Sample-efficient learning with auxiliary class-label information

Quang Nguyen[1], Hamed Valizadegan[1], Amy Seybert[2], Milos Hauskrecht[1]
[1] Department of Computer Science, [2] School of Pharmacy, University of Pittsburgh,
email: *milos@pitt.edu*

Abstract

Building classification models from clinical data collected for past patients often requires additional example labeling and annotation by a human expert. Since example labeling may require to review a complete electronic health record the process can be very time consuming and costly. To make the process more cost-efficient, the number of examples an expert needs to label should be reduced. We develop and test a new approach for the classification learning in which, in addition to class labels provided by an expert, the learner is provided with auxiliary information that reflects how strong the expert feels about the class label. We show that this information can be extremely useful for practical classification tasks based on human assessment and can lead to improved learning with a smaller number of examples. We develop a new classification approach based on the support vector machines and the learning to rank methodologies capable of utilizing the auxiliary information during the model learning process. We demonstrate the benefit of the approach on the problem of learning an alert model for Heparin Induced Thrombocytopenia (HIT) by showing an improved classification performance of the models that are trained on a smaller number of labeled examples.

1 Introduction

The vast amounts of clinical data collected, stored and later archived in electronic health records (EHRs) today provide us with an excellent opportunity to better understand the disease, its dynamics, the efficacies of different treatments, and may eventually lead to new computer models with a potential to impact and improve the decision-making and patient management processes. However, the EHR data archived in practice are often not complete and ready to be applied to a specific problem, and additional human expert assessment or annotation of data and patient cases may be needed before the analysis can be conducted and appropriate models can be built.

Take for example the problem of building a monitoring and alerting system that aims to detect a risk of some adverse condition with the help of data. While some of the temporal data (such as lab test time-series, or medications given) are often archived and collected, the diagnoses or occurrences of some adverse event are either not recorded at all or their record is atemporal and it is not clear at what time or during what time interval the event occurred. Hence, if our goal is to analyze these conditions and build models that are able to predict them, individual patient instances must be first labeled by an expert or a group of experts.

The process of labeling (annotating) patient instances using subjective human assessments can be an extremely time-consuming and costly process, since it requires one to review large amount of information in the EHR. Optimizing the time and cost of this process boils down to reducing the number of examples one must assess. One direction to address this problem explored extensively by the machine learning community in recent years is to develop active learning[1] methods that analyze examples, prioritize them and select those that are most critical for the task we want to solve, while optimizing the overall data labeling cost.

In this work, we explore an alternative solution that is orthogonal to the active learning approach and may alleviate the costly example labeling process in practice. The idea is based on a simple premise, the human expert that gives us a subjective class label (detect or not detect, alert or do not alert) can often provide us with auxiliary information related to the case which reflects his or her confidence about the label or belief about the underlying condition. The acquisition of this additional information often comes at a cost that is insignificant when compared to the cost of the case review and label assessment. To illustrate this point, assume an expert reviewing patient data in order to assess if the alert for some adverse condition is appropriate or not. Clearly the complexity of the data in EHR prompts the expert to spend a large amount of time reviewing and analyzing the case (typically minutes). However, once the decision about alerting or not alerting on some adverse condition is made, it is often possible to refine the decision with additional information related to the underlying adverse condition reflecting how strongly the reviewer believes the condition occurs or how strong the alert should be.

The auxiliary information a human expert can provide in addition to the class label can be represented and acquired in different ways. One possibility is to use numerical values representing directly the chance (or probability) the patient suffers from the target condition, another possibility is to use ordinal categorical values representing qualitative assessment of the belief using a finite number of categories (e.g strong disbelief, weak disbelief, weak belief the adverse condition is present, etc.). Our objective is to develop a framework in which a classification model (that assigns class labels to patient cases) can be learned more efficiently with a smaller number of labeled examples with the help of this auxiliary information.

We develop and present a new learning method based on the support vector machine framework that lets us incorporate the auxiliary information in terms of order constraints. Briefly, our method learns the discriminative boundary by assuring that examples with a higher belief are projected further away from the decision boundary than examples with weaker beliefs. Conceptually, our method draws upon the results and research on the learning to rank problems[2,3] investigated in various information retrieval applications. We test our method on the problem of learning a classification model for generating Heparin induced thrombocytopenia (HIT)[4,5,6] alerts for post surgical cardiac patients. We show that with additional auxiliary information we are able to learn alert models with a smaller number of examples than with just the alert label information. Moreover we show that we are able to learn a model even if the model is trained with examples from just one class, which can be extremely important for problems in which the prevalence of the two classes in the population and data is highly unbalanced.

2 Problem description

We aim to learn a binary classifier $f : X \rightarrow Y$ where X is a feature vector and Y is a binary label, i.e. $Y \in \{0, 1\}$. In the common setting, a set of examples $\mathbf{x}_1, \mathbf{x}_2, .., \mathbf{x}_N$ and their binary labels $y_1, y_2, .., y_N$ are provided for training. Here we assume that we also have access to additional information: a score p_i reflecting one's belief the example $\mathbf{x}_i$ belongs to class 1. p_i can be a discrete score $p_i \in \{0, 1, 2, .., k\}$ or a probability number $p_i \in [0, 1]$. Hence each data entry in the data set $D = \{d_1, d_2, \cdots d_N\}$ consists of three components: $d_i = (\mathbf{x}_i, y_i, p_i)$, an input, a class label and a belief score for class 1.

The belief information can be often obtained when labels are acquired from human assessment. 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 score captures the physician's belief the patient indeed suffers from the condition. The cost of obtaining this additional information is typically small once the patient case is reviewed and assessed by the expert.

Despite possible noise in the human-based assessment, a discrete class label y_i and the score p_i are closely related. Adopting a decision-theoretic perspective, we assume the class label y_i is a function (although unknown) of the belief score.

Our main conjecture in this work is that additional belief information can help us to learn a classifier more efficiently and with a smaller number of examples. This can be particularly useful when the data is unbalanced (the prior probability of one of the classes is small), and when the number of labeled examples is limited.

Surprisingly, not much prior research work has been done combining the class labeling and related belief information. Perhaps the closest to our framework is the research by[7,8] who considers probabilistic information as a vital component of the learning process because of the ambiguities in the class labeling. This work applies the approach to classification of volcanos from radar images of distant planets. The differences from our framework are: they rely only on the probabilistic information to build the models, class labels are ignored; only classification models based on simple neural network and probabilistic models are considered; they make no attempt to correct for the variations and noise in subjective estimates.

3 Learning with binary labels

Before developing the methodology capable of utilizing auxiliary information we briefly review one of the most widely used methods for learning binary classification models: the support vector machine (SVM)[9]. The main reason for this

review is that our solution builds directly upon and extends this methodology.

The support vector machine is an instance of a (discriminative) classification method[10], that aims to learn a function $f : X \rightarrow \mathcal{R}$ that discriminates examples from the two classes. Once the function f is known, the class decision is made with the help of a threshold σ such that for values $f(\mathbf{x}) \geq \sigma$ we classify the example as class 1, otherwise we classify it as class 0.

The (linear) SVM is popular in the machine learning community primarily thanks to its ability to learn high-quality discriminative patterns in high-dimensional datasets. Among many linear decision boundary that can separate the examples from two classes, the linear SVM chooses one that has the maximum margin. The margin is defined as the distance of the decision line to its nearest examples. For two classes that are linearly separable, SVM has the following form:

$$\begin{aligned} \min_{\mathbf{w},b} \quad & Q(\mathbf{w}) \\ \text{subject to:} & \\ \forall \mathbf{x}_i,\ y_i \quad & \mathbf{w}^T \mathbf{x}_i y_i + b \geq 1 \\ \forall i: \quad & \eta_i \geq 0 \end{aligned}$$

where $i = 1, 2, ..., N$ indexes examples, $Q(\mathbf{w})$ a regularization penalty, typically $\frac{1}{2}\mathbf{w}^T\mathbf{w}$. Solving this problem will give us the weight vector $\mathbf{w}$ and the discriminant function $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ that that has the maximum margin. When the two classes are not linearly separable, slack variables are utilized and the SVM takes the following form:

$$\begin{aligned} \min_{\mathbf{w},b} \quad & Q(\mathbf{w}) + C \sum_i \eta_i \\ \text{subject to:} & \\ \forall \mathbf{x}_i,\ y_i \quad & \mathbf{w}^T \mathbf{x}_i y_i + b \geq 1 - \eta_i \\ \forall i: \quad & \eta_i \geq 0 \end{aligned}$$

where C is a constant. This form is called the soft-margin case and allows violating some of the difficult constrains. After computing $\mathbf{w}$ and b in the above formulation, one can classify new examples by either inspecting the sign of $\mathbf{w}^T\mathbf{x} + b$, or by using a threshold σ for defining the class decision.

4 Learning with auxiliary belief information

In this section we develop classification learning algorithms that let us accept and learn from the auxiliary belief labels. We start with a simple and straightforward approach by utilizing regression to learn from the auxiliary belief labels. After that we adapt a well-known ranking algorithm, the Rank-SVM[11], that utilizes both binary and probabilistic labels when constructing the model.

4.1. A simple regression approach

In the standard binary classification setting (see above), the discriminant function is learned from examples with class labels ($\{0, 1\}$) only. In our framework, in addition to class labels, we have access also to auxiliary belief information associated with these class labels. The question is how this information can be used to learn a better model. One relatively straightforward solution is to regress a function f where (x_i, p_i) are the input-output pairs. Assuming the function $f : X \rightarrow R$ is formed by a linear model $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$, the learning problem becomes a linear regression problem solved by minimizing the error function based on the sum of squared residuals.

$$Error(D, \mathbf{w}) = \frac{1}{N} \sum_{i=1}^{N} \left(\mathbf{w}^T \mathbf{x}_i - p_i\right)^2 \tag{1}$$

The solution $\mathbf{w}^* = \arg\min Error(D, \mathbf{w})$ yields a weight vector optimizing the linear model.

Defining the classification threshold. Once the discriminant function is learned a classifier can be defined using a decision threshold σ. To find the optimal threshold, we use true class labels (binary labels) and minimize the overall loss in the training data.

Regularization. The regression methods are quite common and can be enriched with different bells and whistles that let it perform better in different settings. In our data the primary concern is the dimensionality of $\mathbf{x}$ and the number of samples N in the data set. Briefly, if the dimensionality of $\mathbf{x}$ is high and the number of examples in N is small, the possibility of the model overfit. In such a case we can modify and improve the performance of the regression model using one of the regularization approaches, such as the Ridge (or L_2) regularization[12], lasso (or L_1) regularization [13,14], or their elastic network combination[15]. Briefly, the optimization in Equation 1 is modified as the following using the regularization:

$$\mathbf{w}^* = \arg\min \mathbf{Error}(\mathbf{D}, \mathbf{w}) + \mathbf{Q}(\mathbf{w}) \tag{2}$$

such that $Q(\mathbf{w})$ is a regularization penalty. Examples of regularization penalties are: $Q(\mathbf{w}) = \lambda|\mathbf{w}|_1$ for the L1 (lasso) regularization, or $Q(\mathbf{w}) = \lambda|\mathbf{w}|_2$ for the L2 (ridge) regularization.

Sensitivity to the noise in subjective estimates. Learning a regression function directly from auxiliary belief information raises a concern of what happens if these subjective probabilistic assessments are not consistent and subject to noise due to inaccurate subjective human estimates. Clearly, if the estimates differ widely one expects them to impact the quality of the discriminant function. Intuitively, if the noise is too strong, the benefit of auxiliary probabilistic information disappears and the binary label information may become more reliable when learning a classification model.

Another problem with learning a regression function from the belief scores results from the characteristics of the belief scores when they are provided in the form of discrete numbers $p_i \in \{0, 1, 2, \ldots, k\}$. In such cases, the belief information are in form of ordinal class label whose absolute numerical value does not necessary carry meaningful information; i.e. the distance between 0 and 1 might not be the same as the distance between 1 and 2, and etc. This problem is well-known in ranking[3] when the relevancy scores are considered as absolute numerical values.

4.2. Using ranking to improve the noise tolerance

As mentioned, the regression approach introduced in Section 4 learns the model by relying on the numeric value of auxiliary probabilistic information. As a result it may become very sensitive to the noise and inconsistencies in the numerical assessments. Since humans are not very good in providing well calibrated probabilistic estimates[16,17], the deterioration of the performance due to the noise becomes an important issue and methods that are more robust to this noise must be used to alleviate the problem.

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

Let $f : X \rightarrow \mathcal{R}$ be a linear model $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ that discriminates between examples in class 0 and class 1. f can also represent a linear ranking function that order individual data points such that if the instance $\mathbf{x}_1$ is ranked higher than $\mathbf{x}_2$ then $f(\mathbf{x}_1) > f(\mathbf{x}_2)$. Now assuming any two data points $\mathbf{x}_1$ and $\mathbf{x}_2$ are ordered according to their subjective belief scores p_1 and p_2, we expect the ranking function to preserve their order.

The learning to rank algorithms[2] construct a ranking function from the training data by minimizing the number of violated pairwise constrains between the data points and the amount of these violations. Such a formulation of a learning problem makes the problem of learning the discriminative model less dependent on exact subjective value estimates that are used to induce the pairwise ordering. Hence we hope this relaxation provides a tool to better absorb some amount of noise in the subjective probability estimates, eventually leading to more robust learning algorithms.

Let r^* be the target ranking order determined by the belief information p_i associated with each example. Then for every pair of examples $\mathbf{x}_i$ and $\mathbf{x}_j$ that $p_i > p_j$ we say $(\mathbf{x}_i, \mathbf{x}_j) \in r^*$ we can write a constrain $\mathbf{w}^T(\mathbf{x}_i - \mathbf{x}_j) > 0$ that the ranking function $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ needs to satisfy. Just, like in the classification SVM, we allow some flexibility in building the hyperplane by adding slack variables $\xi_{i,j}$ representing penalties for the constraint violation and a constant

C to regularize these penalties. Now the learning-to-rank of N examples is equivalent to the following optimization problem:

$$\begin{aligned} \min_{\mathbf{w}} \quad & Q(\mathbf{w}) + C\sum_{i,j}\xi_{i,j} \\ \text{subject to:} & \\ \forall(\mathbf{x}_i,\mathbf{x}_j)\in r^*: \quad & \mathbf{w}^T(\mathbf{x}_i-\mathbf{x}_j)\geq 1-\xi_{i,j} \\ \forall i\forall j: \quad & \xi_{i,j}\geq 0 \end{aligned}$$

where $i,j=1,2,\ldots,N$ indexes examples, $Q(\mathbf{w})$ is the regularization penalty similar to SVM, and C is a constant. Solving this problem will give us the weight vector $\mathbf{w}$ and the discriminant function $f(\mathbf{x})=\mathbf{w}^T\mathbf{x}$ that violates the smallest number of constraints.

4.3. Optimizing the discriminant function by weighting the importance of belief information

As mentioned, the absolute numerical values provided by p_is are not meaningful and only the relative magnitude of p_is is important. However, in the previous section, we only considered the order information provided by the belief scores and ignored the relative magnitude of the belief assessment. To emphasize the importance of the relative magnitude provided by the reviewer, we recommend to give more weights to those pairs that have bigger difference in their belief assessment; in other word, we weight examples pair xi and x_j in r^* by p_i-p_j difference normalized to interval $[0,1]$ and get the following new objective function.

$$\begin{aligned} \min_{\mathbf{w}} \quad & Q(\mathbf{w}) + C\sum_{i,j}(p_i-p_j)\xi_{i,j} \\ \text{subject to:} & \\ \forall(\mathbf{x}_i,\mathbf{x}_j)\in r^*: \quad & \mathbf{w}^T(\mathbf{x}_i-\mathbf{x}_j)\geq 1-\xi_{i,j} \\ \forall i\forall j: \quad & \xi_{i,j}\geq 0 \end{aligned}$$

In this new formulation, more weight is given to the slack variables that correspond to pairs with a larger difference in their belief assessment. In other words, if p_i-p_j is large, the slack variable $\xi_{i,j}$ gets more weight and contributes more to the minimization process. This is equivalent to emphasizing more the satisfaction of the hard constrain $\mathbf{w}^T(\mathbf{x}_i-\mathbf{x}_j)\geq 1$ by reducing the value of $\xi_{i,j}$.

5 Experiments

We test the performance of our methods on clinical data obtained from EHRs for post-surgical cardiac patients and the problem of monitoring and detection of the Heparin Induced Thrombocytopenia (HIT)[4,18]. HIT is an adverse immune reaction that may develop if the patient is treated for a longer time with heparin, the most common anticoagulation treatment. If the condition is not detected and treated promptly it may lead to further complications (such as thrombosis) and even to patient's death. An important clinical problem is the monitoring and detection of patients who are at risk of developing the condition. Alerting when this condition becomes likely prevents the aggravation of the condition and appropriate countermeasures (discontinuation of the heparin treatment or switch to an alternative anticoagulation treatment) may be taken. In this work, we investigate the possibility of building a detector from patient data and human expert assessment of patient cases with respect to HIT and the need to raise the HIT alert. This corresponds to the problem of learning a classification model from data where expert's alert or no-alert assessments define class labels.

Data collection

In this experiment we have started with data from Electronic health records of approximately $4,500$ post-surgical cardiac patients stored in PCP database[19,20]. Each patient record was sliced in time at 8:00am and was used to generate thousands of patient instances. Out of these we have selected 182 instances and asked an expert – a clinical pharmacist, who routinely attends to and evaluates patients at risk of HIT – to provide us with the following information: (1)

whether she agree with the decision to raise an alert on the risk of HIT, and (2) how strongly she agrees or disagrees with alert decision. We used 5 discrete belief scores 4:'strongly-agree', 3:'agree', 2:'weakly-disagree', 1:'disagree' and 0:'strongly-disagree'. Note, that these labels create a ranking of how much the expert agrees with an alert. In order to make the qualified judgement, the expert was able to see the complete patient medical record, including text reports. Out of 182 labeled instances, there were 4, 33, 47, 32, 66 instances labeled with 'strongly-agree', 'agree', 'weakly-disagree', 'disagree' and 'strongly-disagree' respectively, from which 37 instances (or 20.3%) were positive.

We would like to note that 182 examples selected for the assessment were not selected randomly from all patient instances, instead they were chosen via a stratified sampling approach to assure we observe a larger proportion of positive HIT alerts. Briefly, the incidence of HIT in the postsurgical cardiac population is about 2%[4,18]. Hence, if we were to sample patient instances randomly from all possible time segmented patient cases, the chance of seeing a positive HIT alert on a randomly picked instance would be very low.

Data instances

The data in medical records are high dimensional. For the purpose of this study, we have selected 50 features derived from the patient health record and clinical variables important for the detection of HIT. These features represent time series of labs, medications and procedures. From labs we used Platelet counts, Hemoglobin levels and White Blood Cell Counts and their time series. The features generated for labs in the experiment included: last values observed, time elapsed since the last value was observed, quantitative value trends, apex and nadir values and differences of last values from the nadir and apex values. From medications we used Heparin and its administration record. The medication related features generated for all patient instances reflect whether the patient is currently on the heparin or not, the time elapsed since the medication was started and the time since last change in its administration. Finally, the procedure features included in the data were the indicator of a major heart procedure and the time elapsed since such a procedure. All these features were used to define the patient case. The alert decision by the expert was used as a class label. The degree of belief information collected was the auxiliary information supplementing the class label information.

Methods

To test the benefit of the auxiliary information on the quality of the classification model, we trained the models with training data of different size (from 10 to 100 with a step of 10) and compared them on the data witheld from the training stage. We used the following models in our comparisons:

- **SVM.** The linear SVM with the hinge loss and L2 regularization trained on binary labels only,
- **LinReg.** The linear regression with the lasso regularization trained directly on the auxiliary information,
- **SVM-Rank.** The SVM-Rank method (Section 4.2.) with the hinge loss applied to the pairwise data points and L2 regularization, and
- **SVM-RankW.** The weighted version of the SVM-Rank procedure from Section 4.3.

The constant C in SVM methods was set to 10^{-2}.

We evaluated and compared the performance of the different methods by calculating the Wilcoxon statistic (the area under the ROC curve) on the test data. The results are summarized in Figure 1. Since one of the main objectives of this work was to show the value of auxiliary belief information for unbalanced datasets, we constructed three different sets of training data by limiting the rate of positive samples in the training set to 10%, 5% and 0% of the size of training set (Figures 1(b), 1(c), 1(d) respectively). For all the reported experiments, we used 30 different training and testing data splits and reported the average performance over the corresponding testing datasets.

Discussion

The results of this experiment show that auxiliary belief information can improve the learning process and can help us to obtain better models with a smaller number of training samples. For example, when the frequency of positive

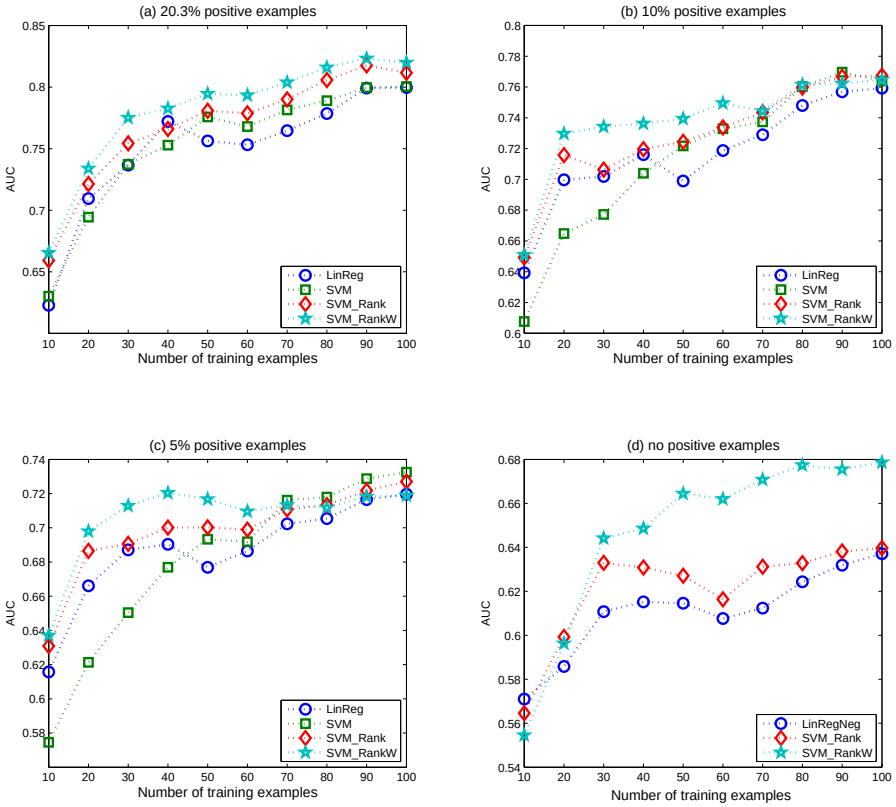

Figure 1: The average area under the ROC curve for different learning methods for varied training-set sizes on the HIT data set. Figures (a)-(d) show the performance of the methods when the percentage of positive examples included in the training data was limited to 20.3% (a), 10% (b), 5% (c)and 0% (d).

samples in the training data is kept at 20.3% (Figure 1(a) a), the SVM_RankW method reaches the AUC score of $.775$ with the training size is 30. To reach the same AUC score, the SVM method needs 65 and the linear regression method 85 examples respectively.

Other experimental results in Figure 1(a) show that the benefit of auxiliary information becomes more pronounced when datasets are more unballanced and the rate of positive samples in the training set is small. For the most extreme case, when there are no positive examples in the training set (Figure 1(d) d), the AUC score for the SVM remains at or close to 0.5 (not reported in the figure) confirming the method is not able to learn anything from such a data. However, other methods, relying on auxiliary information are still able to learn a classification model. Out of the three models, the SVM_RankW method is the best and outperforms other two methods that rely on auxiliary information. More specifically, the weighted SVM-Rank model learned on 30 training examples is better than models learned by other two methods even if they are trained on 100 training examples.

The reported results demonstrate the benefit of auxilary information collected at the time of case review and case annotation for learning classification models. In general the approach can be very useful when: (a) the amount of accessible medical data is limited, (b) the cost of labeling is high and the expected number of labeled examples that can be obtained is low and (c) the two classes of examples in data are unballanced. The results are especially encouraging for classification problems with highly unbalanced classes, which are quite common in medical domain, such as HIT where the rate of positive HIT examples in postsurgical cardiac population is about 2%[4,18]. Intuitively, if the expert is able to rank the cases with respect to the underlying condition we alert on, we are able to benefit and extrapolate from this information and apply it to rare cases.

6 Conclusion

Making use of many real-world data sets often prompts one to fill additional information with subjective human labels. However, this process is often very time consuming and different ways of reducing the labeling costs need to be sought. In this work we investigate a new framework for reducing this cost by reducing the number of examples one must label. The trick is to use an auxiliary probabilistic information that reflects how strongly the human believes in the label which can be extracted cheaply and virtually at no additional cost. We propose multiple methods that use this information to make the learning more sample-efficient. Since the subjective estimates are often inconsistent and noisy we propose and test ranking based methods that are more resilient to the noise. We test the methods and show the improved performance on a real-world medical data set.

7 Acknowledgement

This research work was supported by grants R01LM010019, R01GM088224, and R21LM009102 from the National Institutes of Health. Its content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

1. David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. Active learning with statistical models. *JAIR*, 4:129–145, 1996.

2. Ralf Herbrich, Thore Graepel, and Klaus Obermayer. Support vector learning for ordinal regression. In *ICANN 1999*, pages 97–102, 1999.

3. Hamed Valizadegan, Rong Jin, Ruofei Zhang, and Jianchang Mao. Learning to rank by optimizing ndcg measure. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *NIPS 22*, pages 1883–1891, 2009.

4. TE. Warkentin, JI. Sheppard, and P. Horsewood. Impact of the patient population on the risk for heparin-induced thrombocytopenia. *Blood*, pages 1703 – 1708, 2000.

5. Iyad Batal, Lucia Sacchi, Riccardo Bellazi, and Milos Hauskrecht. A temporal abstraction framework for classifying clinical temporal data. *American Medical Informatics Association Anual Symposium*, 2009.

6. Iyad Batal and Milos Hauskrecht. Mining clinical data using minimal predictive rules. *American Medical Informatics Association Anual Symposium*, 2010.

7. Padhraic Smyth. Learning with probabilistic supervision. *Computational Learning Theory and Natural Learning System 3*, pages 163–182, 1995.

8. Padhraic Smyth, Usama Fayyad, Michael Burl, Pietro Perona, and Pierre Baldi. Inferring ground truth from subjective labeling of venus images. *NIPS 7*, pages 1085–1092, 1995.

9. Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

10. Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition, 2007.

11. Thorsten Joachims. Training linear svms in linear time. In *KDD'06*, pages 217–226, 2006.

12. Arthur E. Hoerl and Robert W. Kennard. Ridge regression—1980. Advances, algorithms, and applications. *American Journal of Mathematical and Management Sciences*, 1(1):5–83, 1981.

13. Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

14. Jerome Friedman. Regularization paths for generalized linear models via coordinate descent. *J. Roy. Statist. Soc. Ser. B*, 33(1), 2010.

15. Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(2):301–320, 2005.

16. Liana Suantak, Fergus Bolger, and William R. Ferrell. The hard-easy effect in subjective probability calibration. *Organizational Behavior and Human Decision Processes*, 67(2):201 – 221, 1996.

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

18. TE. Warkentin. Heparin-induced thrombocytopenia: pathogenesis and management. *Br J Haematology*, pages 535 – 555, 2003.

19. Milos Hauskrecht, Michal Valko, Iyad Batal, Gilles Clermont, Shyam Visweswaram, and Gregory Cooper. Conditional outlier detection for clinical alerting. *American Medical Informatics Association Annual Symposium*, 2010.

20. Michal Valko and Milos Hauskrecht. Feature importance analysis for patient management decisions. *MEDINFO*, 2010.