ECAI 2016
G.A. Kaminka et al. (Eds.)

doi:10.3233/978-1-61499-672-9-1618

Learning of Classification Models from Noisy Soft-Labels

Yanbing Xue [1] and **Milos Hauskrecht** [2]

Abstract. We develop and test a new classification model learning algorithm that relies on the soft-label information and that is able to learn classification models more rapidly and with a smaller number of labeled instances than existing approaches.

1 Introduction

While huge amounts of data in various areas of science, engineering, and every day life are available nowadays, these data alone may not be sufficient for all the learning tasks we are ultimately interested in, and additional data collection is necessary to address them. These learning tasks include classification problems in which class labels are based on subjective human assessment. Examples include various text annotation problems, annotation of images or videos, or annotation of patient cases with diseases, and many others. For all these problems annotation effort is needed to supplement the data. However, the annotation effort may often be too costly limiting the number of instances one may feasibly label. The challenge is to develop methods that can reduce the number of the labeled instances but at the same time preserve the quality of the learned models.

Here we study the sample labeling problem in binary classification settings. Our solution advances a relatively new approach to address the problem: learning with soft label information [7, 8], in which each instance is associated with a soft-label further refining the class label. Soft labels reflect the certainty of human annotators in the specific class label, such as, the probability the patient suffers from a specific disease. The benefit of soft labels is that they distinguish data instances that are strong, weak or marginal representatives of a class, and when properly used in the training phase they can help us learn better models with a smaller number of labeled samples.

In this work we assume that soft-labels given to learners by humans are probabilistic. The caveat of learning models from such labels is that humans are often unable to give consistent probabilistic assessments; a phenomenon well documented in psychology and decision making literature [6, 3]. In such a case, learning methods that are robust to 'noisy' soft-label assessments are necessary. [7, 8, 9] 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 [5, 4]. 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 (with both small and large soft-label differences) are treated equally.

Our objective is to develop a more efficient approach for learning models from noisy soft-label information. Our solution relies on soft-label binning. Briefly, we 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 [5, 4] defined on these bins [1]. This reformulation 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 data instances.

2 Methodology

Our binning approach divides data instances into multiple non-overlapping bins according to their soft label information. The idea 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 to reformulate the optimization problem as an ordinal regression problem [1]. 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.

In general, because of the soft label noise, we cannot expect that all the constraints will be always satisfied. We allow violations of constraints but penalize them via bin-constraint loss function. This leads to the following optimization problem:

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

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

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

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

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 non-negative slack variables permitting violations of binary class and soft-label bins respectively. B and C are constants weighting

[1] University of Pittsburgh, United States, E-Mail: yax14@pitt.edu
[2] University of Pittsburgh, United States, E-Mail:milos@cs.pitt.edu

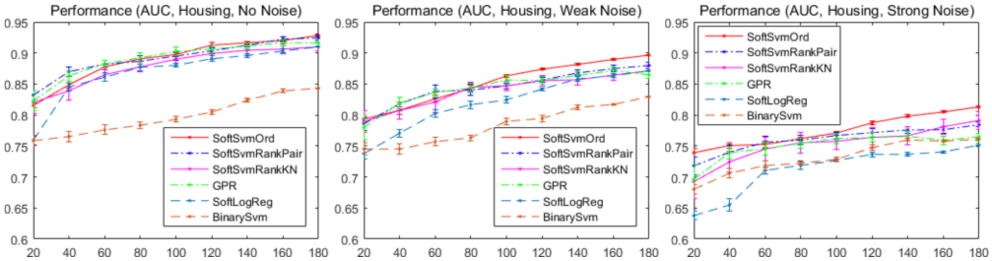

Figure 1. Performance on UCI Housing data set (no, weak, strong noise)

the objective function terms. 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.

One important open question is how to define the bins and how to choose their number. 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 is that the number of bins may affect the quality of the result. Briefly, choosing the number of bins to be equal to N (singleton bins) reduces to $(N-1)$ bin boundaries and the total of $O(N^2)$ constraints in the optimization problem which basically mimics all pairwise orderings. On the other side, having just two bins means we are trying to separate two groups of data points, which 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 a heuristic. Our heuristic is inspired by the results on the optimal binning for discretization of continuous values [2] who determined that the number of bins for N examples should follow $floor(\sqrt[3]{N})$ trend.

3 Experiments and Results

We use UCI Housing data set to test our method. We normalize the real-valued outputs and reinterpreted them as probabilistic scores. We also defined a binary class threshold over the probabilistic scores to distinguish class 0 from class 1. 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. Our experiments compare the following methods:

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

SoftLogReg: The logistic-regression-based model based on [7] that directly fits the soft-label information to the model.

GPR: The Gaussian process regression approach [10] for learning with soft-label information.

SoftSVMRankPair: The soft-label method proposed in [7] that relies on all pairwise ordering of data instances.

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

SoftSVMRankKN: A version of SoftSVMRankPair that uses a random subset of KN pairwise constraints. The value of K is selected to assure the SoftSVMOrd and SoftSVMRankKN methods always use the same number of constraints.

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. To avoid potential train/test split biases, we repeated the training process (splitting) and learning steps 24 times. We report the average AUC. To test 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(left) 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 [7]. Figure 1(left) assumes the soft labels are accurate. However, in practice, probabilistic information (when collected from humans) may be imprecise and subject to noise. In order to generate soft-label with the noise (p') we modify a soft label p derived from the UCI data by injecting a Gaussian noise of different strength. The noise injection levels indicate the average proportion of noise. Figures 1 (middle, right) show results for the noise signal at weak (10%) and strong (30%) levels respectively. The figures demonstrates that the performance of a model may drop when noise is injected. 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 data sets. 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. These experiments demonstrate the robustness of our method on the soft-label learning tasks.

Acknowledgement This work 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] W Chu et al, 'New approaches to support vector ordinal regression', in *Proc. of 22nd Int. Conf. on Machine learning*, 145-152, (2005).

[2] D Freedman et al, 'On the histogram as a density estimator', *Probability Theory and Related Fields*, **57**(4), 453-476, (1981).

[3] D Griffin et al, 'The weighing of evidence and the determinants of confidence', *Cognitive Psychology*, **24**(3), 411-435, (1992).

[4] R Herbrich et al, 'Support vector learning for ordinal regression', in *Int. Conf. on Artificial Neural Networks*, 97-102, (1999).

[5] T Joachims, 'Optimizing search engines using clickthrough data', in *Proc. of ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 133-142, (2002).

[6] P Juslin et al, 'The calibration issue: Theoretical comments on suantak, bolger, and ferrell', *Organizational Behavior and Human Decision Processes*, **73**(1), 3-26, (1998).

[7] Q Nguyen et al, 'Learning classification with auxiliary probabilistic information', in *IEEE Int. Conf. on Data Mining*, 477-486, (2011).

[8] Q Nguyen et al, 'Sample-efficient learning with auxiliary class-label information', in *Proc. of Annu. American Medical Informatics Assoc. Symp.*, 1004-1012, (2011).

[9] Q Nguyen et al, 'Learning classification models with soft-label information', *J. of American Medical Informatics Assoc.*, (2013).

[10] P Peng et al, 'Learning on probabilistic labels', in *SIAM Int. Conf. on Data Mining*, 307-315, (2014).