contribution reduces noise within uncertain label sets without the need for annotator identities or filtering labels. To our knowledge, we are the first to apply this method of appending propagated labels to training sets that fall in the multi-uncertain scenario.

3 Methodology

3.1 Formal Definition of the Multi-Uncertain Scenario

In a standard supervised learning scenario, the training set $D = \{(\mathbf{x_i}, y_i)\}_{i=1}^{N}$ contains N instances, where $\mathbf{x_i} \in X$ is a feature vector of length M and $y_i \in Y$ is the corresponding known label. In machine learning literature, y_i is typically referred to as the gold standard or ground truth label G. In a training set with *noisy labels*, R different annotators, or experts, provide a set of labels, $\mathbf{y_i} = \{\mathbf{y_i}^1, \mathbf{y_i}^2, ..., \mathbf{y_i}^R\}$ for every i^{th} instance. A training set consists of *reference truths* when the gold standard y_G is not guaranteed to be included in a given label set, or, $Pr(y_G \in \mathbf{y_i}) \neq 1$. We make no assumptions for the label sets. We do not assume that the same R annotators are labeling each instance. For each instance we only have a set of labels which may or may not contain the gold standard y_G.

We reserve the term *multi-uncertain* for training sets that possess these three characteristics: 1) the ground truth is unknown, 2) the reference truth is uncertain or noisy, and 3) annotator identity is independent between instances.

We define the gold inclusion percentage P_G as the frequency at which the gold standard appears in the label set for the entirety of the training data. Specifically, $P_G = \frac{\sum_{i=1}^{N} \mathbf{1}(y_G \in \mathbf{y_i})}{N}$ where $\mathbf{1}(\cdot)$ is the indicator function.

It is apparent that a simple majority voting strategy alone will not suffice for datasets with low values of P_G. In the following sections we discuss a strategy for adding adding informed labels that will improve the accuracy of classifiers.

3.2 Informed Labels

We present a method of introducing a new set of labels y_g within the training data to map $Y \Rightarrow Y'$ where $\mathbf{y'_i} = \{\mathbf{y_i}^g, \mathbf{y_i}^1, \mathbf{y_i}^2, ..., \mathbf{y_i}^R\}$ for every instance $\mathbf{y'_i} \in Y'$. We use a cluster-based method that leverages label information from points surrounding an instance to derive a new label. Given a dataset with N instances and M features, we denote $\mathbf{x_i}$ as a feature vector of length M for a given instance i. We start with a hierarchical clustering of the set of feature vectors $\mathbf{x_i} \in X$ and prune the resulting dendrogram from the bottom-up. This pruning method searches for a set of clusters, $\{C_1, C_2, ..., C_k\}$, that are densely packed with respect to a minimum number of features, F.

Definition 1. *Let D be the distribution of all normalized values for a feature in a cluster, C. C is* densely packed *with respect to that feature if the standard deviation of D is no greater than a threshold, t.*

Each cluster $C_i \in \{C_1, C_2, ..., C_k\}$ must meet the requirement in Equation 1. The expression within the first set of brackets of the equation represents Definition 1. The pruning method has two input parameters, F and t, where F is the minimum number of densely packed features in a cluster and t is a specified threshold. We define $F = \frac{M}{2}$ and $t = \frac{1}{3}$ in our experiments. For the LIDC data, a parameter analysis showed our choice of t resulted in clusters that were characterized by similar features and of appropriate size.

$$\sum_{j=1}^{M} \mathbf{1}\left(\sigma\left(\bigcup_{\forall x_i \in C_i} x_{i,j}\right) \leq t\right) \geq F \tag{1}$$

where $x_{i,j}$ is the j^{th} feature of the i^{th} instance.

The entire set of labels within a cluster is then used to assign a unanimous label to all instances in the cluster. For a given C_i we consider the set of all uncertain labels, $l = y_{C_i}$, where $|l| = R|C_i|$. The mode of this set l is used as the informed label: $y^g_{C_i} = \text{mode}(l)$.

3.3 Generating Noise

In order to measure the uncertainty of the label set for a given instance, the golden standard for that instance must be known. In our analysis we generate uncertain label sets from a set of baseline, supervised classification data sets. Note that these supervised data sets will have a gold standard for every observation, $(\mathbf{x_i}, y_i)$ for all instances in X.

Uncertain label sets are generated by first parameterizing a distribution from the known pairs of observations and gold standards, $(\mathbf{x_i}, y_i)$. Let V be the set of all possible values for all labels in the dataset. The distribution f_{x_i} from which each label is randomly drawn is

$$f_{x_i}(y; p) = \begin{cases} p & y = y_i \\ \frac{1-p}{|V|} & y \neq y_i \text{ and } v \in V \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

In this definition, p is a parameter for the probability that the golden standard is included in the label set $\mathbf{y_i}$. This value is adjusted to alter the probability that the true label appears in the overall uncertain label set, P_G.

Definition 2. *P_G is the proportion of label sets within a data set that contain the gold standard label, y_G. Formally, this is denoted as*

$$P_G = \frac{\sum_{i=1}^{N} \mathbf{1}\Big(y_G \in \mathbf{y_i}\Big)}{N}$$

for a data set X with N observations.

More specifically, for a given uncertain label set $\mathbf{y_i} = \{\mathbf{y_i}^1, \mathbf{y_i}^2, ..., \mathbf{y_i}^R\}$ for every i^{th} instance, $P(y_G \notin \mathbf{y_i}) = (1-p)^{|y_i|}$. Thus, increasing p decreases the uncertainty for all generated uncertain label sets. Because we are given the gold standard for every observation, we can directly observe what proportion of the uncertain label sets contain their respective gold labels, or P_G. Adjusting p indirectly controls this value of P_G while maintaining independence with respect to the method in which random labels are selected.

The overall method of generating a multi-uncertain data set from a supervised data set is described by Algorithm 1.

Algorithm 1: Generation of multi-uncertain label sets

```
  input: probability of gold standard $p$
  Result: Uncertain label set $\mathbf{y_i}$ for all $\mathbf{x_i} \in X$
1 for $\mathbf{x_i} \in X$ do
2     set distribution $f_{\mathbf{x_i}}(y; p)$ as defined by Equation 2;
3     choose $R$ labels from this distribution to form a label set $\mathbf{y_i}$;
4 end
```

Line 2 defines a parameter as described in Equation 2. In Line 3, labels are drawn R times from this distribution to simulate receiving labels from R annotators.

4 Experiments

4.1 Experimental Data

We apply our methods to six UCI datasets: breast, iris, wine, class, e. coli, and yeast [6]. We chose well-known datasets within the UCI Repository that consisted of continuous features and discrete classes. Each data set is treated as a classification problem with nominal classes. The datasets vary in dimensionality and in the number of classes, as described in Table 1. We transform the gold standard labels into uncertain datasets as in 3.3. In testing, we use the original gold standard from the pair $(\mathbf{x_i}, y_i)$ to assess accuracy.

4.2 Classification Methods

We apply informed labels in training four different classification methods: CART, SVM, logistic regression, and the *EM* prior algorithm implemented in Jin et al

Dataset	Features	Classes	Instances
Breast	30	2	569
Iris	4	3	150
Wine	13	3	150
Glass	10	5	214
E. Coli	7	7	336
Yeast	8	10	1484

Table 1: Information about six UCI datasets used in the experiments.

[7]. Every classification algorithm is a supervised classifier. We use these classifiers because they are well-known and used classification techniques. We train the classifiers on the *mode* of each label set $\mathbf{y_i}$. The EM prior algorithm utilizes the raw, uncertain label sets. The accuracies reported in this paper are always with respect to the gold standard included in the UCI datasets.

We implement the EM prior algorithm described by Jin et al. because it does not attempt to learn labeler accuracies but still remains relevant and effective in the task of identifying the gold label. We expect to improve the performance of this algorithm for instances where the gold label is not guaranteed to be in a label set. We implement the EM prior algorithm rather than the EM algorithm because it is more appropriate for our label sets which will have prior class distributions, as a result of the process described in Section 3.3.

Logistic regression and EM prior models have no hyper-parameters to tune. CART trees are pruned through cross validation within the training set. SVMs are also tuned via a validation set with a tuning grid of $C \times \Gamma$, where $C = \{10^i\}_{i=-3}^{3}$ and $\Gamma = \{10^i\}_{i=-3}^{3}$.

4.3 Performance Evaluation

In addition to accuracy, we evaluate the classification performance using runtime, transparency, confidence, and interpretability. The runtime for each classifier was derived on a Dell Optiplex 7020 Desktop computer as the average of ten runs reported for each of the six UCI datasets. Transparency refers to how easily the principle of a classification method is understood by human intuition [2], whereas interpretability is the level of clarity to a user on how a classification method derived a prediction from training information [10]. The values for these two arguably subjective standards are provided by [2] and [9], comparative studies on different supervised algorithms. Finally, the confidence is measured through the probabilistic outcome of the learning algorithm itself. Statistical algorithms are specially considered, as predictions are associated with a level of confidence for which a given label can be assessed.

5 Results and Discussion

In this section, we compare the classifiers using the soft performance evaluation metrics described in 4.3, contrast results between simple and complex classification problems, and make a special comparison to the EM prior classifier.

5.1 Comparing Classifiers

We present graphs for each classifier displaying the accuracy as a function of P_G, as defined in 3.1 . Consider Figure 1. There are two lines for each classifier: one for the classifier trained on the mode of the uncertain label set Y (dotted line) and the other trained on the mode of the uncertain label set with informed labels Y'(solid line). All accuracies on a scale of [0,1] reported in the y-axis are with respect to the known golden label set for the data. The classifier curve for this figure is a polynomial fit of degree three on the accuracy data. The method of generating classifier curves and coloring shown in Figure 1 is maintained for the remaining graphs in this paper. Costs of each classifier are shown in Table 2. EM Prior is considered separately in a later section, but is included in the Table 2.

Metric		CART	SVM	Logistic Regression	EM Prior
Runtime (ms)		2.600	3292.413	5.017	1065.173
Transparency		Excellent	Average	Excellent	Poor
Confidence	Probabalistic	No	No	Yes	Yes
	Interpretability	Excellent	Poor	Excellent	Excellent

Table 2: Cost analysis of the classification methods used in the experiments

In Figure 1, it is clear that informed labels do not significantly improve classifier performance on simple classification problems. Note that the breast, iris, and wine datasets are simpler classification tasks with only 2-3 possible classes. We expect that the best achievable accuracy will be similar for these three datasets regardless of classification method and that the resulting classification curves will remain close to one another. The plot of all classifier curves in for each dataset in Figure 2 shows tightly packed, nearly indistinguishable curves for these three datasets . However, the curves for the glass, e. coli, and yeast datasets maintain separation with the addition of informed labels, indicating a clear accuracy boost from these labels. Because of this distinction in behavior between simple and complex classification problems, we continue the comparison of these datasets by splitting them into simple (breast, iris, wine) and complex (glass, e. coli, yeast) categories for the remainder of this discussion.

Significance Tests Table 3 shows the result of adding a single informed label to an uncertain label set for varying numbers of R annotators. In order to asses

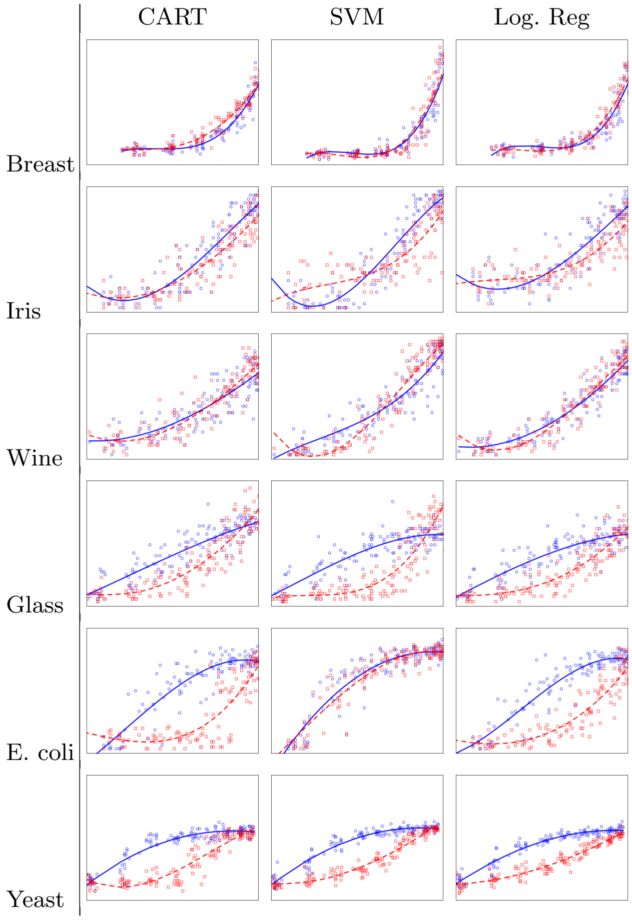

Fig. 1: Graphs of accuracy with respect to varying levels of P_G for all UCI datasets and classification methods for $R = 3$. The dotted line is without informed labels and the solid line is with informed labels.

the effect of informed labels applied to various levels of P_G, we "bin" the values P_G into three groups: $[0, 0.5)$, $[0.5, 0.7)$, and $[0.7, 1]$. This evenly breaks up each graph in Figure 1 into three groups for performing a t-test amongst a similar range of data. Each cell is the p-value of a t-test for the difference of means between each group of classifiers trained on Y and the corresponding Y' within the same bin. Let μ_n be the accuracy of the classifier trained on Y and μ_i be the accuracy of the classifier trained on the Y'. We are testing $H_o : \mu_n = \mu_i$ and $H_a : \mu_n < \mu_i$. We perform 72 tests here, so we use a Bonferroni correction and set $\alpha = 0.05/72 \approx 0.0007$.

In Table 3, for complex problems with two and three annotators, we see a significant increase in accuracy across the values for CART trees and logistic

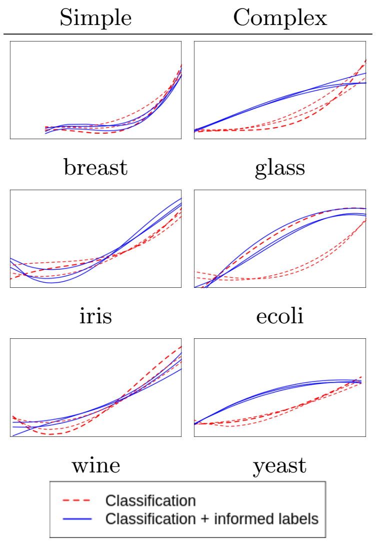

Fig. 2: Plot of fitted curves for all classifiers and all datasets where $R = 3$.

R	Classifier	Simple			Complex		
		[0, 0.5)	[0.5,0.7)	[0.7,1.0]	[0, 0.5)	[0.5,0.7)	[0.7,1.0]
2	CART	1.000	0.722	0.0000	0.000	0.0000	0.000
	SVM	1.000	0.745	0.0000	0.0215	0.0000	0.0000
	LR	1.000	0.150	0.0000	0.0000	0.0000	0.0000
	EM	0.819	0.448	5.94e-02	7.21e-01	9.78e-01	9.97e-01
3	CART	0.376	0.452	0.287	0.0000	0.0000	0.0000
	SVM	0.698	0.170	3.29e-01	0.0000	0.0000	7.17e-03
	LR	0.587	0.222	2.34e-01	**0.0000**	**0.0000**	**0.0000**
	EM	0.977	0.144	1.59e-01	2.97e-01	7.34e-01	1.000
4	CART				1.73e-01	0.000	0.000
	SVM				9.54e-01	2.71e-02	0.0000
	LR				3.07e-01	0.0000	0.0000
	EM				2.57e-01	3.81e-01	1.000

Table 3: p-values of t-test for difference in means of classifiers trained on Y and Y'

regression. Support vector machines also follow this trend with a few exceptions. We believe that the optimization of the SVM parameters narrowed the window of improvement for the informed labels.

When $R = 4$ there are fewer improvements. For all tests we train the models on the mode of Y or Y'. However, we add only one informed label per uncertain set. We expect that if a larger number of informed labels were added to $R = 4$ we would see performance similar to that of $R = 2$ and $R = 3$.

5.2 Comparison to Weak Supervised Classifier: EM Analysis

We specially consider EM prior because it was developed specifically for noisy label set scenarios and is a correlation-based weak supervised classifier. Across all instances, the addition of informed labels does not improve the performance of the EM classifier. The graphs in Figure 3, which display the result of the EM classifier with and without informed labels, confirm this notion. However, because the EM models defined by Jin et al. were designed specifically for noisy label sets where the gold standard is guaranteed, or $P_G = 1$, we expect that our classifiers will achieve similar accuracies as the EM prior for low values of P_G.

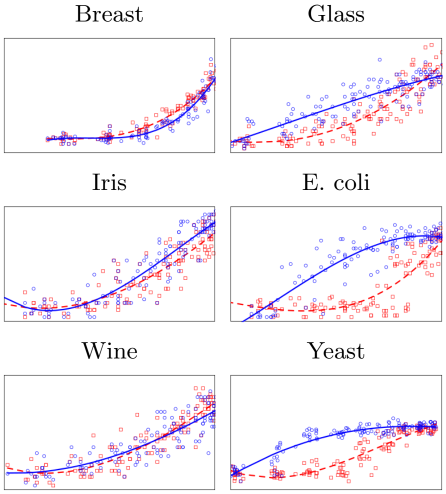

Fig. 3: Graphs of accuracy with respect to varying levels of P_G on EM prior classification method for all UCI datasets and for $R = 3$. Axes, distinguishing colors/lines/shapes, and scaling are all the same as in Figure 1.

Using our method, simple classification models such as CART trees and logistic regression are able to achieve an similar accuracy as EM prior in label sets with high levels of uncertainty. This is outlined in Table 4 and Figure 4, where it is clearly visible that a classifier trained on Y jumps to meet the EM prior curve when trained on Y'. As discussed previously, Table 4 and Figure 4 perform comparisons related to EM Prior only for the complex classification problems. According to the classification costs in Table 2, CART trees and logistic regression models are *considerably cheaper* than EM prior but *maintain comparable accuracy rates* when trained on Y'.

In Table 4, we perform a t-test for the difference of means between the accuracy of the EM prior algorithm and other classification models trained on Y' where $R = 3$. The binning for these tests is identical to the binning performed in Table 3. Let μ_{EM} be the accuracy of the EM prior model trained on Y and μ_n

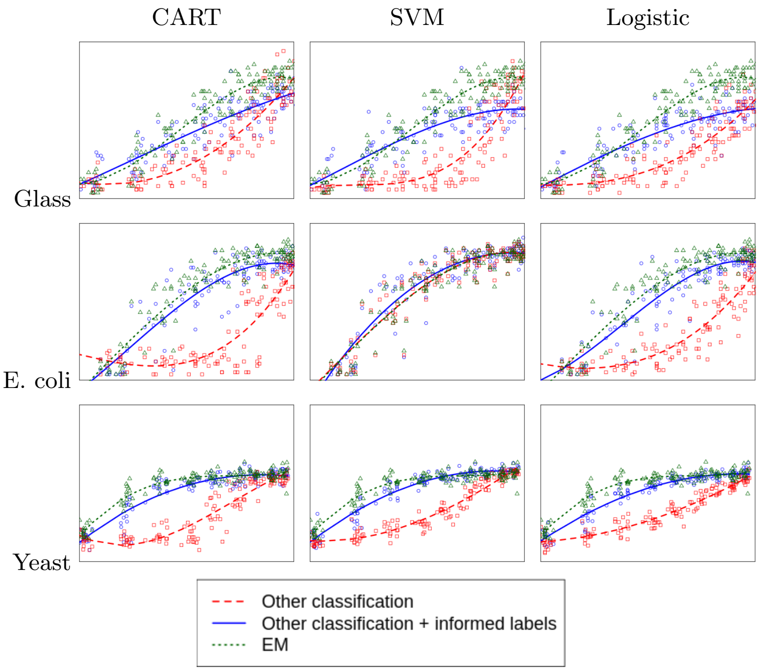

Fig. 4: Graphs of classifier performance compared to EM prior performance for complex datasets.

be the accuracy of the model trained on Y'. Then, each cell represents a p-value for $H_o : \mu_{EM} = \mu_n$ and $H_a : \mu_{EM} > \mu_n$. Table 4 shows that for the lower range of values $P_G = [0, 0.5)$ our method paired with a simple classification method, such as decision trees, achieves the same accuracy as the EM prior model on complex classification tasks. Furthermore, Figure 4 shows the effect of our informed noise on various classifiers compared to the EM model.

Classifier	[0,0.5)	[0.5,0.7)	[0.7,1.0]
CART	0.0933	1.48e-04	0.0000
SVM	0.3040	1.11e-01	0.0000
Logistic	0.0981	0.0000	0.0000
EM Prior	0.7030	0.266e-1	1.46e-04

Table 4: p-values of t-test for difference in means of EM prior model and classifiers trained on Y'

6 Conclusion

In complex classification tasks with three or more classes, we find that our method of using informed labels in a uncertain label set *significantly improves* classifier accuracy across all levels of gold inclusion and annotator numbers for CART and logistic regression classifiers. For SVM we see a similar improvement in performance, but this improvement is not unanimous.

In the case of EM prior we demonstrate that informed labels allow other, simpler classifiers to perform just as well as the EM algorithm in label sets with low values of P_G. Therefore, we assert that by using informed labels, cheap and accessible machine learning models are able to achieve at least the same, if not better, accuracy as their costly but high performing counterparts for scenarios with highly uncertain labels.

We acknowledge the need to experiment with adding more than one informed label to a set of labels with a large number of R annotators. It would be interesting to explore this relationship between R and the number of informed labels added to a set. In addition, there is much potential for probabilistic classifiers to be used in deriving these informative labels.

References

1. Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., et al.: The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. Medical physics **38**(2), 915–931 (2011)
2. Bhavsar, H., Ganatra, A.: A comparative study of training algorithms for supervised machine learning. International Journal of Soft Computing and Engineering (IJSCE) **2**(4), 2231–2307 (2012)
3. Brodley, C.E., Friedl, M.A.: Identifying mislabeled training data. Journal of artificial intelligence research **11**, 131–167 (1999)
4. Ceci, M., Pio, G., Kuzmanovski, V., Džeroski, S.: Semi-supervised multi-view learning for gene network reconstruction. PloS one **10**(12), e0144031 (2015)
5. Dawid, A.P., Skene, A.M.: Maximum likelihood estimation of observer error-rates using the em algorithm. Applied statistics pp. 20–28 (1979)
6. Dheeru, D., Karra Taniskidou, E.: UCI machine learning repository (2017), http://archive.ics.uci.edu/ml
7. Jin, R., Ghahramani, Z.: Learning with multiple labels. In: Advances in neural information processing systems. pp. 921–928 (2003)
8. Kang, F., Jin, R., Sukthankar, R.: Correlated label propagation with application to multi-label learning. In: null. pp. 1719–1726. IEEE (2006)
9. Kotsiantis, S.B., Zaharakis, I., Pintelas, P.: Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering **160**, 3–24 (2007)
10. Lipton, Z.C.: The mythos of model interpretability. arXiv preprint arXiv:1606.03490 (2016)

11. Northcutt, C.G., Wu, T., Chuang, I.L.: Learning with confident examples: Rank pruning for robust classification with noisy labels. arXiv preprint arXiv:1705.01936 (2017)
12. Ratner, A.J., De Sa, C.M., Wu, S., Selsam, D., Ré, C.: Data programming: Creating large training sets, quickly. In: Advances in neural information processing systems. pp. 3567–3575 (2016)
13. Raykar, V.C., Yu, S., Zhao, L.H., Valadez, G.H., Florin, C., Bogoni, L., Moy, L.: Learning from crowds. Journal of Machine Learning Research **11**(Apr), 1297–1322 (2010)
14. Rodrigues, F., Pereira, F., Ribeiro, B.: Learning from multiple annotators: distinguishing good from random labelers. Pattern Recognition Letters **34**(12), 1428–1436 (2013)
15. Sheng, V.S., Provost, F., Ipeirotis, P.G.: Get another label? improving data quality and data mining using multiple, noisy labelers. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 614–622. ACM (2008)
16. Smyth, P., Fayyad, U.M., Burl, M.C., Perona, P., Baldi, P.: Inferring ground truth from subjective labelling of venus images. In: Advances in neural information processing systems. pp. 1085–1092 (1995)
17. Vaswani, S., Ahmed, M.O.: Learning from multiple annotators: A survey
18. Wang, F., Zhang, C.: Label propagation through linear neighborhoods. IEEE Transactions on Knowledge and Data Engineering **20**(1), 55–67 (2008)
19. Yan, Y., Rosales, R., Fung, G., Dy, J.G.: Active learning from crowds. In: ICML. vol. 11, pp. 1161–1168 (2011)
20. Zhang, M.L., Yu, F.: Solving the partial label learning problem: An instance-based approach. In: IJCAI. pp. 4048–4054 (2015)
21. Zhang, M.L., Yu, F., Tang, C.Z.: Disambiguation-free partial label learning. IEEE Transactions on Knowledge and Data Engineering **29**(10), 2155–2167 (2017)
22. Zhang, M.L., Zhou, B.B., Liu, X.Y.: Partial label learning via feature-aware disambiguation. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1335–1344. ACM (2016)
23. Zhu, X., Ghahramani, Z.: Learning from labeled and unlabeled data with label propagation (2002)