Efficient Learning of Classification Models from Soft-label Information by Binning and Ranking

Yanbing Xue
Department of Computer Science
University of Pittsburgh
yax14@pitt.edu

Milos Hauskrecht
Department of Computer Science
University of Pittsburgh
milos@pitt.edu

Abstract

Construction of classification models from data in practice often requires additional human effort to annotate (label) observed data instances. However, this annotation effort may often be too costly and only a limited number of data instances may be feasibly labeled. The challenge is to find methods that let us reduce the number of the labeled instances but at the same time preserve the quality of the learned models. In this paper we study the idea of learning classification from soft label information in which each instance is associated with a soft-label further refining its class label. One caveat of applying this idea is that soft-labels based on human assessment are often noisy. To address this problem, we develop and test a new classification model learning algorithm that relies on soft-label binning to limit the effect of soft-label noise. We show this approach is able to learn classification models more rapidly and with a smaller number of labeled instances than (1) existing soft label learning methods, as well as, (2) methods that learn from class-label information.

Introduction

Meaningful use of data often requires annotation of these data by humans. This is critical for building various kinds of classification models capable of differentiating examples according to human defined categories. Examples of such problems are annotation of text with human-preselected keywords or topics, annotation of time series (e.g. videos) with activities or events of interests captured in the data, annotation of patient instances with diseases, and many others. Unfortunately, due to its cost, the annotation of the data may become the bottleneck of the model building process. Briefly, the annotation effort may be too costly and only a limited number of data instances may be feasibly labeled. The challenge then is to develop methods that can learn high-quality models from a smaller number labeled instances.

Our solution focuses on binary classification and seeks to advance a relatively new machine learning approach proposed to address the sample annotation problem: learning with soft label information (Nguyen, Valizadegan, and Hauskrecht 2011a; 2011b), in which each instance is associated with a soft-label reflecting the certainty or belief of human annotators in the specific class label, such as, a probability the patient suffers from a specific disease. The benefit of soft labels is that they let us distinguish data instances that are strong, weak or marginal representatives of a class, and when properly used in the classification training phase they can help us learn better classification models with a smaller number of labeled samples. .

Throughout this paper we assume that soft-labels provided by humans are probabilistic. However, the caveat of learning from such labels is that humans are unable to give consistent probabilistic assessments; a phenomenon well documented in psychology and decision making literature (Juslin, Olsson, and Winman 1998; Griffin and Tversky 1992). In such a case, learning methods that are robust to 'noisy' soft-label assessments are necessary. (Nguyen, Valizadegan, and Hauskrecht 2011a; 2011b; 2013) address the problem by using probabilistic soft-labels to first determine the relative order of examples in the training data and then build the final classification model by considering all pairwise orderings among them (Joachims 2002; Herbrich, Graepel, and Obermayer 1999). They showed this approach is more robust to the soft-label noise than regression methods trying to directly fit probabilities. However, the limitations of their approach is that (1) the number of pairwise orderings one aims to satisfy is quadratic in the number of data points in the training data, and (2) all orderings are treated equally, that is, orderings induced by data points with smaller soft-label differences are treated equally to orderings with larger differences.

In this work we first show how one can modify the all-pair problem formulation through binning where constraints within each bin are ignored and only constraints among data points in the different bins are enforced. This leads to a smaller number of pairwise constraints to satisfy and exclusion of constraints that are more likely corrupted by the noise. Second, we reformulate the problem of satisfying constraints among data points in different bins as an ordinal regression problem and solve it using ranking-SVM (Joachims 2002; Herbrich, Graepel, and Obermayer 1999) defined on these bins (Chu and Keerthi 2005). This reformulation further reduces the number of constraints one has to satisfy leading to a more efficient solutions where the number constraints to satisfy is linear in the number of samples.

The paper is structured as follows. In Section 2, we review the related soft-label learning work. In Section 3, we introduce our approach and analyze its benefits In Section 4, we conduct experiments on multiple synthetic datasets (cor-

rupted with the different noise levels) and real-world human-labeled datasets in the clinical event prediction domain. In Section 5, we summarize the results and new directions.

Related Work

Learning with soft-label information The problem of learning binary classification models from soft-labels obtained from humans is relatively new and was first explored, to the best of our knowledge, by (Nguyen, Valizadegan, and Hauskrecht 2011a; 2011b; 2013). In their work, 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 (Juslin, Olsson, and Winman 1998; Griffin and Tversky 1992; O'Hagan et al. 2007). To deal with this problem they proposed a more robust method that ignores numerical differences in probability assessments and replaces them with pairwise ordering of data points in the training data. The method learns a parametric discriminative model that maximizes the satisfaction of all the 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 it consists of is quadratic in the number of training data instances.

More recently, (Peng and Wong 2014; Peng, Wong, and Yu 2014) proposed a new non-parametric algorithm for predicting the probability associated with binary classes based on the Gaussian process regression. 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 to 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.

The model we propose in this work builds upon the work of (Nguyen, Valizadegan, and Hauskrecht 2013; 2011a; 2011b), but limits the number of constraints one has to satisfy, by binning the probabilistic labels, and by applying the ordinal regression SVM approach (Chu and Keerthi 2005) for satisfying the constraints among the bins.

Methodology

We start by first defining and formalizing our learning problem. After that we review an algorithm proposed by (Nguyen, Valizadegan, and Hauskrecht 2011a; 2011b) for learning the classification model from data enriched with soft-labels, and gradually modify it to make it (a) more robust to noise and (b) more efficient to solve.

Problem description Our objective is to learn a binary classifier $f : X \rightarrow Y$, where X is an input (feature) space and $Y = \{0, 1\}$ represents class labels one can assign to individual input instances. We want to learn the classifier, starting from an unlabeled dataset D_U that consists of input instances only. The labels to examples are assigned by a human annotator. In this work, we assume that in addition to binary $\{0, 1\}$ labels defining Y we also obtain soft-label information: a probability p_i reflecting annotators belief the example $\mathbf{x}_i$ belongs to class 1. Hence each labeled data entry d_i we can learn from 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 class 1. For example, if $\mathbf{x}$ is a patient and y denotes the presence or absence of a disease or some adverse condition that is based on physician's evaluation of the patient, the probability p_i captures the physician's belief the patient indeed suffers from the condition. The human-label assessment, especially the soft-label part, may not be perfect. This problem is well documented and was discussed in the **Related Work**.

Method for learning with soft label information The approach we follow in this work is motivated by the model proposed by (Nguyen, Valizadegan, and Hauskrecht 2011a; 2011b) that is more robust to soft-label noise. Briefly, instead of fitting the precise probabilities, it models the relation between probabilistic assessments in terms of pairwise order constraints of any two data entries in the labeled data, and uses them to drive the construction of a binary classifier.

To explain the approach in more depth, let us consider a function $f(\mathbf{x}_i) = \mathbf{w}^T\mathbf{x}_i$ allowing us to discriminate between data entries of class 0 and class 1 after picking an appropriate threshold value. Using the soft label information one way we can learn the function is by fitting the examples and probabilistic labels directly via regression. However, because regression is sensitive to the soft-label noise, (Nguyen, Valizadegan, and Hauskrecht 2011a; 2011b) propose to learn this function from pairwise constraints induced by the probabilities. Briefly, if any two data entries $\mathbf{x}_j$ and $\mathbf{x}_k$ in the training data satisfy $p_j > p_k$, we expect the ordering function will preserve the order, that is $f(\mathbf{x}_j) > f(\mathbf{x}_k)$ or $f(\mathbf{x}_j) - f(\mathbf{x}_k) = \mathbf{w}^T(\mathbf{x}_j - \mathbf{x}_k) > 0$. The approach in (Nguyen, Valizadegan, and Hauskrecht 2011a) aims to satisfy (pairwise) constraints for all pairs of examples in the training data. Since in practice some constraints may be violated, the authors' limit the number of pairwise constraint violations by using the pairwise-constraint loss function that is incorporated in the following optimization problem for finding the discriminative model (Nguyen, Valizadegan, and Hauskrecht 2011a):

$$\min_{\mathbf{w},w_0,\eta,\xi} \quad \frac{\mathbf{w}^T\mathbf{w}}{2} + B\sum_{i=1}^{N}\eta_i + C\sum_{j=1}^{N-1}\sum_{k=j+1}^{N}\xi_{j,k}$$
$$y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1 - \eta_i \quad \forall i$$
$$\mathbf{w}^T(\mathbf{x}_j - \mathbf{x}_k) \geq 1 - \xi_{j,k} \quad \forall j,k(p_j > p_k)$$
$$\eta_i, \xi_{j,k} \geq 0 \quad \forall i,j,k$$

where $i = 1, 2, \ldots, N$, $j = 1, 2, \ldots, N-1$ and $k = j+1, j+2, \ldots, N$ index entries in the training data. w_0 defines the bias term and together with $\mathbf{w}$ it defines the binary decision boundary for the model. The first term in the objective function: $\frac{\mathbf{w}^T\mathbf{w}}{2}$ defines a regularization penalty, the second term (single sum) defines the hinge loss for all examples and their binary labels, and the third term (double sum) defines the pairwise-constraint loss function for pairs

of soft labels. η_i are slack variables defining the hinge loss, and $\xi_{j,k}$ slack variables reflecting individual constraint violation penalties for soft label pairs $p_j > p_k$. Finally B and C are constants weighting the different loss and regularization terms in the objective function. The optimization will find the weights $\mathbf{w}$ and w_0 and the corresponding discriminant function that violates the minimum constraints.

Reducing the number of constraints via binning The number of soft-label constrains in the above problem formulation is $O(N^2)$, more precisely $\frac{N(N-1)}{2}$. This negatively affects the efficiency of its solution. In this work we study binning to alleviate the problem.

The gist of the binning approach is that we divide the instances into several consequent, non-overlapping bins according to their soft label information. The constraints for pairs of instances that fall into the same bin are then ignored; the constraints among instances in different bins are kept. One reason for applying this approach is that by binning we are more likely to remove constraints for instances with smaller soft label differences, while preserving constrains for instances with larger soft label differences. This is important since the soft-label noise (due to human variation in soft-label assessment) is more likely to flip the order of instances with small soft-label difference than the order of instances with larger soft-label difference. Hence the net effect of the binning is (1) the reduction in the number of constraints, as well as, (2) the selection of constraints that are more likely to be correct in terms of instance ordering. However, we would like to note that even with binning, the number of pairwise constraints in the formulation remains quadratic or $O(N^2)$. In the following, we develop a more efficient solution based on the ordinal regression that significantly improves the number of constraints one has to satisfy while learning the model.

The idea of binning is to satisfy constraints only among entries placed in the different bins. Optimally we would like to have data entries that are in the same bin according to its probability label fall into the same bin also after the projection. We can use this intuition to reformulate the optimization problem as an ordinal regression problem (Chu and Keerthi 2005). Briefly we want to find the function $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ that puts the data points into bins according to their soft label. We can achieve this by having every example $\mathbf{x}$ project on the correct side of each bin boundary. For example, if the example $\mathbf{x}$ is located in ith bin, then after the projection, $f(\mathbf{x})$ should be smaller than the lower margin (boundary) of bin j in the projected space, whenever $i < j$. In general, assuming m bins labeled from 1 to m, bin boundaries $b_1, b_2, \ldots b_{m-1}$ separating them in the projected space, and bin function $bin(p_i)$ that maps the probability to the bin number (lowest probability maps to lowest number), then, after the projection, the example x_i with soft label p_i should project to value smaller than b_j whenever $bin(p_i) \leq j$, otherwise its value should be larger than b_j. Overall, for N data entries and m boundaries there are $(m-1)N$ constraints, one for each data entry/boundary pair. To guarantee the robustness of our model against soft label noise, we allow violations of constraints by penalizing the loss function of sample/boundary pairs. By adding the constraints for binary class labels, we can formulate the following optimization problem:

$$\min_{\mathbf{w},w_0,\mathbf{b},\eta,\xi} \quad \frac{\mathbf{w}^T\mathbf{w}}{2} + B\sum_{i=1}^{N}\eta_i + C\sum_{j=1}^{m-1}\sum_{i=1}^{N}\xi_{j,i}$$

$$y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1 - \eta_i \qquad \forall i$$

$$\mathbf{w}^T\mathbf{x}_i - b_j \leq \xi_{j,i} - 1 \qquad \forall i, j(bin(p_i) \leq j)$$

$$\mathbf{w}^T\mathbf{x}_i - b_j \geq 1 - \xi_{j,i} \qquad \forall i, j(bin(p_i) > j)$$

$$\eta_i, \xi_{j,i} \geq 0 \qquad \forall i, j$$

where $j = 1, 2, ..., m-1$ indexes bin boundaries in $\mathbf{b}$, and $i = 1, 2, ..., N$ indexes data entries. The first term in the objective function is the regularization term, the second term (single sum) defines the hinge loss with respect to binary labels, and the third term (double sum) defines the bin-constraint loss function. η_i and $\xi_{j,i}$ are slack variables permitting violations of binary class and soft-label bins respectively. B and C are constants weighting the objective function terms. Again, this optimization yields a discriminant function $f(\mathbf{x}_i) = \mathbf{w}^T\mathbf{x}_i + w_0$ that tries to minimize the number of violated constraints, but the number of constrains is reduced to $O(mN)$ as compared to $O(N^2)$ for the pairwise-ordering methods (with or without the binning).

Choosing the number of bins One question that remains open is how to define bins and how to choose their number. One possible solution to define the bins is to use an equal distance binning that splits the range of values (in our case soft-label values) equally. Another possibility choose bins of equal size. In our work, we use equal size binning, that is, the bin boundaries are built such that each bin covers approximately the same number of examples. The challenge, however, is to choose the number of bins. The caveat here is that the number of bins may affect the quality of the result. If we use $N-1$ bins where each bin only contains one data sample, the optimization problem is similar to (Nguyen, Valizadegan, and Hauskrecht 2013; 2011a; 2011b) with $O(N^2)$ constraints. On the other hand, if we only use two bins, the bin/sample pairwise ordering is equivalent to binary classification. The optimal bin choice is somewhere in between these two extremes.

One approach to select the number of bins is to use the internal cross-validation approach. Another is to use a heuristic. In this work we experimented with both approaches. The design of our heuristic is inspired by the results on the optimal binning for discretization of continuous values (Freedman and Diaconis 1981) who determined that the number of bins for N examples should follow $floor(\sqrt[3]{N})$ trend. We analyze this heuristic and compare it to the internal cross-validation approach.

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 data sets which we transform to soft-label problems. We use these data to show the performance of the methods when soft-labels are corrupted with the different level of noise. The second experiment works with real-world clinical data with true (human assessed) probabilistic labels.

Experiments on synthetic UCI-based data In this part we adapted one UCI regression data set (Housing) and

three UCI ordinal classification data sets (Cancer, Wine Red, Wine White) as follows. For the UCI housing regression data set we normalized the outputs ranging in R and reinterpreted them as probabilistic scores. We also defined a binary class threshold over the probabilistic scores to distinguish class 0 from class 1. For example, the outputs in Housing data set represents the attractiveness of houses to the consumers. In this case, we define two classes: houses with high attractiveness (class 1) and houses with low attractiveness (class 0). We use 30% of data entries with top score to define class 1, the rest are assigned to class 0. The UCI ordinal classification data sets come with multiple classes and full-order relations among classes. We generate probabilistic labels by evenly normalizing the class labels according to the total number of classes. The binary thresholds can be set according to the meaning of ordinal classes. For example, Breast Cancer data set 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. We use this information to re-map the class labels into $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$ with a threshold of 0.3 for the binary label.

SoftSVMOrd: Our new SVM-based ordinal regression model that splits the data entries into m bins according to the soft labels and enforces the bin-entry constraints. The bin size m is determined as $floor(\sqrt[3]{N})$.

SoftSVMRankPair: The soft-label method proposed in (Nguyen, Valizadegan, and Hauskrecht 2011a) that learns the model from all pairwise constraints reflecting the ordering of data instances.

SoftSVMRankKN: A special version of SoftSVMRankPair that uses KN pairwise constraints selected randomly from all $\frac{N(N-1)}{2}$ pairwise constraints. Throughout the experiments the constant K is selected to assure the SoftSVMOrd and SoftSVMRankKN methods always use the same number of constraints.

GPR: The Gaussian process regression approach (Peng, Wong, and Yu 2014) for learning with soft-label information.

SoftLogReg: The logistic-regression-based model based on (Nguyen, Valizadegan, and Hauskrecht 2011a) that fits the soft-label information directly to the logistic regression model.

BinarySVM: The standard linear SVM with the hinge loss and quadratic regularization trained on examples with binary labels.

We evaluated the performance of the different methods by calculating the Area under the ROC (AUC) the learned classification model would achieve on the test data. Hence, each data set prior to the learning was split into the training and test set (using $\frac{2}{3}$ and $\frac{1}{3}$ of all data entries respectively). The 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 soft label information on the number of data entries, we trace the performance of all models for the different sizes N of labeled data. Figure 1 shows the performance (AUC) of the models on all four UCI data sets for increasing sizes of N and the different levels of soft label noise.

Benefit of soft-labels. Figure 1(top) shows the performance of methods when simulated soft-labels are not corrupted by additional noise. The results show that all methods that rely on soft-label information outperform the SVM method trained on binary labels only. This demonstrates the sample-size benefit of soft-labels for learning classification models and basically reiterates the point made in (Nguyen, Valizadegan, and Hauskrecht 2011a). Out of all soft-label methods tried there does not seem to be a clear winner and all methods perform comparably well. Please notice that SoftLogReg method which fits the probabilities to the model via regression is comparable to other methods.

Effect of Noise on Soft Labels Figure 1(top) results assumed the soft labels directly reflect the probabilistic information. However, in practice, probabilistic information (when collected from humans) may be imprecise and subject to noise. This in turn may affect the quality of our models. Our synthetic noise experiments aim to show the robustness of the methods to noise in probabilistic scores. 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: $p' = p \times (1 + 0.10 \times N(0, 1))$
Strong noise: $p' = p \times (1 + 0.30 \times N(0, 1))$.

Briefly, the noise injection levels above indicate the average proportion of noise to signal at weak (10%) and strong (30%) levels respectively. Also, we truncated the illegal probabilistic scores (e.g. probabilistic score that are less than 0 or greater than 1) to the interval of $[0, 1]$. The results of the different methods for the weak and strong noise are summarized in the middle and bottom rows of Figure 1 respectively.

When noise is added into the probabilistic labels: Figure 1 (middle) and (bottom), the performance of a model may drop. One of the methods, SoftLogReg that directly fits probabilities is particularly sensitive to the noise and its performance drops significantly for both noise levels and across all datasets. Other soft-label models that use constraints or bins are more robust and do not suffer from such a performance drop. Our new method, SoftSVMOrd, is the most consistent and tends to outperform other SVM-based models: BinarySVM, SoftSVMRankPair and SoftSVMRankKN in both noise injection levels. It also outperforms GPR which is another recently proposed soft-label learning method. These experiments demonstrate the robustness of our method on the soft-label learning tasks.

Experiments on clinical dataset Whilst the experiments on synthetic data sets support the benefits of our soft-label approach, it is unclear whether these results also extend and generalize to 'true' soft-labels assessed by humans. In this set of experiments we test the performance of the methods on the real-world clinical data that were independently reviewed and assessed in terms of soft-labels by three different experts. The target label concerned Heparin induced thrombocytopenia (HIT), an adverse clinical condition that affects patient who are treated with heparin for prolonged periods of time. The clinical data consists of 50 patient state features important for detection of HIT derived from the PCP database (Hauskrecht et al. 2010; 2013). The datasets consists of 579, 571, and 573 labeled patient state instances for Expert 1, 2 and 3, respectively. The labels include both binary and soft-label information.

Figure 2 shows the AUC performance of the same methods and models as used in the previous section on three

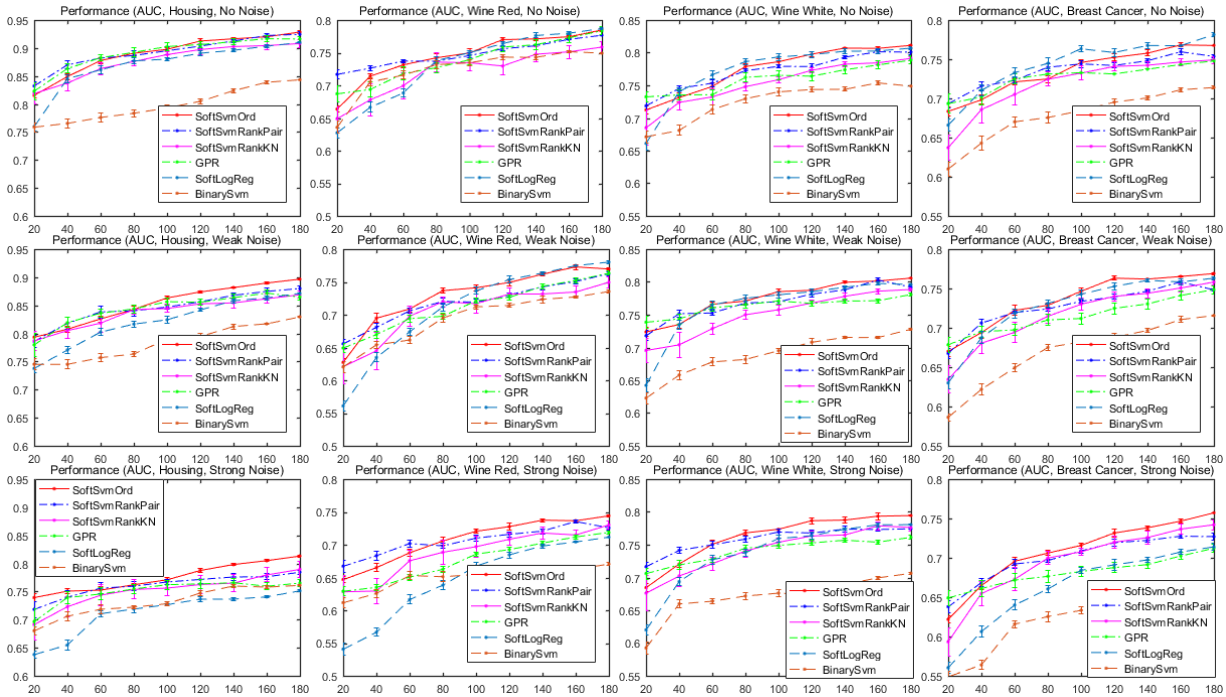

Figure 1: Performance of methods on four UCI datasets with no noise (top), weak noise (middle) and strong noise (bottom).

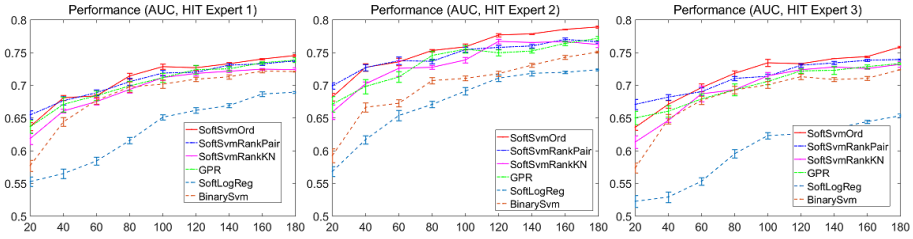

Figure 2: Performance on HIT data annotated by 3 experts

expert-annotated HIT datasets. On all three datasets the performance of our SoftSVMOrd method is the best and it outperforms all other methods. This experiment confirms good performance of our method and the benefit of the soft-labels for more efficient training of binary classification models.

Effect of the number of bins One of the parameters of our method (SoftSVMOrd) is the number of bins used to learn the model. Figure 3 illustrates the performance of the method for the different number of bins on the UCI housing data with two different levels of noise (weak and strong). The number of examples the model was trained on was fixed at $N = 100$. The AUC statistics are averages over 24 train/test splits.

The results in Figure 3 show that the number of bins indeed influences the quality of the model. Briefly, having two bins is equivalent to a binary classifier, so the benefit of soft labels is rather limited. On the other hand, many bins increase the number of constraints for pairs of data points with similar soft-labels which makes the approach more sensitive to the soft-label noise. The optimal operating point is in the middle; it is the bin number that best trade-offs positives and negatives of soft label information: (1) its ability to refine the discriminative model, (2) the soft-label noise

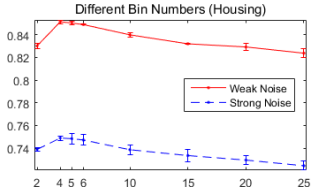

Figure 3: Performance of SoftSVMOrd method on the housing data for the different number of bins

that may switch relative order of two examples. Please also note that our cuberoot heuristic $floor(\sqrt[3]{100} = 4)$ correctly estimates the optimal bin number.

Earlier in the paper we mentioned cross-validation as an alternative to the cube-root heuristic to pick the number of bins. We would like to note, that although we run both the methods, their results were nearly identical.

To show how close these two approaches are, Figure 4 plots average differences in AUC scores for the cross-validation and heuristic approaches on the housing data (with three levels of noise) and three HIT datasets. Clearly the differences in performance across all these experiments are very small, suggesting the cube root heuristic a

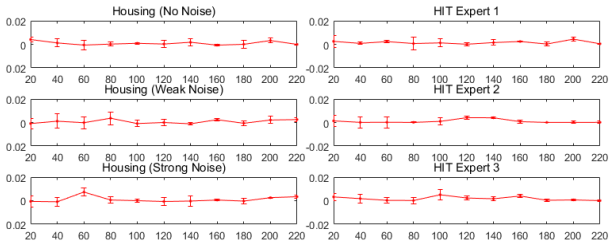

Figure 4: Average AUC difference for two versions of the SoftSVMOrd method on six datasets.

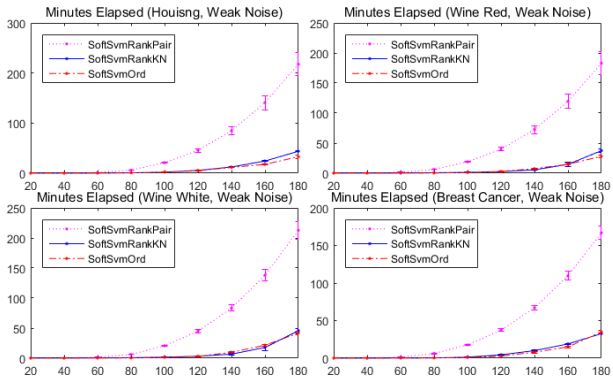

Figure 5: Time consumption (in minutes) of methods on four UCI datasets with weak noise.

good choice for determining the number of bins.

Experiments and results of time complexity One of the reasons for introducing the new binning method was to improve the pairwise constraint solution (SoftSVMRankPair method) proposed by Nguyen (Nguyen, Valizadegan, and Hauskrecht 2011a). Figure 5 shows the time consumption of three soft-label methods used earlier (SoftSVMRankPair, SoftSVMOrd and SoftSVMRankKN) on UCI data sets for increasing sizes of N and different levels of soft label noise.

We evaluated the time consumption of the different learning methods by the total minutes elapsed on the training data. For SoftSVMOrd and SoftSVMRankKN we always keep the same number of soft-label constraints: KN. As expected, SoftSVMOrd and SoftSVMRankKN running times are very close across all experiments. In contrast to these, the performance of SoftSVMRankPair that uses all $\frac{N(N-1)}{2}$ pairwise constraints deteriorates very quickly as N increases, and at $N = 180$ the running time increases about four fold when compared to our SoftSVMOrd approach. This confirms the running-time benefit of SoftSVMRankKN and SoftSVMOrd with the reduced number of soft-label constraints. Please notice that the results in Figure 1 and in Figure 5 combined demonstrate the benefit of our new method SoftSVMOrd. It tends to outperform the baseline SoftSVMRankPair in terms of the solution quality across many sizes N and this with a remarkably lower running time. It also outperforms SoftSVMRankKN in terms of the solution quality at comparable running times.

Conclusion

To obtain labels for classification purposes, we often rely on human annotators. However, the human annotation process may be costly. In such a case, different methods of reducing the labeling cost need to be applied. In this paper we have developed and tested a new robust method that uses soft-label information that is able to enrich the feedback one receives from human and hence improve the number of examples one has to label to get a good classification model. Our results on synthetic and real-world clinical data show that our method (1) can benefit greatly from additional soft-label information (2) is robust to the different levels of soft-label noise.

Acknowledgement

The work presented in this paper was supported 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

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. New York, NY, USA: ACM.

Freedman, D., and Diaconis, P. 1981. On the histogram as a density estimator. *Probability Theory and Related Fields* 57(4):453–476.

Griffin, D., and Tversky, A. 1992. The weighing of evidence and the determinants of confidence. *Cognitive Psychology* 24(3):411 – 435.

Hauskrecht, M.; Valko, M.; Batal, I.; Clermont, G.; Visweswaram, S.; and Cooper, G. 2010. Conditional outlier detection for clinical alerting. In *Proceedings of the Annual American Medical Informatics Association Symposium*, 286 – 290.

Hauskrecht, M.; Batal, I.; Valko, M.; Visweswaran, S.; Cooper, G. F.; and Clermont, G. 2013. Outlier detection for patient monitoring and alerting. *Journal of Biomedical Informatics* 46(1):47–55.

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.

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.

Peng, P., and Wong, R. C.-W. 2014. Selective sampling on probabilistic data. In *SIAM International Conference on Data Mining*, 28–36.

Peng, P.; Wong, R. C.-W.; and Yu, P. S. 2014. Learning on probabilistic labels. In *SIAM International Conference on Data Mining*, 307–315.