Learning classification models with soft-label information

Quang Nguyen, Hamed Valizadegan, Milos Hauskrecht

▶ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/amiajnl-2013-001964).

Computer Science Department, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

Correspondence to
Dr Milos Hauskrecht, Computer Science Department, University of Pittsburgh, 5329 Sennott Square Bldg, 210 S. Bouquet St, Pittsburgh, PA 15260, USA; milos@pitt.edu

Received 22 April 2013
Revised 24 October 2013
Accepted 1 November 2013
Published Online First
20 November 2013

ABSTRACT

Objective Learning of classification models in medicine often relies on data labeled by a human expert. Since labeling of clinical data may be time-consuming, finding ways of alleviating the labeling costs is critical for our ability to automatically learn such models. In this paper we propose a new machine learning approach that is able to learn improved binary classification models more efficiently by refining the binary class information in the training phase with soft labels that reflect how strongly the human expert feels about the original class labels.

Materials and methods Two types of methods that can learn improved binary classification models from soft labels are proposed. The first relies on probabilistic/numeric labels, the other on ordinal categorical labels. We study and demonstrate the benefits of these methods for learning an alerting model for heparin induced thrombocytopenia. The experiments are conducted on the data of 377 patient instances labeled by three different human experts. The methods are compared using the area under the receiver operating characteristic curve (AUC) score.

Results Our AUC results show that the new approach is capable of learning classification models more efficiently compared to traditional learning methods. The improvement in AUC is most remarkable when the number of examples we learn from is small.

Conclusions A new classification learning framework that lets us learn from auxiliary soft-label information provided by a human expert is a promising new direction for learning classification models from expert labels, reducing the time and cost needed to label data.

BACKGROUND AND SIGNIFICANCE

Large clinical data sets provide us with a great opportunity to better understand diseases and the efficacy of different treatments, and afford the possibility to build high-quality automated diagnosis systems. However, many of these real-world data sets are not perfect and lack information that we currently are unable to collect automatically. One type of such information is subjective labels provided by human experts, for example, in electronic health records (EHRs). While some of the data, such as laboratory tests and medications given, are archived and collected, the diagnoses of some conditions or adverse events that occur during hospitalization are not. In the context of supervised learning, in order to analyze these conditions and predict them, individual patient examples must be first labeled by an expert. However, the process of labeling examples using subjective human assessments can be very time-consuming, especially in the medical domain where data are complicated and their assessment requires a high level of expertise. As a result, the amount of labeled training data available for learning can be limited. The development of new methods that reduce dependency on the number of labeled examples becomes critical for their practical deployment.

To address this issue, we propose and study a new machine learning framework in which the binary class label information that is used to learn binary classification models is enriched by soft-label information reflecting a more refined expert's view on the class an instance belongs to. We expect the soft-label information, when applied in the training (learning) phase, will let us learn the binary classification models more efficiently with a smaller number of labeled patient instances. In general, the soft-label information can be represented either in terms of (1) a probabilistic (or numeric) score, for example, the chance of the patient having the disease is 0.7, or (2) a qualitative category, such as weak or strong agreement with the patient having the disease. The cost of obtaining this additional information is typically small once the patient case is reviewed by the expert.

Formally, we want to learn a binary classifier $g: X \rightarrow Y$, where X are feature vectors and Y are binary {0,1} labels. In the training phase, in addition to binary labels, we also have access to additional information: a soft label c_i reflecting one's belief that the example $\mathbf{x}_i$ belongs to class 1. Hence each data entry in the data set $D = \{d_1, d_2, \ldots, d_N\}$ we learn from consists of three components: $d_i = (\mathbf{x}_i, y_i, c_i)$, an input, a class label, and a soft label refining the class assignment. Our conjecture is that soft-label information, when properly used in the model training phase, can help us to learn a classification model more efficiently (with a smaller number of labeled examples) than with binary labels only.

In this paper we show how to adapt a number of existing machine learning frameworks to the new learning task. We demonstrate the benefits of these methods on clinical data by focusing on heparin induced thrombocytopenia (HIT)[1] and the construction of classification models that can, given a set of patients' observations, predict as accurately as possible patients who are at risk of HIT, and for whom the human expert would like to generate an HIT alert.

Related work

Solutions for reducing the cost of labeling have been studied extensively by the machine learning community. One of the most popular research directions for this problem is active learning.[2–4] The goal of active learning research is to develop methods that analyze unlabeled examples, prioritize them, and select those that are most critical for the task to be solved, while optimizing the overall data labeling cost.

To cite: Nguyen Q, Valizadegan H, Hauskrecht M. *J Am Med Inform Assoc* 2014;**21**: 501–508.

Our research direction is orthogonal to active learning: we aim to obtain more useful information from selected examples through soft labels and utilize it to learn better classification models. The research work most relevant to our problem was carried out by Smyth *et al*,[5] who studied ways of incorporating probabilistic information into a simple generative classification model in order to learn classification of volcanoes from radar images of distant planets. Our work and methods apply to a broader class of discriminative frameworks. Methodologically, our learning problem and solutions are related to and built upon classification and regression models, but modify them to permit learning with both binary class and soft-label information. Briefly, standard classification algorithms (eg, logistic regression,[6] support vector machines (SVMs)[7]) use only class labels, and do not accept soft labels. On the other hand, regression models[8] can learn from soft labels represented by continuous quantities, but do not use categorical information.

Another direction we build upon in this work is preference/rank learning.[9–11] The rank-learning algorithms rely on prior ordering of data examples and learn a ranking function that respects the ordering. The ranking function can be then used to define a classification model.

MATERIALS AND METHODS

Our goal is to learn a classification model $g: X \rightarrow Y$, where X denotes the input (or feature) space and $Y = \{0, 1\}$ are two classes one can assign to each input. Typically we approach this task by learning a discriminant function $f: X \rightarrow R$. A classifier is then defined using a decision threshold σ, such that if $f(\mathbf{x}) \geq \sigma$ then $y = 1$, otherwise $y = 0$. In the standard binary classification setting, the discriminant function is learned from examples with class labels {0,1} only. In our framework, in addition to class labels y, we also have access to auxiliary soft labels c associated with these class labels. In the following, we first present methods that learn binary classification models from auxiliary probabilistic label information, and after that, methods that rely on ordinal categorical labels.

Algorithms for learning with auxiliary probabilistic labels

In this section we develop classification learning algorithms that let us learn classifiers from probabilistic labels $c_i \in [0, 1]$.

Discriminative linear regression

One relatively straightforward solution is to assume the discriminant function is defined directly in terms of these auxiliary probabilities. In such a case, the learning of the discriminant function can be converted into a regression problem. One way to learn the function is to regress the features directly to probabilities, that is, we can learn a regression mapping f where $(\mathbf{x}_i, c_i)$ are the input–output pairs.

Assuming the function $f: X \rightarrow R$ is formed by a linear model $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$, where $\mathbf{w}$ are the parameters (or weights) of the model, the learning problem becomes a linear regression problem solved by minimizing the error function based on the sum of squared residuals:

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{N}\sum_{i=1}^{N} (\mathbf{w}^T\mathbf{x}_i - c_i)^2 + Q(\mathbf{w})$$

The term $Q(\mathbf{w})$ denotes an optional regularization term that may help to prevent the model over-fit.[8 12] Given the weights, the final classifier is defined using a decision threshold σ, such that if $\mathbf{w}^T\mathbf{x}_i \geq \sigma$ then $y_i = 1$, otherwise $y_i = 0$. We refer to this method as LinRaux (linear regression with auxiliary information).

Logistic regression model

By fitting a regression function model, the outputs of the model may fall outside the [0,1] range, and hence be inconsistent with probabilities. An alternative is to regress inputs to a new space of reals R obtained by transforming the probabilistic space, such that the transformation is monotonic in c_i, and its inverse lets us revert back to probabilities. An example of such a transformation is $t(c_i) = \ln\frac{c_i}{1 - c_i}$, which is the inverse of the logistic function. In such a case, the regression model is trained on $(\mathbf{x}_i, t(c_i))$ pairs. Now, the learning problem becomes a linear regression problem:

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{N}\sum_{i=1}^{N} (\mathbf{w}^T\mathbf{x}_i - t(c_i))^2 + Q(\mathbf{w})$$

where $Q(\mathbf{w})$, similarly to the previous model, defines a regularization term. The solution $\mathbf{w}^*$ yields a weight vector defining the optimal discriminant function $f(\mathbf{x})$. We refer to this method as LogRaux (logistic regression with auxiliary information).

Using ranking to improve the noise tolerance

O'Hagan *et al*[13] surveyed various methods for eliciting the subjective uncertain assessments of physicians. They note that 'subjective probabilities can be well calibrated, but often they are not,' and cite many studies that revealed inaccurate judgments of uncertainty, including those of Poses *et al*,[14] Tierney *et al*,[15] and Dolan *et al*.[16] Since the above regression approach learns the model solely using the auxiliary probabilistic information, it may become very sensitive to the inconsistencies in the numerical assessments and as a result, its performance may deteriorate.

To address the problem, we propose to adapt ranking methods[9 10] that are more robust and better tolerate the noise in the estimates. Briefly, instead of relying strongly on exact probabilistic estimates, we try to model the relation between the two probabilistic assessments only qualitatively, in terms of pairwise order constraints.

Let $f: X \rightarrow R$ be a linear model $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ representing the ranking function that lets us order pairs of data points such that if instance $\mathbf{x}_i$ is ranked higher than $\mathbf{x}_j$, then $f(\mathbf{x}_i) > f(\mathbf{x}_j)$. The two data points are ordered according to the subjective probability c_i and c_j, hence we expect the ranking function to preserve their order. The learning to rank algorithms[9 10] let us find the ranking function f from the training data by minimizing the number of violated pairwise constraints between the data points. Now assume the ranking function defines a discriminant function that lets us discriminate between examples in class 0 and class 1. This formulation makes the learning of a discriminative model less dependent on the exact subjective value estimates that are used to induce the pairwise ordering. Hence we expect this relaxation would allow us to better absorb some amount of noise in subjective probability estimates, eventually leading to more robust learning algorithms.

Let r^* be our target ranking order determined by the probabilistic information c_i associated with each example. Then for every pair of examples $\mathbf{x}_i$ and $\mathbf{x}_j$: $(\mathbf{x}_i, \mathbf{x}_j) \in r^*$ if $c_i \geq c_j$, we can write a constraint $\mathbf{w}^T(\mathbf{x}_i - \mathbf{x}_j) \geq 0$ that we want the ranking

function $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ to satisfy. Similarly to the standard SVM learning problem formulation,[7] we allow some flexibility in building the solution $\mathbf{w}$ by adding slack variables ξ_{ij} representing penalties for constraints violation and a constant to regularize these penalties. We combine the constraints for satisfying ranking orders with constraints for satisfying binary class labels in one optimization problem. In particular, we propose to optimize:

$$\min_{w,} Q(\mathbf{w}) + A_1 \sum_{i=1}^{N} \eta_i + A_2 \sum_{i,j=1}^{N} \xi_{ij}$$

$$\forall i = 1 \ldots N : y_i(\mathbf{w}^T\mathbf{x_i} + b) \geq 1 - \eta_i$$

$$\forall (x_i, x_j) \in r^* : y_i\mathbf{w}^T(\mathbf{x_i} - \mathbf{x_j}) \geq 1 - \xi_{ij}$$

$$\forall i, j = 1, \ldots N : \eta_i > 0, \xi_{ij} > 0$$

where A_1 and A_2 are constants and $Q(\mathbf{w})$ is a regularization penalty. This formulation assumes two sets of constraints and two corresponding loss terms: $A_1 \sum_i \eta_i$ that defines the loss for not respecting binary labels, and $A_2 \sum_{i,j} \xi_{ij}$ that defines the loss for not respecting the orders induced by subjective probabilistic estimates. Solving this problem will give us the weight vector $\mathbf{w}$ and the discriminant function $f(\mathbf{x})$ that violates the smallest number of constraints. Note that by changing scaling constants A_1 and A_2, one can stress more either the label or the probabilistic order information. In general, the setting of these parameters is optimized using the internal cross-validation approach. We refer to this approach as SVMaux (SVM with auxiliary soft labels).

Algorithms for learning with auxiliary categorical labels

As mentioned earlier, auxiliary soft labels c_i may be present not only in the form of probabilistic estimates, but also in the form of qualitative ordinal categories. For example, we can ask an expert to assess the certainty in diagnosing a disease using four ordinal categories: 'strongly-disagree,' 'weakly-disagree,' 'weakly-agree,' and 'strongly-agree.' Since the auxiliary soft labels are no longer probabilities/numbers, the regression methods described in the previous section (LogRaux and LinRaux) cannot be applied directly. In this section we describe two methods that can accept auxiliary labels in the form of qualitative ordinal categories. We use the following notation: r to denote the number of ordinal categories and $\mathbf{C} = \{c_1, c_2, \ldots, c_r\}$ the set of ordinal categories.

Ordinal regression approach

Ordinal regression[17] is an approach which takes outputs, represented by an ordered set of categories, and constructs a regression function $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ that maps examples $\mathbf{x}$ onto a real line such that examples in each category project to a compact and well separated real-valued region. Assuming the ordered categories reflect the increased support for the binary label 1, the function f defines a discriminant function one can apply to the binary classification task. The method we propose in this work is based on the SVM approach for ordinal regression.[18]

Let $\mathbf{b} = \{b_1, b_2, \ldots, b_{r-1}\}$ denote the boundaries that separate categories in $\mathbf{C}$ on a real-valued line formed by f. For $\mathbf{x}$ that is assigned to category c_j we want f to satisfy $b_1 < \cdots < b_{j-1} < f(\mathbf{x}) < b_j < b_{j+1} < \cdots < b_{r-1}$. We encode these inequalities by a set of margin constraints. More specifically, we encode each inequality $f(\mathbf{x}) < b_k$ using a lower margin constraint: $f(\mathbf{x}) \leq b_k - 1$, and each inequality $f(\mathbf{x}) > b_k$ using an upper margin constraint: $f(\mathbf{x}) \geq b_k + 1$. Overall, for each boundary $b_1, \ldots, b_{r-1}$ there are N constraints, for the total of $(r-1) * N$ constraints. In addition to order constraints for all examples and boundaries of ordinal categories, we also include constraints for binary class labels, that is, if $y_i = 0$ and $y_j = 1$, then $f(\mathbf{x_i}) < f(\mathbf{x_j})$. Since the perfect discriminant function (satisfying all constraints) may not exist, we permit violations of constraints, but penalize them using sets of slack variables $\boldsymbol{\xi}$ (for ordinal constraints) and $\boldsymbol{\mu}$ (for binary class constraints). This leads to the following optimization problem:

$$\min_{\mathbf{w},d,\mathbf{b},\boldsymbol{\mu},\boldsymbol{\xi}} Q(\mathbf{w}) + A_1 \sum_{i=1}^{N} \mu_i + A_2 \sum_{i=1}^{N} \sum_{j=1}^{r-1} \xi_{ij}$$

$$\forall i = 1 \ldots N : y_i(\mathbf{w}^T\mathbf{x_i} + d) \geq 1 - \mu_i$$

$$\forall j = 1, \ldots, r-1;\ \forall (\mathbf{x_i}, c_i) \in \{(\mathbf{x_i}, c_i) | c_i \leq k\} : \mathbf{w}^T\mathbf{x_i} \leq b_j - 1 + \xi_{ij}$$

$$\forall j = 1, \ldots, r-1;\ \forall (\mathbf{x_i}, c_i) \in \{(\mathbf{x_i}, c_i) | c_i > k\} : \mathbf{w}^T\mathbf{x_i} \geq b_j + 1 - \xi_{ij}$$

$$\forall i = 1 \ldots N;\ \forall j = 1, \ldots, r-1 : \xi_{ij} \geq 0$$

where $Q(\mathbf{w})$ is a regularization term, and A_1 and A_2 are constants that define the trade-off between penalty for violating class-label constraints and penalty for violating categorical order constraints, respectively. We refer to this method as SVMaux_cat (SVM with auxiliary categorical labels).

Linear regression with local search

The idea of this method is to first convert the ordinal categories to real values and then apply a regression model to learn the discriminant function. To preserve the ordering of categories we seek a mapping $h : \mathbf{C} \rightarrow R$ that assigns each category to a real number such that $h(c_i) < h(c_j)$ where $i < j$. In addition, we want this mapping to yield the optimal classification performance once it is regressed using a regression model f.

We find the optimal mapping h using the following local search algorithm. In the initial step, the algorithm assigns all ordinal categories $\{c_1, c_2, \ldots, c_r\}$ to numbers in interval [0,1], such that $h(c_i) = \frac{i-1}{r-1}$. This assignment distributes all ordinal categories uniformly across [0,1] while preserving their order. After that, we run a local re-optimization procedure which repeatedly relaxes and re-optimizes the mapping $h(c_i)$ for $i = 2, \ldots, r-1$ one at a time, while keeping the rest of the mappings fixed. The optimization of $h(c_i)$ is implemented as a line search restricted to the subinterval $[h(c_{i-1}), h(c_{i+1})]$. The area under the receiver operating characteristic curve (AUC)[19] based on the threefold cross-validation is used to measure the quality of each mapping. The procedure stops if the improvement in the AUC after we re-optimize all $h(c_i)$ for $i = 2, \ldots, r-1$ is smaller than 0.01. We refer to this method as

LinRaux_localsearch (linear regression with auxiliary information and local search).

Experiments

We demonstrate the benefits of the proposed methods on the problem of monitoring the risk of HIT.[1] HIT is an adverse immune reaction that may develop if a patient is treated with heparin for a long time. If the condition is not detected and treated promptly, it may lead to further complications, such as thrombosis, and even death. An important problem is the monitoring and detection of patients who are at risk of developing the condition. We investigate the possibility of building an HIT alert model from patient data using assessment of the risk of HIT by an expert. This corresponds to the problem of learning a binary classification model from data.

HIT data

The data used in the experiments were extracted from the Post-Surgical Cardiac (PCP) database[20 21] of the EHRs of 4486 post-surgical cardiac patients. The extracted data consisted of over 51 000 patient-state instances obtained from EHRs using the 24 h segmentation procedure proposed in Hauskrecht *et al*.[20] Out of these we selected 377 instances that were labeled (independently) by three clinical pharmacists with respect to HIT. The instances for the study were selected using a special stratified sampling approach aimed at increasing the proportion of HIT alert instances the expert would agree with in the dataset. For example, one stratum with a larger proportion of positives was built using patient instances for which an HPF4 test (ie, used to confirm the HIT) was ordered in the next 24 h. The strata covered the full instance space and the sampling was biased to strata with expected higher proportions of positive instances. All 377 examples sampled by this procedure were recorded with weights reflecting how likely they would be if they were obtained by an unbiased random sampling process. The weights let us correct biases due to stratified selection. Please note that the reason for introducing stratified sampling for labeling the patient instances was that the data and their labels were intended also for other related projects and the aim of one of them was to obtain and analyze a larger sample of positive HIT alerts. Given the fixed review budget, the stratification of patient instances and stratified sampling was the best option to achieve our goals.

HIT data assessment

For each patient instance, we asked three experts in clinical pharmacy the following two questions:

Question 1: How strongly does the clinical evidence indicate that the patient is at risk of HIT? The answer was as a numeric score in the range from 0 to 100, which we interpreted as a probabilistic score by converting it to interval [0,1].

Question 2: Assume you have received an HIT alert for this patient. Please indicate to what extent you agree/disagree with the alert? The answer was one of the four ordinal categories: 'strongly-disagree,' 'weakly-disagree,' 'weakly-agree,' and 'strongly-agree.'

We used answers to question 2 to define the binary class label 'Agree with alert.' Briefly, the label is positive if the expert agrees (weakly or strongly) with the alert in question 2, otherwise it is negative. Our ultimate objective was to build a classification model that predicts this binary class label. Answers to questions 1 and 2 can be then viewed as its soft-label refinement: answers to question 1 define probabilistic labels, and answers to question 2 ordinal category labels.

In order to make the qualified judgment, the experts were able to see EHRs up to the time of the alert assessment, which is what the experts would see if the instances were encountered prospectively. Questions 1 and 2 were asked, and the answers to both questions were recorded on the same electronic input form. The answers were submitted at the same time by the expert by pressing the submit button on the form. We note that inclusion of both questions on the same form and their specific ordering could have influenced their respective answers. However, at this point we do not have any evidence for or against to believe this skewed the results in any significant way.

Table 1 gives basic statistics of the review process and answers for the three experts. The last column shows the probability of a positive alert normalized by taking into account the data weights. Table 2 shows the basic statistics related to auxiliary soft-label information collected from the experts and used in the experiments. The second and third columns are related to probabilistic label information, and the remaining columns to the ordinal category labels. We see that while the strong HIT alert (strong positive) option was used only rarely by the experts, strong and weak no-alerts are quite frequent and likely to be useful for learning a good discriminative model.

Online supplementary appendix 3 gives detailed agreement matrices for all pairs of experts, and their corresponding agreement, κ[22] and weighted κ[23] statistics. Fleiss' κ[24] statistic is used to assess the agreement for all three experts combined. Briefly, for binary labels the pairwise κ ranges between 0.33 and 0.57, while Fleiss' κ is 0.47. For ordinal categorical labels, the weighted κ for pairwise analysis ranges between 0.31 and 0.51, and Fleiss' κ is 0.34.

Data features

Medical records consist of complex time-series data. Extracting temporal information from the time-series that is useful for building a good classification model is a challenging task.[25–29] For the purpose of this study, in order to build classification models, we represented each patient-state instance using 50 features derived from EHRs using extraction routines from Hauskrecht *et al*[21] and Valko and Hauskrecht[30] that convert time-series data for different clinical variables to fixed feature sets. More specifically, we applied these routines to generate features for platelet counts, hemoglobin levels, white blood cell counts, heparin administration record, and major heart surgeries that are important for HIT detection. Examples of generated features are last observed platelet value, last platelet trend, and the length of time the patient is on heparin medication. Online supplementary appendix 1 gives a complete list of the clinical variables and features used in our models.

Experimental setup

We evaluated the performance of the soft-label classification methods using the AUC score.[19] To perform the evaluation, the data set of 377 examples with weights (reflecting the

Table 1 Basic statistics for the review process for experts 1, 2 and 3

	Mean time (SD) to review a case (s)	Number of positives	Number of negatives	Normal probability of a positive alert*
Expert 1	77.23 (111.54)	88	289	0.06
Expert 2	77.16 (81.47)	53	324	0.02
Expert 3	110.27 (178.31)	106	271	0.15

*Values in this column are calculated by normalizing the answers with respect to the data weights.

Table 2 Basic statistics for the auxiliary soft-label information collected from the experts

	Mean (SD) probability for positives	Mean (SD) probability for negatives	Number of strong positives	Number of weak positives	Number of weak negatives	Number of strong negatives
Expert 1	0.51 (0.16)	0.14 (0.17)	3	85	141	148
Expert 2	0.59 (0.06)	0.28 (0.16)	0	53	235	89
Expert 3	0.48 (0.13)	0.39 (0.14)	1	105	181	90

The second and third columns show the distribution of probabilities for positive and negative examples. The remaining columns show the counts of strong and weak subcategories for positive and negative examples.

stratification) was first split randomly (by ignoring the weights) using a 2:1 ratio into the disjoint training and testing groups. This split assured that instances in the training group were used for training the models, while instances in the testing group were used for their evaluation. The weights associated with the instances were then used to subsample the two groups in order to generate the training and test sets. This process lets us generate un-biased and non-overlapping training and testing datasets. The process (data splitting and subsampling of data according to weights) was repeated 100 times and the average AUC and 95% CI on the test set were calculated.

To investigate the impact of the soft-label information on the sample complexity, we trained the new models that accept soft-label information with data of different sizes and compared them to models that learn from the binary 'Agree with alert' and 'Disagree with alert' labels only. In all experiments, the soft-label information was used only to aid the training (learning) process, and it was never used to test the binary model. The training sizes were varied from 20 to 250. The upper limit (250) was chosen because the performance of the methods at that point stabilized and further increases in the training size did not result in any significant changes. We performed two experiments testing the impact of learning with auxiliary probabilistic labels (experiment 1) and ordinal categorical labels (experiment 2). Table 3 summarizes all methods in the experiments. An L1 regularization penalty was used to implement Q(**w**) for all models. Please note that because the number of samples in the strong positive subcategory is very small, the multiclass method is trained only on three subcategories (positives, weak negatives, and strong negatives).

RESULTS AND DISCUSSION

The results of all our experiments are shown graphically in figures 1A–C and 3A–C. In addition, the same results are tabulated together with the pairwise statistical significance test in online supplementary appendix 2.

Experiment 1: learning with probabilistic labels

Figure 1A–C compares the performance of different methods on HIT data and soft labels expressed in terms of probabilistic scores (answers of the experts to question 1 for expert 1, expert 2, and expert 3, respectively). The x-axis shows the number of examples the models are trained on and the y-axis shows the AUC. The baseline methods (SVM and LogR) are illustrated using dashed lines. Methods that utilize soft labels are shown using solid lines.

The results in figure 1 show that the two simple approaches utilizing the probabilistic information (LinRaux and LogRaux) are worse than classification models (SVM and LogR) that learn from the binary label information only. This is true for all three experts. The new ranking approach (SVMaux) that ignores exact probabilistic assessments but at the same time tries to preserve the relative order of patient instances, performs the best and outperforms all alternatives on two of the experts (experts 1 and 2) and is comparable to binary classifiers for expert 3.

Discussion

The LinRaux and LogRaux methods that attempt to fit probabilistic estimates to define the discriminant function perform the worst. At first glance this is somewhat surprising. However, these results can be explained by inconsistencies and biases in subjective probability estimates provided by the experts. The fact that subjective probability estimates are often not well calibrated is a widely documented problem in the literature.[14–16] Hence, it is not realistic to expect that probabilistic assessments for all instances are perfect both in absolute terms (ie, each probability assessment is a perfect estimate of the true probability) and in relative terms (when pairs of instances and their probability differences are considered). This in turn may influence the model based on such assessments, especially when the model tries to 'closely' fit the estimates. Figure 2 illustrates this problem on the data in our study. It shows the distribution of probability estimates for positive and negative examples for

Table 3 Summary of all methods used in the experiments

Labels used	Method	Short description	Experiments 1	Experiments 2
Binary labels	LogR (baseline)	Standard binary logistic regression	●	●
	SVM (baseline)	Standard binary SVM	●	●
Binary and auxiliary probabilistic labels	LinRaux	Linear regression with probabilistic labels	●	
	LogRaux	Logistic regression with probabilistic labels	●	
	SVMaux	Rank-based SVM with probabilistic labels	●	
Binary and auxiliary ordinal categories	Multiclass (baseline)	Soft-max regression model[31] that learns from categorical labels but ignores the ordinal information		●
	LinRaux_localsearch	Linear regression with ordinal categories and local search		●
	SVMaux_cat	SVM regression with ordinal categories		●

SVM, support vector machines.

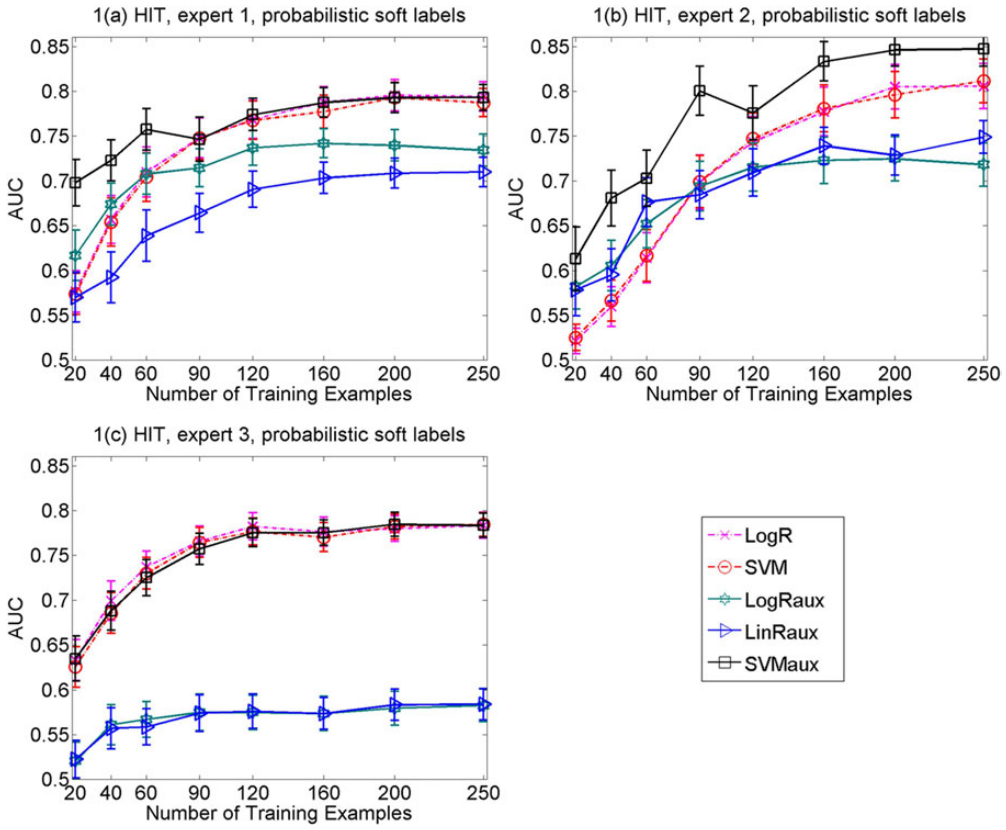

Figure 1 Area under the receiver operating characteristic curve (AUC) for the different learning methods trained on probabilistic soft labels from three different experts and for the different training sample sizes.

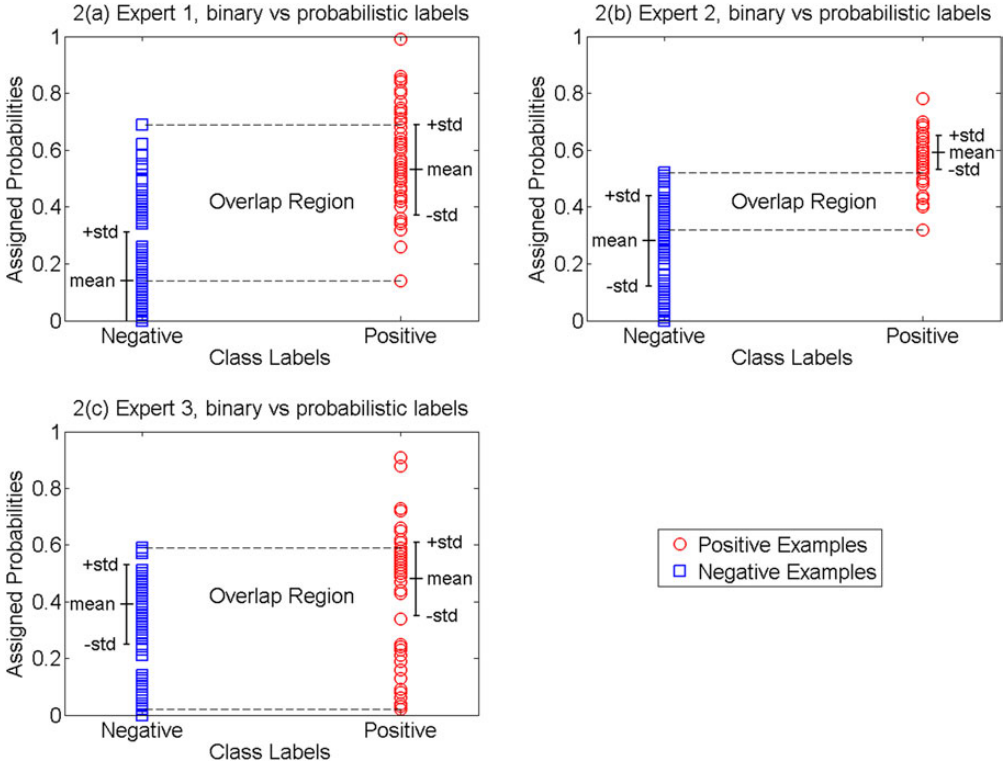

Figure 2 Distribution of probabilities assigned to negative (left) and positive (right) examples by experts 1, 2, and 3, respectively. Left and right vertical bars show the mean and SD of assigned probabilities to negative and positive examples, respectively. The 'overlap region' between horizontal dash lines is where probability estimates for positive and negative examples overlap. These inconsistencies may lead to deterioration of models trained based on such probabilities.

experts 1, 2, and 3, and overlaps of the two regions. These overlaps and inconsistencies influence the discriminatory performance of the models trained on probabilistic soft labels and translate to the results observed in figure 1. More specifically, notice that while the probability estimates of expert 2 assigned to positive and negative labels in figure 2 are rather well separated, the estimates of expert 1 and particularly of expert 3 are much harder to separate, and there is a great deal of overlap in the regions defining positive and negative labels. These differences translate to the results in figure 1. In particular, LinRaux and LogRaux that learn their models only from probability estimates get more benefit from probabilistic information provided by expert 2 than by expert 1 or expert 3. SVMaux (our top performing model) that relies on both probability estimates and binary labels, is more robust and is able to absorb the inconsistencies much better. However, please notice that for expert 3 (with the most inconsistent probability assignments), the benefit of probabilistic information even for this model diminishes and the model is comparable to models one can learn from the binary information only.

In summary, the new ranking approach (SVMaux) that ignores exact probabilistic assessments, but at the same time preserves the relative order of patient instances, is more robust to noisy probabilistic estimates and outperforms all alternatives, hence demonstrating the benefit of soft probabilistic labels for learning the classification models.

Experiment 2: learning with ordinal categories

Figure 3 compares the performance of the methods on HIT data and the soft labels expressed in terms of ordinal categories: 'strongly-disagree,' 'weakly-disagree,' 'weakly-agree,' and 'strongly-agree' with the HIT alert. Figure 3A–C shows the AUC of models learned for expert 1, expert 2, and expert 3, respectively.

The linear regression aided by the local search is the best method for all three experts, followed by the SVM ranking method modified to ordinal categories. The two baselines that rely on the binary class information are comparable, and come next. Finally, the multiclass learning method that tries to learn different categories but ignores their order, is the worst and is outperformed by all other methods.

Discussion

The results demonstrate that binary labels, when they are further refined to ordinal subcategories, can lead to improved classification models. In particular, splitting positive and negative labels into strong and weak positives and negatives, and mapping them on the same discriminative function while assuring their order helped us to converge faster to a better discriminative function. However, we also observe that when these subcategories are taken in isolation (without ordering) as in the multiclass method, the performance tends to be worse. This can be attributed to the fact that multiclass models require more parameters and in general more samples are needed to fit them accurately.

The refinement of the two alert classes to four ordinal subcategories clearly helped us to learn better models for all three experts. In contrast to this, the probabilistic information in experiment 1 was helpful, but the margin of the improvement was much smaller, and for expert 3, who was the least consistent in assigning probabilities to examples, the benefit was marginal. Our experiments with probabilistic assessments (experiment 1)

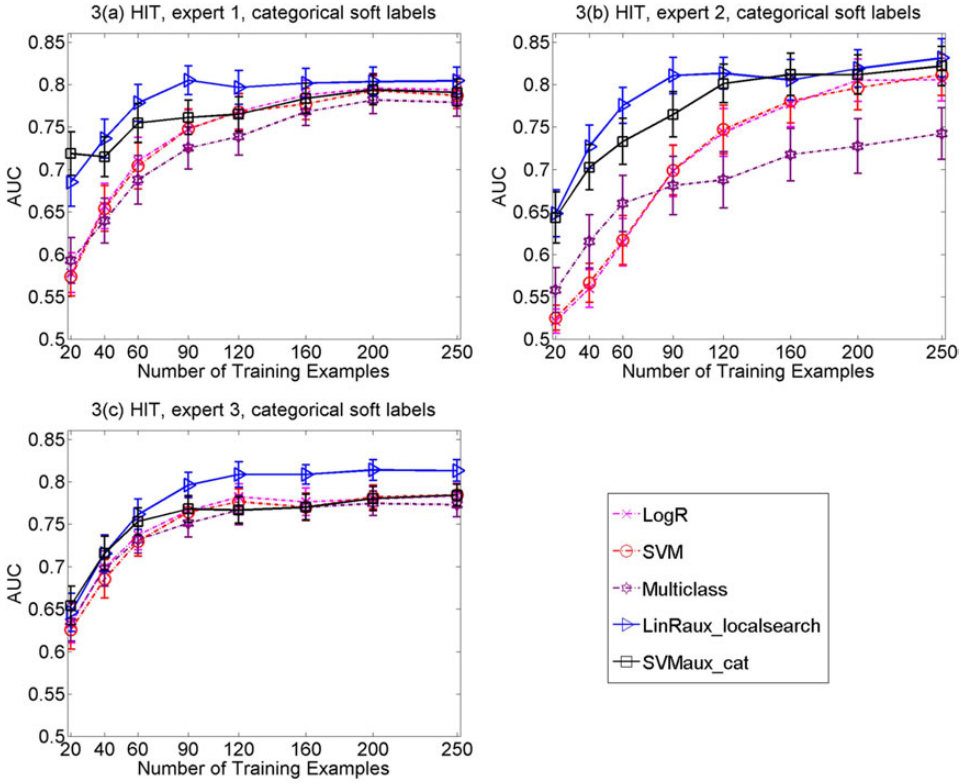

Figure 3 Area under the receiver operating characteristic (ROC) curve (AUC) for the different learning methods trained on ordinal categorical labels from three different experts and for the different training sample sizes.

given by humans suggest the assessments may suffer from consistency/calibration problems which may reduce the utility of the probabilistic information for aiding the learning process. This also suggests that the utility of soft labeling may differ and vary with the resolution and the number of soft categories the expert may choose from. An interesting and open question is how many categories to use in order to benefit from the soft labeling the most.

CONCLUSION

Making use of real-world data sets often prompts one to fill additional information with subjective human labels. However, this process is often time-consuming, so different ways of reducing the labeling effort need to be sought. In this work we investigate a new framework for reducing this cost by using auxiliary soft labels that reflect how strongly the human expert believes in the class label, which can be extracted quickly and with virtually no additional time effort.

We proposed and studied different methods that incorporate soft labels into the classification learning process. The experimental results on HIT data show that our methods outperform traditional binary classifiers, which supports our hypothesis that auxiliary soft-label information may lead to improved learning efficiency. We have also found that soft-label information expressed in terms of four subcategories helped the models the most, and the improvement was more consistent than for models aided with soft probabilistic information. An interesting open research question is whether this is the optimal number of soft categories one can use in order to learn better classification models.

Finally, we would like to note that soft-label approaches can be used to learn discriminative models even in the case when instances for one of the two target classes are not observed in the training data. This feature is especially useful when classes are unbalanced and the chance of observing the minority class in a random sample is low.

Acknowledgements We would like to thank all the reviewers and the associate editor for their helpful comments and feedback on the first version of this manuscript.

Contributors All three authors listed on the paper satisfy the ICJME criteria for authorship and all authors contributed to (1) the conception and design of the new methodology, and analysis and interpretation of data, and (2) writing and revising the article, and (3) approved the revised version submitted.

Funding This work was supported by National Institutes of Health (NIH) grants R01-LM010019 and R01-GM088224.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Warkentin T, Sheppard J, Horsewood P. Impact of the patient population on the risk for heparin-induced thrombocytopenia. *Blood* 2000;96:1703–8.
2 Cohn D, Atlas L, Ladner R. Improving generalization with active learning. *Machine Learn* 1994;15:201–21.
3 Lewis D, Gale W. A sequential algorithm for training text classifiers. *SIGIR*; 1994:3–12.
4 Seung H, Opper M, Sompolinsky H. Query by committee. *COLT*; 1992:287–94.
5 Smyth P, Fayyad U, Burl M, *et al*. Learning with probabilistic supervision. *COLT*; 1995;vol 3:163–82.
6 McCullagh P, Nelder J. *Generalized linear models*. Chapman & Hall, 1989.
7 Cortes C, Vapnik V. Support-Vector Networks. *Machine Learn* 1995;20:273–97.
8 Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B* 1996;58:267–88.
9 Herbrich R, Graepel T, Obermayer K. Support vector learning for ordinal regression. *ICANN*; 1999:97–102.
10 Joachims T. Optimizing search engines using clickthrough data. *SIGKDD*; 2002:133–42.
11 Furnkranz J, Hullermeier E. *Preference learning*. Springer, 2010.
12 Hoerl A, Kennard R. Ridge regression—advances, algorithms, and applications. *Am J Math Manage Sci* 1981;1:5–83.
13 O'Hagan A, Buck C, Daneshkhan A, *et al*. *Uncertainty judgements eliciting experts' probabilities*. John Wiley and Sons, 2007.
14 Poses R, Cebul R, Collins M, *et al*. The accuracy of experienced physicians' probability estimates for patients with sore throats. *JAMA* 1985;254:925–9.
15 Tierney W, Fitzgerald J, McHenry R, *et al*. Physicians' estimates of the probability of myocardial infarction in emergency room patients with chest pain. *Med Decis Making* 1986;6:12–17.
16 Dolan J, Bordley D, Mushlin A. An evaluation of clinicians' subjective prior probability estimates. *Med Decis Making* 1986;6:216–23.
17 McCullagh P. Regression models for ordinal data. *J R Stat Soc Ser B* 1980;42:109–42.
18 Chu W, Keerthi S. New approaches to support vector ordinal regression. *ICML*; 2005:145–52.
19 Hanley J, McNeil B. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;1:29–36.
20 Hauskrecht M, Valko M, Batal I, *et al*. Conditional outlier detection for clinical alerting. *AMIA Annu Symp Proc* 2010:286–90.
21 Hauskrecht M, Batal I, Valko M, *et al*. Outlier detection for patient monitoring and alerting. *JBI* 2013;46:47–55.
22 Cohen J. A coefficient of agreement for nominal scales. *EPM* 1960;20:37–46.
23 Cohen J. Weighed kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 1968;70:213–20.
24 Fleiss J. Measuring nominal scale agreement among many raters. *Psychol Bull* 1971;76:378–82.
25 Hauskrecht M, Fraser H. Modeling treatment of ischemic heart disease with partially observable markov decision processes. *AMIA Annu Symp Proc* 1998:538–42.
26 Batal I, Sacchi L, Bellazzi R, *et al*. *Multivariate time series classification with temporal abstractions*. FLAIRS, 2009.
27 Combi C, Keravnou-Papailiou E, Shahar Y. *Temporal information systems in medicine*. Springer, 2010.
28 Batal I, Valizadegan H, Cooper G, *et al*. A pattern mining approach for classifying multivariate temporal data. *BIBM*; 2011:358–65.
29 Batal I, Fradkin D, Harrison J, *et al*. Mining recent temporal patterns for event detection in multivariate time series data. *SIGKDD*; 2012:280–8.
30 Valko M, Hauskrecht M. Feature importance analysis for patient management decisions. *Med Info* 2010;160:861–5.
31 Lin C, Weng R, Keerthi S. Trust region Newton method for logistic regression. *JMLR* 2008;9:627–50.