Active Learning of Multi-Class Classifiers with Auxiliary Probabilistic Information

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

Abstract

Our ability to learn accurate classification models from data is often limited by the number of available data instances. This limitation is of particular concern when data instances need to be labeled by humans and when the labeling process carries a significant cost. Recent years witnessed increased research interest in developing methods capable of learning models from a smaller number of examples. One such direction is active learning. Another, more recent direction showing a great promise utilizes auxiliary probabilistic information in addition to class labels. However, this direction has been applied and tested only in binary classification settings. In this work we first develop a multi-class variant of the auxiliary probabilistic approach, and after that embed it within an active learning framework, effectively combining two strategies for reducing the dependency of multi-class classification learning on the number of labeled examples. We demonstrate the effectiveness of our new approach on both simulated and real-world datasets.

Introduction

Classification problems are ubiquitous in our everyday life. Machine learning field provides an opportunity to further enrich the spectrum of classification problems one can tackle, enhance their construction, as well as, automate their execution. However, the successful deployment of classification models built by machine learning methods is often limited by the amount of available training data. This is of particular concern when data for classification must be annotated by humans and when annotation process carries a significant time and economical cost. In such a case, the key challenge is to reduce the sample dependency of learning methods as much as possible.

The majority of classification learning methods are limited to only the class-label information. However, the class label decisions the human annotators make are often not straightforward and there may be some uncertainty associated with an instance belonging to one of the classes. This uncertainty information could provide a valuable feedback for training a better classification model. One way to represent such an information is to use probabilistic scores (Nguyen, Valizadegan, and Hauskrecht 2011a; 2011b) , where each data instance is associated with a soft probabilistic label indicating the certainty of human annotators in the given class label, such as, a probability of the patient having a disease. However, it is also well documented that humans are unable to give consistent probabilistic assessments (Juslin, Olsson, and Winman 1998; Griffin and Tversky 1992), leading to noisy probabilistic scores. To solve this problem, (Xue and Hauskrecht 2017) proposed to convert the learning with soft labels to an ordinal regression problem using binning (Chu and Keerthi 2005) and solve it via ranking-SVM (Joachims 2002; Herbrich, Graepel, and Obermayer 1999). The method, however, applies only to binary classification tasks. In this work, we show how to improve and extend this approach to multi-class classification settings. The new method is one of the contributions of this paper.

Another technique to alleviate the annotation effort is active learning (Lewis and Gale 1994; Settles 2010; Roy and McCallum 2001). Briefly, active learning learns from a subset of labeled instances. This set is gradually grown by selecting an unlabeled instance that appears the most informative for refining the current model and by requesting its label. Numerous active learning strategies have been developed. In this work, we develop a new active learning strategy assuming the feedback also includes the auxiliary probabilistic score in addition to class label. Our active learning strategy implements a variant of the expected model change approach. The expected model change approach requires costly recalculation of models every time an instance is considered during the example selection process. We address it by developing its efficient gradient-based approximation.

Through experiments, we show that our new multi-class classification framework achieves improved classification performance and, at the same time, it is able to speed up the selection of instances to be queried next by its active learning component. These results are obtained on both simulated data derived from data in UCI repository and real-world image data. We demonstrate the ability of our active learning and auxiliary label information solutions to reduce the data labeling cost both individually and in combination.

Related Work

In this section, we briefly review the topics related to our framework: soft-label information, multi-class support vec-

Copyright © 2018, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

tor machine, and active learning.

Probabilistic soft-label information The problem of learning from probabilistic soft-label information is relatively new and was initiated by (Nguyen, Valizadegan, and Hauskrecht 2011a; 2011b; 2013). This series of work assumes supplemental probabilistic information can be provided by annotator along with class labels at negligible cost. Such probabilistic information indicates the confidence with which the annotator believed the class label will occur and is expressed using a probabilistic score. To utilize such information, the author first developed several methods to fit the probabilistic scores directly via regression and showed that these methods are highly vulnerable to the noise which is common in probabilistic scores (Juslin, Olsson, and Winman 1998; Griffin and Tversky 1992; O'Hagan et al. 2007). To solve this problem, the authors developed a robust method based on the pairwise orderings among all data examples. Basically, this method learns a parametric discriminative model by attempting to satisfy pairwise score orderings among all data examples while ignoring their exact probabilistic scores. The limitation of the approach is that the number of constraints the orderings induce is quadratic in the number of data examples.

More recently, another robust method was proposed by (Xue and Hauskrecht 2017). The method first splits the range of probabilistic scores into multiple consequent and non-overlapping bins, and after that, it learns an ordinal regression model by attempting to satisfy the constraints from pairwise orderings between each data example and each bin boundary. Similarly to (Nguyen, Valizadegan, and Hauskrecht 2011a), this method ignores exact probabilistic scores to improve its robustness against soft-label noise. However, the method is more efficient: it reduces the number of constraints to be linear in the number of data examples.

All of the above references are focused on binary classification models. However, the problem of learning multi-class classification models from probabilistic soft-label information remains open. In this paper, we attempt to fill this gap. Building upon (Xue and Hauskrecht 2017), we propose a multi-class classifier that learns from auxiliary probabilistic (soft-label) information in addition to their class labels. That is, each data example comes with a class label and a probability indicating the confidence with which the annotator believes the class label indeed occurs. The annotation cost of such a soft-label information is negligible since only one soft label is associated with each data example. We show that our soft-label multi-class classifier substantially reduces annotation effort by achieving higher performance compared with multi-class classifiers based only on class labels.

Multi-class support vector machine Multi-class support vector machine was proposed by (Vapnik 1998; Weston et al. 1999). For a K-classification problem, multi-class support vector machine trains K one-vs-all binary classifiers in one optimization problem. That is, the sum of the regularization term and the penalty for slack variables of all the K one-vs-all classifiers is minimized jointly. Compared with previous methods, one-vs-all, that trains K one-vs-all classifiers and, one-vs-one, that trains $\frac{K(K-1)}{2}$ one-vs-one classifiers independently, multi-class support vector machine achieves higher performance especially when the amount of the labeled data is limited.

The work in this paper builds upon multi-class support vector machine and enhances it to accept also probabilistic score information. This will significantly reduce the annotation effort. We further show how one can design an active learning strategy compatible with such classifier.

Active learning In active learning frameworks, model training and data instance annotation process are conducted alternately. Basically, active learning first labels an initial set of data and sequentially selects the data instances that are most informative to be labeled next. There are numerous criteria to measure the "informativeness" of an unlabeled instance. Perhaps the most famous strategy is *uncertainty sampling* (Lewis and Gale 1994). In multi-class scenarios, three different standards are applied to measure the uncertainty: (1) lowest confidence, that queries the unlabeled instance with the lowest confidence in its highest class prediction, and (2) marginal confidence, that queries the unlabeled instance with the lowest discrepancy in its top-two class predictions, and (3) information entropy, that queries the unlabeled instance with the highest information entropy among all of its class predictions. However, uncertainty sampling is incompatible with soft labels, since such uncertainty is already given in soft labels. Another popular strategy is *query-by-committee* (Seung, Opper, and Sompolinsky 1992) that trains a committee of models and selects an unlabeled instance on which these models disagree the most. The models in the committee can be acquired bootstrapping (Breiman 1996) of the training set. The limitation of the above query-by-committee is the bias since the models in the committee are highly under-fit.

Other more sophisticated instance selection strategies are based on expectation. The *Expected model change* (Tong and Koller 2000) queries unlabeled instance that brings the highest change to the model parameters when labeled. The limitation of this strategy is the overestimate of informativeness. *Expected error reduction* (Roy and McCallum 2001) seeks to minimize the generalization error of the model by assuming an unlabeled instance is labeled.

The development of active learning strategies for multi-class soft-label settings has not been explored in the literature. In this paper, we propose an effective active learning approach for multi-class classifiers with probabilistic soft-label information. It selects the unlabeled instance with the highest expected projection change. To avoid the model retraining, we approximate the projection change via gradients, which remarkably reduces its running times. We show that our active learning strategy can substantially reduce the number of examples it needs to query.

Methodology

Soft-label multi-class support vector machine

Problem settings Our goal is to learn a multi-class classifier $f : X \to Y$, where X is the feature space and $Y \in \{1, 2, \ldots, k\}$ represents class labels of a data instance. We assume that in addition to class labels $\{1, 2, \ldots, k\}$ defining y_i we also obtain soft-label information: a probability p_i reflecting annotator's confidence the example $\mathbf{x}_i$ belongs to class y_i. Hence each labeled data entry D_i consists of three components: $D_i = (\mathbf{x}_i, y_i, p_i)$, an input, a class label and an estimate of the probability of the class label.

Learning soft-label multi-class classifier To elaborate our soft-label multi-class classifier, we start by modifying multi-class support vector machine, which we build our model upon. Basically, in multi-class support vector machine, we learn k binary support vector machine jointly, one for each class. For each data instance, the projection from the binary classifier of the class label should be higher than the projection from other classes. Formally, we would like to get k projection mappings $f_1(\cdot), f_2(\cdot), \ldots, f_k(\cdot)$, such that for each data instance $\mathbf{x}_i$, the projection $f_{y_i}(\mathbf{x}_i)$ is greater than $f_l(\mathbf{x}_i)$ for $l \in \{1, 2, \ldots, k\} \setminus y_i$. To permit some flexibility, we allow violations of the constraints but penalize them through the loss function. Therefore, multi-class support vector machine is formulated as follows:

$$\min_{W, \mathbf{w}_0, \boldsymbol{\xi}} G = \frac{1}{2} \sum_{j=1}^{k} \mathbf{w}_j^T \mathbf{w}_j + C \sum_{i=1}^{N} \sum_{j \neq y_i} \xi_{i,j}$$
$$(\mathbf{w}_{y_i} - \mathbf{w}_j)^T \mathbf{x}_i + (w_{0,y_i} - w_{0,j}) \geq 1 - \xi_{i,j} \qquad (\xi_{i,j} \geq 0),$$

where y_i is the class label of $\mathbf{x}_i$; $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k$ and $w_{0,1}, w_{0,2}, \ldots, w_{0,k}$ are the parameters and biases for the k binary classifiers. For prediction, the class with the highest projection value is selected as the predicted class.

Now we need to incorporate the soft labels into the model. Perhaps the most straightforward intuition is to incorporate the exact soft-label values. For example, we may reformulate the k binary classifiers into k regression models based on the soft labels. However, it is well known that humans are often unable to give consistent probabilistic assessments (Juslin, Olsson, and Winman 1998; Griffin and Tversky 1992). In other words, soft labels from human annotators are usually noisy which may backfire if we dwell too strongly on their exact values. To handle this, we incorporate the soft labels via constraints derived from ordinal regression (Chu and Keerthi 2005), which was first proposed by (Xue and Hauskrecht 2017) for binary classifiers. Briefly, we first split the soft-label space into multiple consequent and non-overlapping bins for each one-vs-all classifier. Then we try to enforce the pairwise orderings between each bin boundary and each soft label in this class. Formally, for each one-vs-all classifier $f_j(\cdot)$ and each data instance $\langle \mathbf{x}_i, y_i, p_i \rangle$ such that $y_i = j$, we try to enforce its projection $f_j(\mathbf{x}_i)$ will fall into the bin consistent with its soft label y_i. Meanwhile, we still try to enforce that $f_j(\mathbf{x}_i)$ is the highest among all one-vs-all classifiers. For example, if a data instance $\mathbf{x}$ belongs to class 3 and soft label 0.4, we want to enforce that the projection f_3 distinguishing class 3 will not only put $\mathbf{x}$ into the bin consistent with its soft label 0.4, but also is greater than any other projection $f_l(\mathbf{x})$ where $l \in \{1, 2, \ldots, k\} \setminus 3$ to guarantee that $\mathbf{x}$ will still be predicted as class 3. Also, we allow violations of both kinds of constraints by penalizing the loss function. By combining two kinds of constraints, we can formulate the following optimization problem:

$$\min_{W, \mathbf{w}_0, \boldsymbol{\eta}, \boldsymbol{\xi}, \mathbf{b}} G = \frac{1}{2} \sum_{l=1}^{k} \mathbf{w}_l^T \mathbf{w}_l + B \sum_{i=1}^{N} \sum_{l \neq y_i} \eta_{i,l} + C \sum_{i=1}^{N} \sum_{j=1}^{m-1} \xi_{i,j}$$
$$(\mathbf{w}_{y_i} - \mathbf{w}_l)^T \mathbf{x}_i + (w_{0,y_i} - w_{0,l}) \geq 1 - \eta_{i,l} \qquad (\eta_{i,l} \geq 0)$$
$$z_{i,j}(\mathbf{w}_{y_i}^T \mathbf{x}_i + w_{0,y_i} - b_j) \geq 1 - \xi_{i,j} \qquad (\xi_{i,j} \geq 0), \quad (1)$$

where y_i is the class label of $\mathbf{x}_i$, $z_{i,j}$ is an indicator whether the projection of $\mathbf{w}_{y_i}^T \mathbf{x}_i$ is supposed to be greater or less than the jth bin boundary b_j (-1 for less and 1 for greater); $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k$ and $w_{0,1}, w_{0,2}, \ldots, w_{0,k}$ are the parameters and biases for the k binary classifiers.

Active learning

In this part, we develop an active learning framework that builds a multi-class classifier by actively querying a human annotator for assessing the instances using both the class and associated soft labels. 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 approximate projection change.

Expected approximate projection change The expected approximate projection change (EAPC) is inspired by the expected model change (Tong and Koller 2000). Breifly, expected approximate projection change selects the unlabeled instance that brings the greatest expected projection change when it is assumed labeled. Such strategy consists of two key quantities: projection change and expectation. When an unlabeled instance is assigned an assumed label, all the k one-vs-all classifiers will change, leading to changes in projections of all unlabeled instances. The projection change measures the absolute change of all unlabeled instances on all the k one-vs-all classifiers. Since in soft-label multi-class scenario, an assumed label contains a discrete class label and a continuous soft label, given the probability of each class label and conditional distribution of soft label, we can calculate the expectation of projection change over the space of assumed label for the unlabeled instance. Formally, when an unlabeled instance $\mathbf{x}^+$ is assigned an assumed label $\langle y^+, p^+ \rangle$, the current models $f_{i,L}(\cdot)$ built on labeled data L will change to $f_{i,L \cup \langle \mathbf{x}^+, y^+, p^+ \rangle}(\cdot)$ for all i. Given the probability $P(y^+|\mathbf{x}^+)$ and conditional density $p(p^+|\mathbf{x}^+, y^+)$, we can calculate the expected projection change $\Delta(\mathbf{x}^+)$ as:

$$\Delta(\mathbf{x}^+) = \sum_{y^+} (y^+|\mathbf{x}^+) \int_0^1 p(p^+|\mathbf{x}^+, y^+)$$
$$\sum_{i=1}^{k} \sum_{j \in U} |f_{i,L \cup \langle \mathbf{x}^+, y^+, p^+ \rangle}(\mathbf{x}_j) - f_{i,L}(\mathbf{x}_j)| dp^+$$

We select the unlabeled instance with highest expected projection change to be labeled next.

Approximating expectation One critical problem is the expectation. Unfortunately, since the soft-label space is continuous, it is typically unfeasible to obtain the soft-label distribution of an unlabeled instance directly. To solve this problem, we propose an approximation which splits the soft-label range into multiple consequent and non-overlapping segments, then calculate the conditional probability that the unlabeled instance falls into each segment. Since such approximation is similar to the binning strategy in (1), we can directly adopt the bins for the conditional probabilities. Formally, instead of conditional density, we split the soft-label range into m bins $\{q_1, q_2, \ldots, q_m\}$ and calculate the conditional probability $P(p^+ \in q^+|\mathbf{x}^+, y^+)$ for all i and q^+. Therefore, the expectation $\Delta(\mathbf{x}^+)$ can now be estimated as:

$$\Delta(\mathbf{x}^+) = \sum_{\substack{y^+\\k}} (y^+|\mathbf{x}^+) \sum_{q^+} P(p^+ \in q^+|\mathbf{x}^+, y^+)$$
$$\sum_{i=1} \sum_{j \in U} |f_{i, L \cup \langle \mathbf{x}^+, y^+, p^+ \in q^+ \rangle}(\mathbf{x}_j) - f_{i,L}(\mathbf{x}_j)|$$

Another problem is measurement of $P(y^+|\mathbf{x}^+)$ and $P(p^+ \in q^+|\mathbf{x}^+, y^+)$. In this work, we adopt the idea of density weight (Settles, Craven, and Friedland 2008). Briefly, if an unlabeled instance is closed to a labeled instance, they are of high probability with the identical label. Formally, for an unlabeled instance $\mathbf{x}^+$ and each labeled instance $\langle \mathbf{x}_i, y_i, p_i \rangle$, the probability they are with identical label is proportional to the inverse of their Euclidean distance $||\mathbf{x}_i - \mathbf{x}^+||_2$. Therefore, the joint probability $P(y^+|\mathbf{x}^+)P(p^+ \in q^+|\mathbf{x}^+, y^+) = P(y^+, p^+ \in q^+|\mathbf{x}^+)$ can be estimated as:

$$P(y^+|\mathbf{x}^+)P(p^+ \in q^+|\mathbf{x}^+, y^+) = \frac{1}{Z} \sum_{i \in L}^{y_i = y^+, p_i = p^+} \frac{1}{||\mathbf{x}_i - \mathbf{x}^+||_2},$$

where $Z = \sum_{i \in L} \frac{1}{||\mathbf{x}_i - \mathbf{x}^+||_2}$ is the normalization factor.

Approximating projection change Another concern is the projection change over the unlabeled data. When adding an unlabeled instance with an assumed label, the new "add-one" model should be retrained. Given U unlabeled instances, k classes, m bins (in (1), soft labels in the same bin give the identical optimization), we need to retrain kmU "add-one" models. To avoid retraining, we propose an approximation via gradient inspired by stochastic gradient descent (Bottou and Bousquet 2008). Briefly, when adding an unlabeled instance with an assumed label, we can treat the other (labeled) instances as constants, calculate the difference compared with the current model and take the gradient to approximate the projection change over the unlabeled data. Formally, when adding $\langle \mathbf{x}^+, y^+, p^+ \in q_j \rangle$ into (1), the new "add-one" model G^+ can be written via rectified function $[\cdot]_+$ (we omit the bias $w_{0,l}$ for convenience) as:

$$\min_{W, \mathbf{w}_0, \boldsymbol{\eta}, \boldsymbol{\xi}, \mathbf{b}} G^+ = \frac{\mathbf{w}^T \mathbf{w}}{2} + B \sum_{l \neq y^+} [1 - (\mathbf{w}_{y^+} - \mathbf{w}_l)^T \mathbf{x}^+]_+ +$$
$$B \sum_{i=1}^{N} \sum_{l \neq y_i} [1 - (\mathbf{w}_{y_i} - \mathbf{w}_l)^T \mathbf{x}_i]_+ +$$
$$C \sum_{j=1}^{m-1} [1 - z_j^+ (\mathbf{w}_{y^+}^T \mathbf{x}^+ - b_j)]_+ + C \sum_{j=1}^{m-1} \sum_{i=1}^{N} [1 - z_{i,j} (\mathbf{w}_{y_i}^T \mathbf{x}_i - b_j)]_+,$$

where z_j^+ is determined from p^+ for all j. Comparing with (1), we get:

$$\Delta G^+ = B \sum_{l \neq y^+} [1 - (\mathbf{w}_{y^+} - \mathbf{w}_l)^T \mathbf{x}^+]_+ + C \sum_{j=1}^{m-1} [1 - z_j^+ (\mathbf{w}_{y^+}^T \mathbf{x}^+ - b_j)]_+$$

Therefore, the gradient for each one-vs-all classifier can be calculated as:

$$\frac{\partial \Delta G^+}{\partial \mathbf{w}_l} = B \mathbf{x}^+ \mathbb{1}_{(\mathbf{w}_{y^+} - \mathbf{w}_l)^T \mathbf{x}^+ < 1} \quad (l \neq y^+)$$
$$\frac{\partial \Delta G^+}{\partial \mathbf{w}_{y^+}} = -B \mathbf{x}^+ \sum_{l \neq y^+} \mathbb{1}_{(\mathbf{w}_{y^+} - \mathbf{w}_l)^T \mathbf{x}^+ < 1}$$
$$- C \mathbf{x}^+ \sum_{j=1}^{m-1} z_j^+ \mathbb{1}_{z_j^+ (\mathbf{w}_{y^+}^T \mathbf{x}^+ - b_j) < 1}$$

In the stochastic gradient descent, the negative gradient determines the step length for learning. Therefore, we claim the gradient is approximately proportional to the change of the parameter of each one-vs-all classifier:

$$\Delta \mathbf{w}_l^+ \underset{\sim}{\propto} \frac{\partial \Delta G^+}{\partial \mathbf{w}_l} \qquad l = 1, 2, \ldots, k$$

Given an arbitrary unlabeled instance $\mathbf{x}_j$, we can approximate the absolute projection change on $\mathbf{w}_i$ before and after $\langle \mathbf{x}^+, y^+, p^+ \in q^+ \rangle$ as:

$$|f_{i, L \cup \langle \mathbf{x}^+, y^+, p^+ \in q^+ \rangle}(\mathbf{x}_j) - f_{i,L}(\mathbf{x}_j)| = |\frac{\partial \Delta G^+}{\partial \mathbf{w}_i}^T \mathbf{x}_j|$$

Experiments and Results

We test our framework on both synthetic and real-world data. The first experiment adapts data from two UCI multi-class data sets which we transform to soft-label multi-class classification tasks. The second experiment works with a real-world image data with human assessed labels from multiple annotators.

Experiments on simulated data

Data simulation We adapted two UCI multi-class datasets (see Table 1 for details) as follows: We take half of the data to train a multi-class support vector machine and obtain probabilistic scores on the other half via soft-max function on their predictions. In the experiments we use only the second half of the data, retain the class labels and keep the corresponding probabilistic scores as soft labels.

Experimental settings To demonstrate the benefits of our soft-label model and expected approximate projection change strategy, we compare it with multi-class classifiers trained only on class labels, soft-label multi-class logistic regression and active learning that retrains to calculate the exact projection change when adding an unlabeled instance. Our experiments compare the following classifiers (we use random sampling by default):

MSVM: multi-class support vector machine (Vapnik 1998; Weston et al. 1999) where K one-vs-all classifiers are trained jointly;

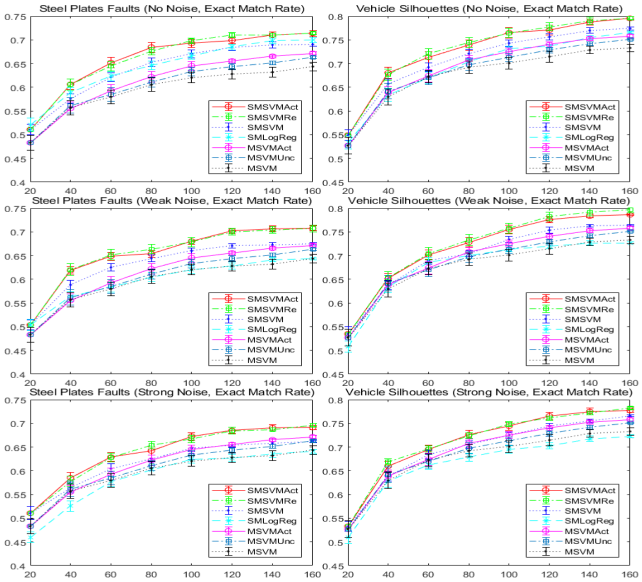

Figure 1: Performance (EMR) on UCI data with no (top), weak (middle) and strong (bottom) noise.

MSVMUnc: multi-class support vector machine (Vapnik 1998; Weston et al. 1999) where K one-vs-all classifiers are trained jointly with uncertainty sampling;
MSVMAct: multi-class support vector machine (Vapnik 1998; Weston et al. 1999) where K one-vs-all classifiers are trained jointly with expected approximate projection change strategy;
SMLogReg: soft-label multi-class logistic regression where K one-vs-all classifiers are trained independently on exact soft labels;
SMSVM: soft-label multi-class support vector machine where K one-vs-all classifiers are trained jointly with soft-label constraints;
SMSVMRe: soft-label multi-class support vector machine where K one-vs-all classifiers are trained jointly with soft-label constraints and retraining the model to calculate the exact projection change when an unlabeled instance is added;
SMSVMAct: soft-label multi-class support vector machine where K one-vs-all classifiers are trained jointly with soft-label constraints and expected approximate projection change strategy.
We evaluate the performance of the different methods in the exact match rates (EMR) on the test data. All data sets before learning are split into the training and test set (using $\frac{2}{3}$ and $\frac{1}{3}$ of all data instances). The learning considered training data only; the EMR is always measured in the test set. We also repeat the splitting and learning 24 times. The average EMRs of different classifiers on UCI data regarding increasing sizes of N are reported in Figure 1.

Experimental results Figure 1 (top) shows the benefit of our framework *SMSVMAct* with a combination of probabilistic soft labels and expected approximate projection change strategy. Both *SMSVMAct* and *SMSVMRe* outperform *MSVMAct* and *MSVMUnc*; both *SMSVM* and *SMLogReg* outperform *MSVM*. These two comparisons show soft labels will achieve better performance than original class label models with the same training sizes. Meanwhile, both *SMSVMAct* and *SMSVMRe* outperform *SMSVM*; *MSVMAct* outperforms *MSVMUnc* and *MSVM*. These two comparisons show the effectiveness of our expected approximate projection change strategy. Overall, both *SMSVMAct* and *SMSVMRe* are always of highest performance, showing that our framework remarkably raises the performance on the same sizes of training data.

Dataset	# Instances	# Features	# Classes
Steel Plates Faults	1941	27	7
Vehicle Silhouettes	946	18	4

Table 1: Properties of UCI data in synthetic experiments.

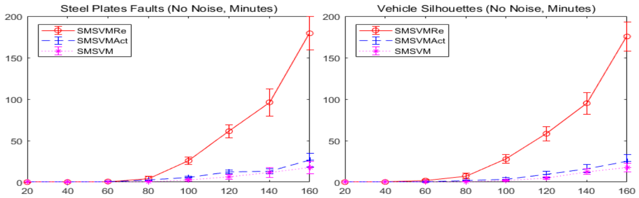

Figure 2: Time consumption (minutes) on UCI data with no noise.

Noise simulation In order to generate soft-label noise, each soft label p derived from the UCI data was modified into p' by injecting a Gaussian noise of different strength:
Weak noise: $p' = p \times (1 + 0.1 \times N(0, 1))$;
Strong noise: $p' = p \times (1 + 0.3 \times N(0, 1))$.
Briefly, the noise injection levels above indicate the average proportion of noise to no, weak (10%) and strong (30%) levels respectively. Also, we truncated the illegal probabilistic scores (e.g. probabilistic scores that are less than $\frac{1}{k}$ or greater than 1) to the interval of $[\frac{1}{k}, 1]$.

Experimental results with noise When soft-label noise is added, the performance of soft-label models may deteriorate. Figure 1 (middle) and (bottom) shows the robustness of our framework *SMSVMAct*. The regression based model *SMLogReg*, which is trained on exact soft labels, is vulnerable to noise and deteriorates remarkably. While other soft-label models are more robust and do not suffer from much performance drop. Our framework *SMSVMAct* are still of top two performance comparable with *SMSVMRe*, showing the robustness of our framework.

Experiments on time consumption The reason we use gradient to approximate projection change is to reduce time consumption. Figure 2 shows the time consumption of three soft-label multi-class classifiers in experiments on UCI data sets for increasing sizes of training data.

We evaluated the time consumption of the different learning methods by the total minutes elapsed on the training data. Because of the calculation of projection change, both *SMSVMAct* and *SMSVMRe* spend more time than *SMSVM*. However, the time consumption of *SMSVMAct* is tolerable, while the time consumption of *SMSVMRe* is seven times as *SMSVMAct* and ten times as *SMSVM*. Overall, our framework *SMSVMAct*, which combines soft labels and active learning, is of both higher performance than other models

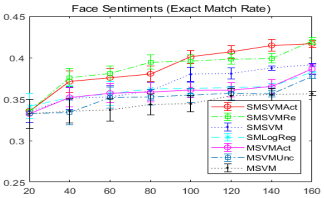

Figure 3: Performance (EMR) of Fact Sentiment.

that utilize at most one of the two methods, and far more satisfactory time consumption since it prevents retraining.

Experiments on real-world data

We also run experiments on Face Sentiment data, a real-world crowd-sourced dataset from Tsinghua University.

Experimental settings Face Sentiment data contains 584 data instances, where each instance is a 128 × 120 grayscale photo of the facial expression. The class label is one of the four moods indicating the mood in the photo. Each data instance is annotated by nine annotators. The true label of each data instance is also given. We use a convolutional neural network to extract 256 features for each data instance. For probabilistic soft-label models, we take the vote ratio of the true class among nine annotators as the soft label. We split all data instances into $\frac{2}{3}$ training and $\frac{1}{3}$ testing data, and measure average exact match rate over 24 trials.

Experimental results Figure 3 shows the benefit of our framework *SMSVMAct* with a combination of probabilistic soft labels and expected approximate projection change strategy on real-world face sentiment data. Both *SMSVMAct* and *SMSVMRe* outperform *MSVMAct* and *MSVMUnc*; both *SMSVM* and *SMLogReg* outperform *MSVM*. These two comparisons show soft labels will achieve better performance than original class label models with the same training sizes. Meanwhile, both *SMSVMAct* and *SMSVMRe* outperform *SMSVM*; *MSVMAct* outperforms *MSVMUnc* and *MSVM*. These two comparisons show the effectiveness of our expected approximate projection change strategy. Overall, both *SMSVMAct* and *SMSVMRe* are always of highest performance, showing that our framework remarkably raises the performance on both simulated and real-world data.

Conclusion

In this work, we proposed a new framework for multi-class classification models incorporating probabilistic soft-label information and a novel active learning strategy with efficient approximation that: (1) can learn more efficiently and from a smaller number of examples than existing methods, (2) is of higher performance than models that rely on only soft labels or active learning individually, and (3) can highly reduce time consumption than active learning strategy that requires retraining when adding new data instances.

Acknowledgement The work in this paper was supported by grant R01GM088224 from the NIH. Its content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

Bottou, L., and Bousquet, O. 2008. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems*, volume 20. 161–168.

Breiman, L. 1996. Bagging predictors. *Machine Learning* 24(2):123–140.

Chu, W., and Keerthi, S. S. 2005. New approaches to support vector ordinal regression. In *Proceedings of the 22nd international conference on Machine learning*, ICML '05, 145–152.

Griffin, D., and Tversky, A. 1992. The weighing of evidence and the determinants of confidence. *Cognitive Psychology* 24(3).

Herbrich, R.; Graepel, T.; and Obermayer, K. 1999. Support vector learning for ordinal regression. In *International Conference on Artificial Neural Networks*, 97–102.

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

Juslin, P.; Olsson, H.; and Winman, A. 1998. The calibration issue: Theoretical comments on suantak, bolger, and ferrell. *Organizational Behavior and Human Decision Processes* 73(1):3–26.

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

Nguyen, Q.; Valizadegan, H.; and Hauskrecht, M. 2011a. Learning classification with auxiliary probabilistic information. In *IEEE International Conference on Data Mining*, 477–486.

Nguyen, Q.; Valizadegan, H.; and Hauskrecht, M. 2011b. Sample-efficient learning with auxiliary class-label information. In *Proceedings of the Annual American Medical Informatics Association Symposium*, 1004–1012.

Nguyen, Q.; Valizadegan, H.; and Hauskrecht, M. 2013. Learning classification models with soft-label information. *Journal of American Medical Informatics Association*.

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

Roy, N., and McCallum, A. 2001. Toward optimal active learning through sampling estimation of error reduction. In Brodley, C. E., and Danyluk, A. P., eds., *Proceedings of the 18th International Conferenceon on Machine Learning*, 441–448. Williams College,Williamstown, MA, USA: Morgan Kaufmann.

Settles, B.; Craven, M.; and Friedland, L. 2008. Active learning with real annotation costs. In *Proceedings of the NIPS workshop on cost-sensitive learning*, 1–10.

Settles, B. 2010. Active learning literature survey. Technical report.

Seung, H. S.; Opper, M.; and Sompolinsky, H. 1992. Query by committee. In *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, 287–294. Pittsburgh, Pennsylvania: ACM Press.

Tong, S., and Koller, D. 2000. Active learning for parameter estimation in bayesian networks. In *Advances in Neural Information Processing Systems*, 647–653. MIT Press.

Vapnik, V. N. 1998. *Statistical Learning Theory*. New-York: Wiley.

Weston, J.; Gammerman, A.; Stitson, M.; Vapnik, V.; Vovk, V.; and Watkins, C. 1999. Support vector density estimation.

Xue, Y., and Hauskrecht, M. 2017. Efficient learning of classification models from soft-label information by binning and ranking. *Proceedings of the 30th International Florida AI Research Society Conference. Florida AI Research Symposium* 2017:164–169.