Active Learning of Classification Models with Likert-Scale Feedback

Yanbing Xue Milos Hauskrecht
Department of Computer Science, University of Pittsburgh, Pittsburgh, PA
{yanbing, milos}@cs.pitt.edu

Abstract

Annotation of classification data by humans can be a time-consuming and tedious process. Finding ways of reducing the annotation effort is critical for building the classification models in practice and for applying them to a variety of classification tasks. In this paper, we develop a new active learning framework that combines two strategies to reduce the annotation effort. First, it relies on label uncertainty information obtained from the human in terms of the Likert-scale feedback. Second, it uses active learning to annotate examples with the greatest expected change. We propose a Bayesian approach to calculate the expectation and an incremental SVM solver to reduce the time complexity of the solvers. We show the combination of our active learning strategy and the Likert-scale feedback can learn classification models more rapidly and with a smaller number of labeled instances than methods that rely on either Likert-scale labels or active learning alone.

1 Introduction

Classification problems are part of our everyday life. While tremendous progress has been made in the development of methods for building classification models from data, the problem of data annotation and its costs may still hinder their construction and consequently prevent their wider deployment. The objective of this work is to study ways of reducing the data annotation effort so that high-quality classification models can be built. Our focus throughout the paper is on the construction of binary classification models.

In this work, we explore two strategies to alleviate the annotation effort and their combination. First, we study solutions for incorporating a more refined feedback on the class label, and its confidence humans may provide when making label assessment. Second, we study and develop an active learning framework that utilizes this information and selects examples most promising for classification model refinement.

Traditional classification model learning methods assume the feedback provided by humans is restricted to the class information. However, in practice, especially when the data object classification is not straightforward humans can differentiate among examples that are clear, weaker or marginal representatives of a class. It is this type information we seek to collect and incorporate into the model building process. In terms of human feedback, this information can be obtained and expressed in various forms, for example, various numerical or probabilistic scores [1, 2, 3]. In this paper, we assume the feedback expressed in terms of Likert-scale categories [4]. Briefly, Likert-scale categories define a set of ordinal categories humans can use to provide information about the strength of agreement (or belief) in the respective class labels. For example, when obtaining a feedback from a physician on whether the patient suffers from a particular disease or not, the binary true/false feedback can be refined by obtaining physician's belief in the presence of the disease on a 5-point Likert scale by asking if he/she agrees, weakly agrees, is neutral, weakly disagrees, or disagrees with the disease. We develop methods based on ordinal regression [5] and ranking [6, 7] to learn the classification model from such information and demonstrate its benefits over feedback that is based only on class label information.

Another widely used approach to alleviate the data annotation problem is active learning [8, 9, 10] . In active learning, the learner selects the examples that appear to be most promising for the refinement of the classification model and asks the user to provide their label. Many different active learning strategies have been developed to address the problem when information about the class label is queried. In this work, we develop a new active learning strategy that attempts to optimize the example selection by considering the Likert-scale feedback. Our example-selection strategy seeks the example with the greatest expected model change and relies on the Bayesian estimates to calculate the expectations.

We test our new framework on multiple classification problems based on UCI and real-world clinical decision problem data. We demonstrate the ability of our solutions to reduce the data labeling cost both individually and in combination.

2 Related Work

In this section, we briefly review the work related to our approach. We start by discussing the existing work in which label uncertainty information is utilized to make the learning of classification models more sample-effective. After that, we review related work on active learning.

2.1 Learning with soft-label information The problem of learning binary classification models from auxiliary soft label information is relatively new was first explored by [1, 2, 3]. This line of work assumes the label information can be supplemented by a probability with which the annotator believed the class label occurs. The human feedback was provided in terms of the probability estimate. To exploit the probabilistic feedback the authors first studied and developed multiple approaches for learning classification models from probabilistic labels via regression. They showed that these approaches may not be the best and may not learn a good model when soft labels are subject to noise, which is very likely when probabilities are based on human assessments [11, 12, 13]. To deal with this problem, they proposed a more robust method that ignores numerical differences in probability estimates and replaces them with the pairwise ordering of data points in the training data. The method learns a parametric discriminative model that maximizes the satisfaction of all these order constraints. An advantage of the approach is that it builds a classification model one can easily apply to classify future data. Its limitation is that the number of constraints the model tries to satisfy is quadratic in the number of training data points. In addition to soft-label learning framework by [1, 2, 3] where soft labels come from just one annotator, [14] explored a framework where soft labels are derived from multiple annotators with potential disagreements. This work, however, comes with multiple limitations. First, it does not or consider any soft label noise. Second, it does not give a clear approach how the soft label information should be used to learn a better classification model. Finally, it lacks empirical support to demonstrate the impact on model learning. More recently, [15, 16] proposed a new non-parametric algorithm for learning the classification model from probabilistic estimates. The approach works by predicting the probability associated with binary classes based on the Gaussian process regression. Briefly, the method defines the mean function of the Gaussian process to be 0.5 and the covariance function using the Radial basis kernel. The model lets one predict the probability p_i for any new point x_i by calculating the posterior distribution of the Gaussian process. The limitations of the approach are the design of the covariance function (restricted to the radial basis functions), and a non-parametric nature of the model when it is applied to prediction tasks. Overall, while all of the above methodologies showed the benefit of additional probabilistic information in learning classification models more efficiently, they also dealt with the problem of noise in the numerical probabilistic estimates based on human feedback. While in this work we build upon the existing work, we are different in that we assume the feedback on label uncertainty information is limited to a relatively small number of ordinal Likert-scale categories. This type of label uncertainty feedback is more realistic, and it is also less likely to suffer from the noisy estimates. Indeed the literature on the design of user studies recommends the number of Likert scale categories not to exceed nine, with five or seven being a typical design choice. We propose and develop efficient classification algorithms for learning the binary classification models for the Likert scale class label feedback.

2.2 Active learning Active learning integrates model learning and data instance annotation processes into one learning framework. Active learning sequentially labels an initially unlabeled set of examples and chooses which example should be labeled next. By selecting the examples most informative for building the classification model it can minimize the cost of labeling (number of examples that need to be labeled) necessary for building a good model. Different strategies to define the "informativeness" of examples have been proposed in the literature. *Uncertainty sampling* [8] is the simplest and most widely used strategy. It queries the example that the current model predicts with the lowest class confidence. Another popular and often very well performing strategy is *query-by-committee* strategy [17] that picks the example to be queried with the help of multiple predictors, by finding the example these predictors disagree on the most. The construction of the Committee may be done in many different ways. For example, *query-by-bagging* [18] repeatedly samples subsets of labeled instances (using bagging [19]) and trains committee models on them. In terms of the size of the committee, previous works have shown that even a small committee (size two or three) could work well in practice [9, 17, 20]. The *expected model change* strategy [21, 22] queries examples which cause the largest change to the current model if we knew their labels. The disadvantage of the strategy is that "informativeness" can be overestimated. Other representative strategies are *expected error reduction* [10] and *variance reduction* [23]. The first one seeks an example that would let it reduce the generalization error of the model. The second one seeks an example that would minimize the prediction variance of the current model the most.

Our goal is to develop an effective active learning approach that works well with ordinal categories reflecting Likert-scale. Attempts to integrate active learning strategies with such information have been very limited so far. Moreover, many of the above query strategies developed for binary class labels do not work in settings of Likert-scale labels. For example, uncertainty sampling relies on the confidence of the predictions. This strategy works well for labels without confidence information, say, binary labels. However, it does not make much sense if confidence may be confirmed by the feedback. In this work, we develop a new instance selection strategy that collaborates with ordinal category Likert-scale feedback and picks the instance with the largest expected impact (change) on the model.

3 Methodology

In this part, we develop an active learning framework that builds a classification model by actively querying an annotator who provides feedback to the framework for assessing the instances using Likert-scale categories. We start by first defining and formalizing the problem of learning from ordinal Likert-scale category labels. After that, we present an algorithm for learning the classification model from such feedback. Second, we show how this algorithm can be included in the active learning framework that aims to improve the model by wisely selecting the examples to be assessed next. The criterion used to choose from among unlabeled candidate instances is based on the highest expected classifier prediction change. We also briefly describe solutions to two partial problems: (1) how to model the distribution which is used to calculate the expected change, and (2) how to speed up training via incremental solver when adding one unlabeled sample.

3.1 Problem settings Our objective is to learn from data a binary classifier: $C : X \rightarrow Y$, where X is a feature space and $Y \in \{0, 1\}$ is one of the two class labels. At the very beginning, all the examples in dataset D are unlabeled. However, we can sequentially query a human annotator to provide information for individual examples and use this feedback to build a classification model. We assume that in addition to traditional binary labels $Y = \{0, 1\}$, each data example is also assessed in terms of ordinal Likert-scale categories characterizing the degree of agreement of the annotator in its assignment to one of the classes. Therefore, a labeled data sample d_i is a vector consisting of three parts $(\mathbf{x}_i, y_i, u_i)$, that is, a vector of features, a traditional binary label and a Likert-scale label indicating the level of agreement that the data example falls into one of the two classes. Both y_i and u_i are based on human annotator feedback. For example, if a human expert is asked to assess a patient whether he or she suffers from a particular disease, $\mathbf{x}$ represent the labs, symptoms and observations describing the patient state, y is expert's disease/no-disease decision, and u represents the degree of which the expert believes in (and agree) with the disease diagnosis.

3.2 Learning classifier from Likert-scale labels Let us focus first on the task of learning a classification model from the data represented by the triplets $(\mathbf{x}_i, y_i, u_i)$, that is, we assume the data with this information are available and can be used. One way to learn the classification model would be to adapt and build upon approach proposed by [1, 2] for probabilistic soft-label feedback with noise. Briefly, their approach seeks to find a ranking function $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}_i$ that aims to satisfy all pairwise constraints among data points ordered according to their 'noisy' probability estimate reflecting the confidence of the annotator in the binary class label. They formulate and solve the problem using an SVM-like optimization task that seeks to satisfy as many constraints as possible. The key trick in their approach is that the same ranking function can also be used to define a discriminative projection that lets us discriminate between class 0 and class 1 data instances. We can quickly adapt their approach and apply it to Likert-scale assessments by creating pairwise ordering constraints only among data points that fall into the different Likert-scale categories. Briefly if two data entries $\mathbf{x}_i$ and $\mathbf{x}_j$ in the dataset are assigned ordinal category labels such that $u_i > u_j$, we expect that the same order will be preserved also by the the ranking function: $f(\mathbf{x}_i) > f(\mathbf{x}_j)$. Similarly to [1, 2], a classifier, and its discriminative projection can be then defined using the same ranking function.

Unfortunately, the above solution suffers from a drawback: the number of pairwise constraints one wants to satisfy grows quadratically with the number of data instances which negatively affects the time-complexity and scalability of the solution. To alleviate the scalability problem, we try to abridge the number of constraints imposed on the ordinal Likert-scale labels. In this paper, we propose an improvement based on 'binning' of values of the ranking function f. The idea of the solution is that after the projection (via ranking function), all examples with the same ordinal category label should, in the ideal case, fall into the same value region or bin.

Let us assume that for each ordinal label u we have a bin defined by a lower bound value b_{u-1} and an upper bound value b_u. Our objective is to find a projection f from the feature space to the space of real numbers, for which instances that are in the same ordinal category fall after the projection into the same bin. More formally, for any data instance $\mathbf{x}_i$ and its ordinal label u_i, we expect to obtain a function $f(\cdot)$ so that $bin(f(\mathbf{x}_i)) = u_i$, where $bin(\cdot)$ is a function where the argument is a prediction value and the return value is the bin where this prediction value belongs. Then each data example with ordinal label u should be projected such that its value is greater than all bin bounds b_j such that $j < u$ and less than all b_k such that $k \geq u$. Since the ordinal label and the projection of its feature vector to the bin are always expected to match, the projection should have the same greater-or-less relationship with all bin boundaries for other ordinal categories. Formally, for a data instance $\mathbf{x}$ and Likert-scale label u, we expect to learn a function so that

$f(\cdot)$ so that $b_j < f(\mathbf{x})$ for any $j < u$ and $b_k > f(\mathbf{x})$ for any $k \geq u$.

However, in reality, we cannot expect that all the constraints will always be satisfied with a linear projection function. Hence, we permit violations of constraints but penalize them via bin-sample loss function. By adding the constraints for standard binary class labels, we can formulate the following optimization problem:

$$\begin{gathered}
\min_{\mathbf{w}, w_0, \mathbf{b}, \eta, \xi} \frac{\mathbf{w}^T \mathbf{w}}{2} + B \sum_{i=1}^{N} \eta_i + C \sum_{j=1}^{m-1} \sum_{i=1}^{N} \xi_{j,i} \\
y_i(\mathbf{w}^T \mathbf{x}_i + w_0) \geq 1 - \eta_i \qquad \forall i \\
z_{j,i}(\mathbf{w}^T \mathbf{x}_i - b_j) \geq 1 - \xi_{j,i} \qquad \forall i, j \\
\eta_i, \xi_{j,i} \geq 0 \qquad \forall i, j
\end{gathered}$$

where $j = 1, 2, \ldots, m-1$ indexes bin bounds in $\mathbf{b}$, and $i = 1, 2, \ldots, N$ indexes data entries. The first term in the objective function is the regularization term, the second term (single sum) defines the hinge loss on binary labels, and the third term (double sum) defines the loss function between each pair of bin bound and each Likert-scale label. η_i and $\xi_{j,i}$ are slack variables permitting violations of binary class and soft-label bins respectively. B and C are constants weighing the objective function terms. $z_{j,i}$ is an indicator whether the projection of feature vector $\mathbf{x}_i$ is supposed to be greater or less than the bin bound b_j. If $j < u_i$, indicating the projection of $\mathbf{x}_i$ is supposed to be greater than b_j, $z_{j,i} = 1$, otherwise $z_{j,i} = -1$. In this model, the number of constraints is reduced to roughly $M = mN$. Since Likert-scales typically comes as from 2 to 10 ordering categories with 5 or 7 being the most common, we have $m << N$. Considering the $O(M^3)$ complexity of convex quadratic optimization problems, the time complexity is reduced to $O(m^3 N^3)$.

Removing empty bins One practical concern related to the above optimization problem occurs when the size $|L|$ of the labeled data is small, and some Likert-scale categories are absent in L. Fortunately, this problem has an easy fix. If a Likert-scale category is missing in L, it is not necessary to consider it, and we should only try to enforce ordering constraints among non-empty ordinal categories. Effectively this translates to a smaller number of bins and their boundaries in the optimization problem.

3.3 Active learning The next challenge is to embed the above learning algorithm in a practical active learning framework. The heart of any active learning method is a strategy that is used to select examples to be queried next. In this work, we propose and experiment with a strategy called expected performance change (EPC) that evaluates and measures the potential of an unlabeled data instance to change the model by estimating its impact on instance predictions. Expected performance change (EPC) is closely related to expected model change (EMC) proposed by [24, 25] where the potential of an unlabeled data instance is estimated from the change in the model weights after it is assumed to be labeled. Another difference is that our expected performance change (EPC) uses a Bayesian posterior to calculate the expectation.

Expected performance change Briefly, the Expected Performance Change (EPC) of an unlabeled sample $\mathbf{x}$ can be measured as follows: Suppose that, for the labeled data L, we have already trained a model f_L. For $\mathbf{x}$, there are m possible Likert-scale labels (m is the number of Likert-scale ordinal categories). For each possible Likert-scale label u, if we add $\langle \mathbf{x}, u \rangle$ into L, we will obtain an add-one model $f_{L \cup \langle \mathbf{x}, u \rangle}$. The performance change of $f_{L \cup \langle \mathbf{x}, u \rangle}$ compared with f_L is denoted as $\delta(\mathbf{x}, u)$. Since there are m possible Likert-scale labels, we will have m performance changes $\delta(\mathbf{x}, u)$ where $u = 1, 2, \ldots, m$, each corresponding to one add-one model. The Expected Performance Change $\Delta(\mathbf{x})$ of $\mathbf{x}$ is then calculated as:

$$\Delta(\mathbf{x}) = \sum_{u=1}^{m} p(u|\mathbf{x}) \delta(\mathbf{x}, u)$$

Measuring performance change One critical question of the Expected Performance Change framework is, how to measure the performance change $\delta(\mathbf{x}, u)$ for an unlabeled example $\mathbf{x}$ and one possible label u the example can be assigned to. In this work, we adopt the measurement based on the discrepancy of the predictions over unlabeled data for cases before and after $\mathbf{x}$ and u are added into L and used to learn a new model. More formally, this measurement is calculated as follows: Let the model for L be f_L, and the model after $\langle \mathbf{x}, u \rangle$ is added to L be $f_{L \cup \langle \mathbf{x}, u \rangle}$. For any unlabeled sample $\mathbf{x}_i$, we measure the performance change as the discrepancy of the bin predictions $||bin(f_L(\mathbf{x}_i)) - bin(f_{L \cup \langle \mathbf{x}, u \rangle}(\mathbf{x}_i))||$. By considering every unlabeled example, the net performance change $\delta(\mathbf{x}, u)$ can be calculated by averaging its impact on all unlabeled data as:

$$\delta(\mathbf{x}, u) = \sum_{i \in U} ||bin(f_L(\mathbf{x}_i)) - bin(f_{L \cup \langle \mathbf{x}, u \rangle}(\mathbf{x}_i))||$$

where i indexes all examples in the unlabeled dataset U.

Approximating the expectation After calculating performance changes $\delta(\mathbf{x}, u)$ for all possible ordinal labels u, one important question is how to calculate the expectation needed for the expected performance change score. In this work, we adopt a Bayesian method to estimate the expectation.

Our calculation is based on the model f_L learned from the labeled set L of data instances. Briefly, a model f_L together with its bin boundaries defines a model for all ordinal categories. We can use this model and its bins to estimate the empirical distribution of labeled examples in these bins. More specifically, each bin that is associated with the projection f_L may receive (labeled) examples from all categories (that is, even categories that do not match the category corresponding to the bin). Assuming there are m categories, in general, each bin may see examples from m different categories. We can use the observed counts of the examples with these categories that fall into the same bin to calculate the necessary expectations for an unlabeled data point $\mathbf{x}$. Briefly, we take an unlabeled data point and use the projection f_L to identify the bin it falls into. The count of category labels for this bin is then used to approximate their probability distribution and hence calculate the expected value.

More formally, let $\mathbf{x}$ be an unlabeled instance we are considering to query and $j = bin(f_L(\mathbf{x}))$ be the bin category the instance falls into based on f_L. Our objective is to estimate the probability distribution $(p_1^j, p_2^j, \ldots, p_m^j)$ for bin j, which represents the probability of an example in bin j to be assigned to one of the m Likert-scale categories. One way to estimate this probability would be to use the maximum likelihood approach and calculate the probabilities from counts of Likert-scale labels $q_1^j, q_2^j, \ldots, q_m^j$ in L that fall into bin j, and by assuming they follow a multinomial distribution with parameters $(p_1^j, p_2^j, \ldots, p_m^j)$. However, this estimate may not work well if the number of labeled examples is small which would lead to a biased estimate. Hence, instead of the maximum likelihood based estimate, we base our estimate on the posterior distribution.

To estimate the posterior distribution of $(p_1^j, p_2^j, \ldots, p_m^j)$ we use a Dirichlet prior which is the conjugate choice for the multinomial sampling distribution. Since we do not have any prior information about the distribution of categories in the bin; we choose $Dirichlet(1, 1, \ldots, 1)$ where all Likert-scale categories are assigned the same prior probability. Given the conjugate prior, the posterior of $(p_1^j, p_2^j, \ldots, p_m^j)$ for L follows a Dirichlet distribution:

$$(p_1^j, p_2^j, \ldots, p_m^j)_L \sim Dirichlet(1+q_1^j, 1+q_2^j, \ldots, 1+q_m^j).$$

Given the posterior distribution, we can approximate the probability $p(u|\mathbf{x})$, that is, the probability that $\mathbf{x}$ is assigned label u, by the expected value of $E(p_u^j)$ from the posterior distribution:

$$E(p_u^j) = \frac{(1+q_u^j)}{(m + \sum_{i=1}^m q_i^j)}.$$

Substituting the result, we can finally calculate the Expected Performance Change for an unlabeled sample $\mathbf{x}$ as:

$$\begin{aligned} \Delta(\mathbf{x}) &= \sum_{u=1}^m p(u|\mathbf{x})\delta(\mathbf{x}, u) \\ &= \sum_{u=1}^m \frac{(1+q_u^j)}{(m + \sum_{i=1}^m q_i^j)}\delta(\mathbf{x}, u) \end{aligned}$$

Counting to preserve ordering information One concern of adopting a multinomial distribution to model the data is that all categories in the multinomial model are assumed to be independent. However, our approach uses Likert-scale categories, which are ordinal categories. One way to modify the multinomial model to reflect such dependencies is to use partial counts and let categories close to the category assigned for example $\mathbf{x}_i$ take partial credit for it. To implement this idea we modify the counts $q_1^j, q_2^j, \ldots, q_m^j$ associated bin j as follows: if an observed example $\mathbf{x}_i$ that falls into bin j is assigned a Likert-scale label u_i then it contributes 1 to the count $q_{u_i}^j$ and 0.5 to the counts $q_{u_i-1}^j$ and $q_{u_i+1}^j$ (that is, two Likert-scale categories next to the observed category).

3.4 Training of add-one models

Another critical question is the running time complexity to obtain an add-one model $f_{L\cup\langle\mathbf{x},u\rangle}$ after adding an unlabeled sample $\mathbf{x}$ and possible Likert-scale label u into labeled data L. If we train the add-one model from scratch, the time complexity is $O(m^3|L|^3)$ where m is the number of Likert-scale labels. Since, in order to select the sample to be labeled next, we need to obtain an add-one model for each unlabeled sample and each possible Likert-scale label, the total time complexity is $O(m^4|L|^3|U|)$ (U is the unlabeled data), which is extravagant and does not scale well as the size of L grows. To solve this problem, we develop an incremental solver learning classifiers from ordinal category feedback. This solution extends the incremental SVM solver proposed in [26]. By using the incremental solver when training $f_{L\cup\langle\mathbf{x},u\rangle}$, instead of starting from scratch, we always start from f_L, which remarkably reduces the total time complexity to $O(m^4|L|^2|U|)$.

4 Experiments and Results

We test our approach on both synthetic and real-world data. The first set of experiments uses data from several UCI regression and ordinal classification datasets which we transform to problems with Likert-scale categories. The second experiment works with real-world clinical data with true (human assessed) ordinal categorical labels.

4.1 Experiments on synthetic UCI-based data

In this part, we adapted three UCI regression datasets (Housing,

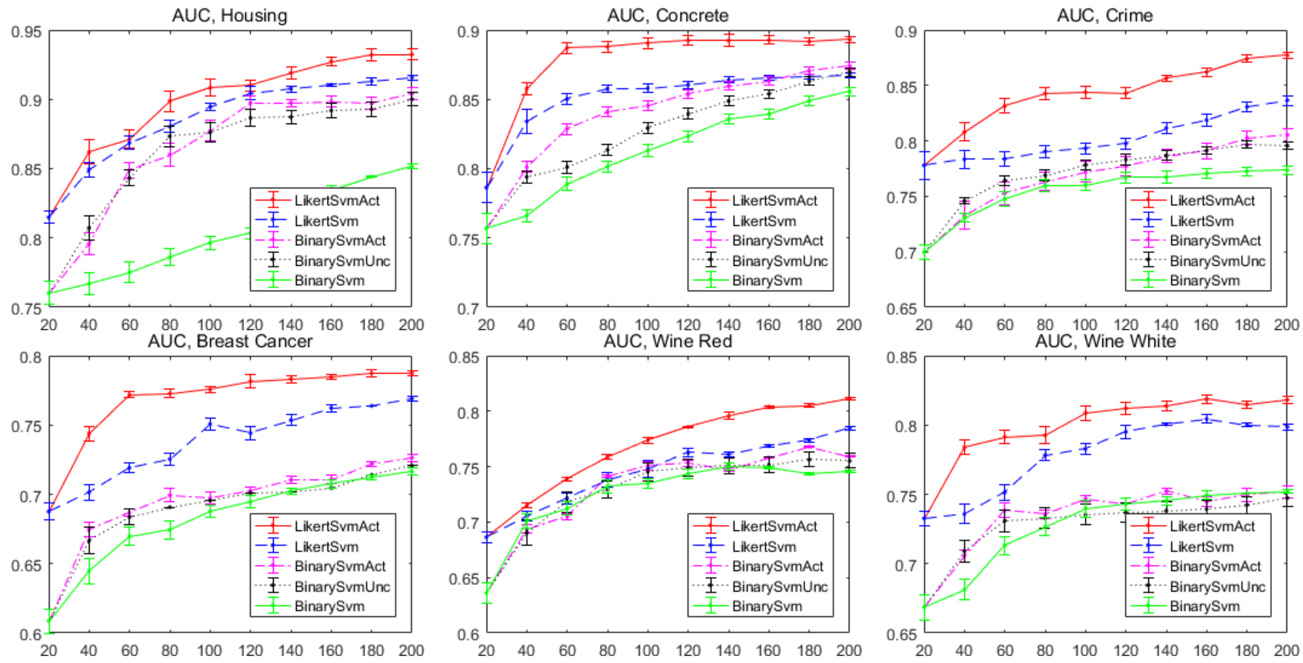

Figure 1: Performance of our active learning framework on six UCI datasets.

Dataset	# Samples	# Features	# Categories
Housing	506	13	Regression
Concrete	1030	9	Regression
Crime	1994	122	Regression
Breast Cancer	699	10	6
Wine Red	1599	12	11
Wine White	4898	12	11

Table 1: Properties of UCI data in synthetic experiments.

Concrete, and Crime) and three UCI ordinal classification datasets (Cancer, Wine Red, and Wine White) that are summarized in Table 1 as follows.

For the regression datasets, we discretized the real-valued outputs into 7 Likert-scale levels by dividing the range of output values into equal length bins. We defined a binary class label by considering the examples that fell into three higher-value bins as representatives of class 1 and examples in four lower-value bins as examples from class 0. For example, in Housing dataset this discretization would represent houses with high attractiveness (class 1), and houses with low attractiveness (class 0) and Likert scales represent different degrees of attractiveness. The UCI ordinal classification datasets come with multiple (ordinal) classes so that they can be used as Likert-scale levels directly. The binary thresholds can be set according to the meaning of these ordinal classes. For example, Breast Cancer dataset contains six ordinal classes $\{1, 2, 3, 4, 5, 6\}$, where $\{1, 2\}$ are healthy, and $\{3, 4, 5, 6\}$ represent the different stages of malignancy, so we map Likert-scale levels 3,4,5,6 to Class 1 and the rest to Class 0.

The objective of our experiments is to demonstrate the benefits of our active learning strategy for models of Likert-scale labels by comparing it to different classification models trained on Likert-scale versus binary labels, and labeling strategies based on the random versus active sampling. Our experiments compare the following models:

BinarySvm: The standard linear SVM with the hinge loss and quadratic regularization factor trained on examples with binary labels only that were sampled randomly.

BinarySvmUnc: The standard linear SVM with the hinge loss and quadratic regularization factor trained on examples with binary labels only, but sampled actively based on the uncertainty sampling selection criterion.

BinarySvmAct: The standard linear SVM with the hinge loss and quadratic regularization factor trained on examples with binary labels only, but sampled actively based on the expected performance change (EPC) selection criterion. To apply the criterion to binary classification settings, we treat class 0 and class 1 as two bins.

LikertSvm: Our SVM-based for Likert-scale labels that enforces both binary and bin-label constraints. Examples to be labeled next, are selected randomly.

LikertSvmAct: A combination of our SVM-based for Likert-scale labels and our Expected Performance Change for selecting examples to be labeled next.

We evaluated the performance of the different methods

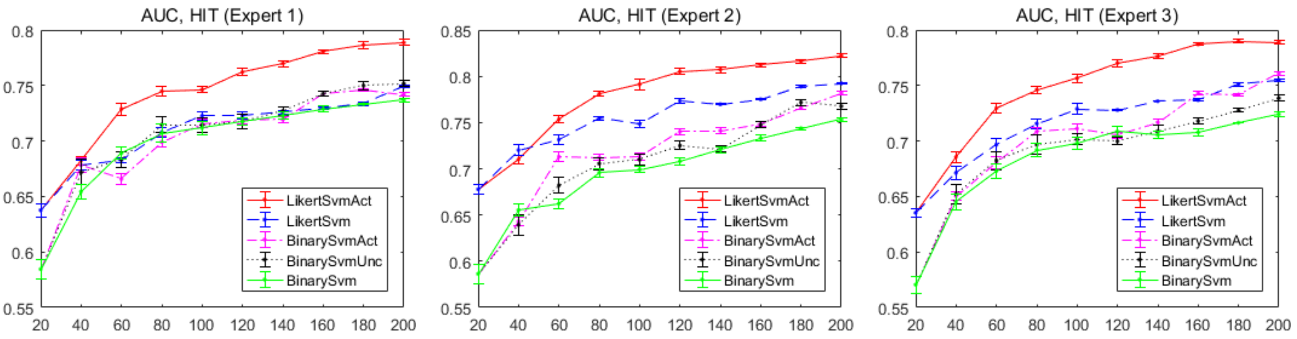

Figure 2: Performance on HIT data annotated by 3 experts.

by calculating the Area under the ROC (AUC) the learned classification model would achieve on the test data. Hence, each dataset before the learning was split into the training and test set (using $\frac{2}{3}$ and $\frac{1}{3}$ of all data entries respectively). The active learning considered training data only; the AUC was always calculated on the test set. The test set performance reflects how well the model generalizes to future data. To avoid potential train/test split biases, we repeated the training process (splitting) and learning steps 24 times. We report the average AUC obtained on these test sets. To test the benefits of our active learning strategy and the impact of Likert-scale label information on the number of data entries, we trace the performance of all models for the different sizes M of labeled data. Figure 1 shows the performance (AUC) of the models on all six UCI datasets for increasing sizes of M.

Figure 1 shows the benefit of LikertSvmAct with a combination of our active learning strategy and soft labels. Both LikertSvmAct and LikertSvm outperform BinarySvmAct, BinarySvmUnc, and BinarySvm, indicating that soft label models will achieve better performance than original binary label models with the same training sizes. LikertSvmAct also outperforms LikertSvm, validating the effectiveness of our querying strategy. Meanwhile, LikertSvmAct greatly outperforms BinarySvm, indicating the combination of active learning and Likert-scale labels clearly raises the performance on the same sizes of training data.

4.2 Experiments on clinical data While the experiments on synthetic datasets appear to support the benefits of our active learning approach based on ordinal Likert-scale labels, it is unclear whether synthetic labels generated for the UCI datasets do not make any unreasonable assumptions and whether good performance also generalizes to "true" feedback provided by humans. In this set of experiments, we test the performance of the methods on a real-world clinical data that were independently reviewed and assessed in terms of soft-labels by three different experts. The target label concerns clinician's agreement with raising an alert on Heparin-induced thrombocytopenia (HIT), an adverse clinical condition that affects the patient who is treated with heparin for prolonged periods of time. The data and features for the experiment were derived from the PCP database of Electronic records of post-cardiac surgical patients [27, 28, 29]. The clinical data consists of 50 patient state features essential for detection of HIT. The datasets consist of 579, 571, and 573 labeled patient state instances for Expert 1, 2 and 3 (see [30]), respectively. The labels include Likert-scale labels on four levels indicating the agreement, weak agreement, weak disagreement, and disagreement of the expert with the HIT alert.

Figure 2 shows the AUC performance of the same methods and models as used in the previous section on three expert-annotated HIT datasets. The performance of LikertSvmAct outperforms LikertSvm, BinarySvmAct, BinarySvm on all three datasets, confirming good performance of our method on synthetic data and the benefit of both the Likert-scale labels and active learning for a more efficient training of binary classification models.

5 Conclusion

In this work, we proposed a new framework for learning binary classification models from human feedback that utilizes a refined human feedback expressed in terms of ordinal Likert-scale categories and novel active learning strategy. Our results on synthetic and real-world clinical data show that our learning framework (1) can learn more efficiently and from a smaller number of examples than existing methods (2) is better than models that rely on Likert-scale labels or active learning individually.

6 Acknowledgement

The work presented in this paper was supported in part by grants R01GM088224 and R01LM010019 from the NIH. The content of the paper is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

[1] Quang Nguyen, Hamed Valizadegan, and Milos Hauskrecht. Learning classification with auxiliary probabilistic information. In *IEEE International Conference on Data Mining*, pages 477–486, 2011.

[2] Quang Nguyen, Hamed Valizadegan, and Milos Hauskrecht. Sample-efficient learning with auxiliary class-label information. In *Proceedings of the Annual American Medical Informatics Association Symposium*, pages 1004–1012, 2011.

[3] Quang Nguyen, Hamed Valizadegan, and Milos Hauskrecht. Learning classification models with soft-label information. *Journal of American Medical Informatics Association*, 2013.

[4] Rensis Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 140:1–55, 1932.

[5] Wei Chu and S. Sathiya Keerthi. New approaches to support vector ordinal regression. In *Proceedings of the 22nd international conference on Machine learning*, ICML '05, pages 145–152, New York, NY, USA, 2005. ACM.

[6] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 133–142, 2002.

[7] Ralf Herbrich, Thore Graepel, and Klaus Obermayer. Support vector learning for ordinal regression. In *International Conference on Artificial Neural Networks*, pages 97–102, 1999.

[8] David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th ACM International Conference on Research and Development in Information Retrieval*, pages 3–12, Dublin, IE, 1994. Springer Verlag, Heidelberg, DE.

[9] Burr Settles. Active learning literature survey. Technical report, 2010.

[10] Nicholas Roy and Andrew McCallum. Toward optimal active learning through sampling estimation of error reduction. In Carla E. Brodley and Andrea Pohoreckyj Danyluk, editors, *Proceedings of the 18th International Conferenceon on Machine Learning*, pages 441–448, Williams College,Williamstown, MA, USA, 2001. Morgan Kaufmann.

[11] Peter Juslin, Henrik Olsson, and Anders Winman. The calibration issue: Theoretical comments on suantak, bolger, and ferrell. *Organizational Behavior and Human Decision Processes*, 73(1):3–26, 1998.

[12] Dale Griffin and Amos Tversky. The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, 24(3):411 – 435, 1992.

[13] A. O'Hagan, C. Buck, A. Daneshkhan, R. Eiser, P. Garthwaite, D. Jenkinson, J. Oakley, and T. Rakow, editors. *Uncertainty judgements Eliciting experts' probabilities*. John Wiley and Sons, 2007.

[14] Christian Thiel. *Classification on Soft Labels Is Robust against Label Noise*, pages 65–73. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[15] Peng Peng and Raymond Chi-Wing Wong. Selective sampling on probabilistic data. In *SIAM International Conference on Data Mining*, pages 28–36, 2014.

[16] Peng Peng, Raymond Chi-Wing Wong, and Phillp S. Yu. Learning on probabilistic labels. In *SIAM International Conference on Data Mining*, pages 307–315, 2014.

[17] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, pages 287–294, Pittsburgh, Pennsylvania, 27–29 July 1992. ACM Press.

[18] Naoki Abe and Hiroshi Mamitsuka. Query learning strategies using boosting and bagging. In *Proceedings of the 15th International Conference on Machine Learning*, ICML '98, pages 1–9, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[19] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, August 1996.

[20] Burr Settles and Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1070–1079, 2008.

[21] Simon Tong and Daphne Koller. Active learning for parameter estimation in bayesian networks. In *Advances in Neural Information Processing Systems*, pages 647–653. MIT Press, 2000.

[22] Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. In *Advances in Neural Information Processing Systems*, pages 1289–1296. MIT Press, 2008.

[23] Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, January 1992.

[24] Burr Settles and Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 1070–1079, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[25] Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. In *Advances in neural information processing systems*, pages 1289–1296, 2008.

[26] T. Poggio and G. Cauwenberghs. Incremental and decremental support vector machine learning. In *Advances in Neural information processing systems*, volume 13, page 409, 2001.

[27] Milos Hauskrecht, Michal Valko, Iyad Batal, Gilles Clermont, Shyam Visweswaram, and Gregory Cooper. Conditional outlier detection for clinical alerting. In *Proceedings of the Annual American Medical Informatics Association Symposium*, pages 286 – 290, 2010.

[28] Milos Hauskrecht, Iyad Batal, Michal Valko, Shyam Visweswaran, Gregory F. Cooper, and Gilles Clermont. Outlier detection for patient monitoring and alerting. *Journal of Biomedical Informatics*, 46(1):47–55, 2013.

[29] Michal Valko and Milos Hauskrecht. Feature importance analysis for patient management decisions. In *Proceedings of the 13th International Congress on Medical Informatics*, pages 861–865, 2010.

[30] Hamed Valizadegan, Quang Nguyen, and Milos Hauskrecht. Learning classification models from multiple experts. *Journal of Biomedical Informatics*, pages 1125–1135, 2013.