# Learning classification with auxiliary probabilistic information

Quang Nguyen, Hamed Valizadegan, Milos Hauskrecht
*Department of Computer Science*
*University of Pittsburgh*
*Pittsburgh, United States*
*e-mail:{quang,hamed,milos}@cs.pitt.edu*

*Abstract*—**Finding ways of incorporating auxiliary information or auxiliary data into the learning process has been the topic of active data mining and machine learning research in recent years. In this work we study and develop a new framework for classification learning problem in which, in addition to class labels, the learner is provided with an auxiliary (probabilistic) information that reflects how strong the expert feels about the class label. This approach can be extremely useful for many practical classification tasks that rely on subjective label assessment and where the cost of acquiring additional auxiliary information is negligible when compared to the cost of the example analysis and labelling. We develop classification algorithms capable of using the auxiliary information to make the learning process more efficient in terms of the sample complexity. We demonstrate the benefit of the approach on a number of synthetic and real world data sets by comparing it to the learning with class labels only.**

*Keywords*-**classification learning; sample complexity; learning with auxiliary label information**

## I. Introduction

Nowadays, large real-world data sets are collected in various areas of science, engineering, economy, health care and other fields. These data sets provide us with a great opportunity to understand the behaviour of complex natural and man-made systems and their combinations. However, many of these real-world data sets are not perfect and come with missing information we currently have no means to collect automatically.

One type of information that is often not collected and recorded in the real-world data are subjective labels provided by an expert in the field that assign examples in the data to one of the classes of interest. Take for example a patient health record. While some of the data (such as lab test data, medications given) are archived and collected, diagnoses of some conditions, or adverse events that occurred during the hospitalization are not. If the goal is to analyse these conditions and predict them, individual patient examples must be first labelled by an expert or a group of experts.

The process of labelling examples using subjective human assessments can be an extremely time-consuming and costly process, especially when examples are non-trivial and high-dimensional. 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 analyse examples, prioritize them and select those that are most critical for the task we want to solve, while optimizing the overall data labelling cost.

In this work, we explore another direction that is orthogonal to the active learning approach that can help us to alleviate the costly example labelling process in practice. The idea is based on a simple premise, the human expert that gives us a subjective label can often provide us with auxiliary information related to the case and its assessment which reflects his/her certainty in the label decision and may take the form of belief or assessment confidence or similar measure, and this at cost that is insignificant when compared to the cost of the example review and label assessment. To illustrate this point, assume an expert reviewing electronic health record data in order to provide some diagnostic assessment of the patient case. Clearly the complexity of the data prompts him/her to spend a large amount of time reviewing and analysing the case (typically 3-5 minutes). Once the example is understood and the label decision is made, the cost of providing additional assessment of the confidence in this decision is relatively small and insignificant.

In this work we propose and study a machine learning framework in which a classification learning problem relies on both a class label and a probabilistic assessment of the confidence or belief in this label. We first show how one can easily modify one of the basic learning algorithms (the logistic regression), to accept the probabilistic assessments and learn a high quality classifiers with a smaller number examples. Since the new model depends strongly on the accuracy of subjective probabilistic estimates it may become sensitive to the assessment inconsistencies and noise. To address the problem we propose a novel method based on the support vector machine and learning to rank approach that is more robust to this noise and is able to learn a high quality classifier with a smaller number of examples. We demonstrate the benefits of our method on a number of UCI data sets, while adding noise to probabilistic assessments. Finally we test the method on a real-world clinical diagnosis problem.

## II. PROBLEM DESCRIPTION

We want to learn a binary classifier $f : X \to Y$. In addition to binary $\{0, 1\}$ labels defining $Y$ we also have access to additional information: a probability $p_i$ reflecting one's belief the example $\mathbf{x}_i$ belongs to class 1. Hence each data entry in the data set $D = \{d_1, d_2, \cdots d_N\}$ we 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.

The probabilistic information we assume in our problem 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 probability 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 probability $p_i$ are closely related. Adopting a decision-theoretic perspective, we assume the class label $y_i$ is determined by a threshold on the class posterior probability $p(y_i|\mathbf{x})$ that reflects different loss applied to different types of misclassification errors.

Our main conjecture in this work is that additional probabilistic 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 we learn from are unbalanced (the prior probability of one of the classes is small), and when the number of labelled examples we can learn from is limited.

### A. Related work

The process of labelling examples using subjective human assessments can be an extremely time-consuming and costly process, especially when examples are non-trivial and high-dimensional. Optimizing the time and cost of this process boils down to reducing the number of examples one must assess. In the following we briefly review research work that addresses the problem and contrast them to our framework.

One of the research directions relevant to our work is transfer learning [2], [3], [4], [5], [6]. Transfer learning relies on auxiliary data sources and labels related to the target problem. The auxiliary data and their labels are either combined with the existing (typically small-size) data for the target task or used to refine the model or at least some of the model components and hence simplify the learning. The limitation of this approach is that we do not always have auxiliary data that are relevant to the current task, and also there is often a need to tune some parameter that regularizes the "importance" of auxiliary training data with that of the training data. Our framework is different from the transfer learning work; the auxiliary information it uses is directly related to the target labels.

Another relevant research direction explored extensively by the data mining and machine learning communities in recent years is the development of active learning [1] methods that analyse examples, prioritize them and select those that are most critical for the task we want to solve, while optimizing the overall data labelling cost. We would like to note that our approach is orthogonal to this effort, since we try to gain more useful information from selected examples with little additional cost. A combination of active learning with auxiliary confidence information is a possibility and we leave this direction for the future work.

Learning with labels based on subjective human estimates is central also to the learning from crowds framework. In this framework a case is reviewed and labelled by multiple reviewers [7], [8], [9] and a 'consensus' model is sought. In our work, we study ways of better learning model from one expert, not a crowd, so these methods are complementary. A combination of our approach with the crowd learning is a possible extension.

The closest to our framework is the research by [10], [11] who considers probabilistic information as a vital component of the learning process because of the ambiguities in the class labelling. This work applies the approach to classification of volcanoes 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.

## III. ALGORITHMS FOR LEARNING WITH PROBABILISTIC LABELS

In this section we develop classification learning algorithms that let us accept and learn from probabilistic labels. We start with modifying a simple discriminative model, and keep modifying the model to account for possible noise and inconsistencies in subjective probability estimates.

### A. Discriminative model

In the discriminative classification approach we want to learn a function $f : X \to \mathcal{R}$ that lets us discriminate examples in the two classes. Once the function $f$ is known, the class decision is made with the help of a threshold $\sigma$ such that for values $f(\mathbf{x}) \geq \sigma$ we classify the example as class 1, otherwise the class is 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, we have access also to auxiliary probabilistic 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 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 $(x_i, p_i)$ are the input-output pairs. Obviously, using an arbitrary function model, the outputs of the regression may not be consistent with probabilities. For example, by applying a linear regression directly to the input-probability pairs we may not guarantee the consistency of probabilistic labels once the model is learned, that is, some data points may fall outside $[0, 1]$ interval. An alternative is to regress inputs to a new space in $\mathcal{R}$ obtained by transforming the probabilistic space, such that the transformation is monotonic in $p_i$, and its inverse lets us revert back to probabilities. An example of such a transformation is $t(p_i) = \ln \frac{p_i}{1-p_i}$ which is the inverse of the logistic function. In such a case the regression model is trained on $(x_i, t(p_i))$ pairs. The results of the regression can be transformed back to the probability space by using the logistic function $g(s) = \frac{1}{1+e^{-s}}$ and the probabilities are consistent.

*1) Linear regression:* Assuming the function $f : X \to R$ 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.

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

The solution $\mathbf{w}^* = \arg\min Error(D, \mathbf{w})$ yields a weight vector optimizing the linear model. If needed, the posterior probabilities are recovered as: $p(y = 1|\mathbf{x}, \mathbf{w}) = \frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}}$.

**Defining the classification threshold.** Once the weights of the discriminant function are learned, a classifier can be defined using a decision threshold $\sigma$. To find the optimal threshold, we can use true class labels and minimize the overall loss in the training data.

*2) Regularization:* Regression methods are quite common and can be enriched with different bells and whistles that fit better different data settings. Our 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 $N$ is small, a possibility of the model over-fit arises. In such a case we can modify and improve the performance of the regression model using one of regularization approaches, such as the ridge (or $L_2$) regularization [12], the lasso (or $L_1$) regularization [13], [14], or their elastic network combination [15]. Briefly, using regularization, the optimization in Equation 1 is modified to:

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

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.

*3) Sensitivity to the noise in subjective estimates:* Learning of the discriminant function directly from auxiliary probabilities 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. Figure 1 shows the performance of the linear regression model on one of the data sets (Concrete) we analysed during the study (see Section IV). It shows the performance of the logistic regression model learnt from original probabilistic estimates and estimates corrupted by additional noise. The logistic regression model learnt from binary labels is shown as the baseline. 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.

### B. Using ranking to improve the noise tolerance

The above regression approach learns the model solely using the auxiliary probabilistic information. As a result it may become very sensitive to the noise and inconsistencies in the numerical assessments as illustrated in Figure 1. Since humans are not very good in providing well calibrated probabilistic estimates [16], [17], [18], 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 probabilistic estimates, we try to model the relation in between the two probabilistic assessments only qualitatively, in terms of pairwise order constraints.

Let $f : X \to \mathcal{R}$ be a linear model $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ that lets us discriminate between examples in class 0 and class 1. Now assume the same model represents a linear ranking function that lets us 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 probability $p_1$ and $p_2$, we expect the ranking function to preserve their order.

The learning to rank algorithms [19], [20] let us find the 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 would allow us to better absorb some amount of noise in subjective probability estimates, eventually leading to more robust learning algorithms.
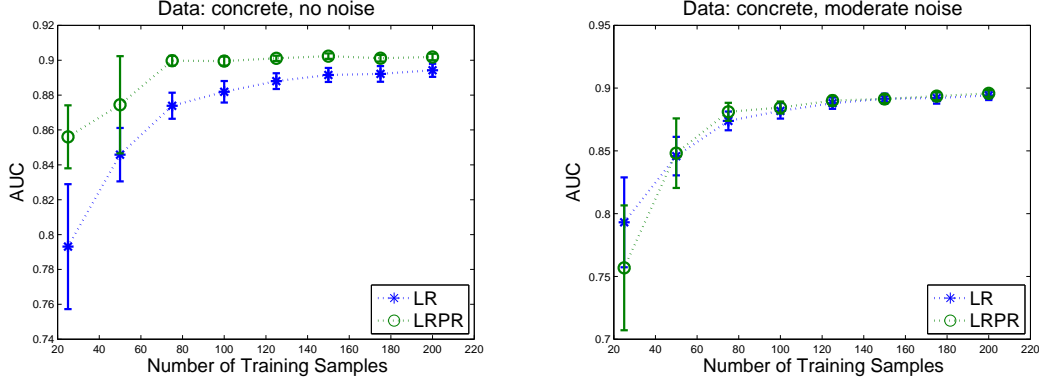
Figure 1. The sensitivity of the logistic regression model learnt with auxiliary probabilistic information to the noise. Left: AUC for the model learnt with the auxiliary information. Right: AUC for the model learnt when the auxiliary information was corrupted by a moderate Gaussian noise.

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

$$\min_{\mathbf{w}} \quad Q(\mathbf{w}) + C \sum_{i,j} \xi_{i,j}$$

$$\text{subject to:}$$
$$\forall(\mathbf{x}_i, \mathbf{x}_j) \in r^* : \quad \mathbf{w}^T(\mathbf{x}_i - \mathbf{x}_j) \geq 1 - \xi_{i,j}$$
$$\forall i \forall j : \quad \xi_{i,j} \geq 0$$

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

*Assuring class-label consistencies:* The basic ranking solution presented above relies purely on the auxiliary probabilistic information and ignores the class labels. Because of the subjective estimates, it is not uncommon that two different class labels may get probabilities that rank them opposite of their expected order, that is, the probability assigned to a class 0 example is higher than the subjective probability assigned to a class 1 example. To address the problem we may set the priority and define the pairwise ordering constraints such that they respect the class label as the primary criterion and the auxiliary information as the secondary criterion.

## C. Optimizing the discriminant function by combining the label and auxiliary probabilistic information

The ranking approach presented above improves the noise tolerance of the model to subjective probability estimates by ignoring their exact values and taking into account only their relative order. Another feature of the model is that the binary label information is incorporated in the model only in terms of the order constraints. The main question that arises is whether it is possible to incorporate and benefit from both the class labels and auxiliary information by combining the loss function for the class label and the loss from the auxiliary information into a single coherent optimization criterion. The hope for doing this is to assure the model is driven by the class label first and refined with the auxiliary probabilistic information, if it is consistent with the labels.

In particular, we propose to optimize:

$$\min_{\mathbf{w}} \quad Q(\mathbf{w}) + B \sum_i \eta_i + C \sum_{i,j} \xi_{i,j}$$

$$\text{subject to:}$$
$$\forall \mathbf{x}_i y_i \quad \mathbf{w}^T \mathbf{x}_i y_i + b \geq 1 - \eta_i$$
$$\forall(\mathbf{x}_i, \mathbf{x}_j) \in r^* : \quad \mathbf{w}^T(\mathbf{x}_i - \mathbf{x}_j) \geq 1 - \xi_{i,j}$$
$$\forall i : \quad \eta_i \geq 0$$
$$\forall i \forall j : \quad \xi_{i,j} \geq 0$$

where $B$ and $C$ are constants and $Q(\mathbf{w})$ is a regularization penalty. Briefly, this formulation assumes two sets of constraints, one defining the hinge loss for all examples and their labels, the other one loss for not respecting the order induced by subjective probabilistic estimates. Once again 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. Note that by changing scaling constants $B$ and $C$ one can stress more either the label or the probabilistic order information. For example, if the noise in probabilistic labels is large then its influence can be decreased by decreasing $C$. In general the settings of

| Data set | # examples | # features |
|----------|------------|------------|
| ailerons | 7154 | 40 |
| concrete | 1030 | 8 |
| kinematics | 8192 | 8 |
| puma32 | 4499 | 32 |

these parameters can be optimized using the internal cross-validation approach. In the paper we refer to this approach as to the *SVM-Combo* approach.

## IV. EXPERIMENTS

We have conducted two sets of experiments to test our framework and methods. The first set of experiments uses four UCI Irvine data sets. We modify these data to fit our framework and provide both the label and auxiliary information with different levels of noise. We use the data to first demonstrate the benefits of auxiliary information for learning the classification models. Second, we show the robustness of our methods to noise in the auxiliary information. Finally we use the data to show how our approach can alleviate the learning problem when the data are unbalanced. The second experiment applies the framework to real-world clinical data and human assessments of a risk of a life threatening condition – the heparin induced thrombocytopenia [21], [22] and demonstrates the improved learning of classification models.

### A. Experiments with UCI data sets

In this set of experiments we use four UCI regression data sets. The data sets and their properties are summarized in Table I. For all these data sets we modified the continuous output variables and interpreted them as probabilities. We defined a binary class variable by using a threshold on the underlying continuous variable. For example, the variable representing the strength of concrete in the Concrete data set was used to define two classes: a concrete with a good strength and a concrete with a bad strength. Specific thresholds used to define the binary class variable and are discussed below.

*1) Effect of the training data size on the model quality:*
To test the benefit of the methods and the impact of auxiliary probabilistic information on the sample complexity we trained the new models with training data of different size and compared them to models that learn from the binary labels only. We used the following models in our comparisons:

- **LR.** The logistic regression with lasso regularization trained on binary labels,
- **LRPR.** The logistic regression model with lasso regularization trained on the auxiliary probabilistic information (from Section III-A),

- **SVM.** The linear SVM with the hinge loss and L2 regularization trained on binary labels only, and
- **SVM-Combo.** The new SVM model (from Section III-C) with the L2 regularization penalties and two hinge losses: one for binary labels and the other one for pairwise ordering constraints.

The constants $C$ and $B$ for SVM models were optimized using 3-fold cross-validation approach.

We evaluated performance of the different methods by calculating the Wilcoxon statistic (the area under the ROC curve). Each data set was split into training and testing set (2/3 and 1/3 of all data respectively). We fixed the testing set and randomly selected samples from the training set to train the models. The training process was repeated 30 times. We reported the average AUC on the fixed test set. Figure 2 compares the performance of the different methods on all data sets by varying the number of samples selected for training. The error bars show 95% confidence interval.

**Discussion.** The results on all four data sets clearly show the benefit of learning with auxiliary probabilistic information. All methods trained with the auxiliary information outperformed their binary label counterparts and the sample complexity for training the model of equivalent quality was greatly reduced. On three of the data sets the best method was the SVM-Combo method, the logistic regression with auxiliary information was the best performing method on one data set.

*2) Effect of the noise on the auxiliary information:* Our first experiment assumed the class label is defined directly by the probabilistic information. It meant to show that the auxiliary information may help. However, in practice the probabilistic information is often imprecise and subject to noise. This may effect its utility for learning the classification models. Our second experiment aims to demonstrate the robustness of our method to such a noise.

Figure 3 shows the performance of four methods from the previous experiment when auxiliary probabilistic information is corrupted by a noise. We assume four different levels of noise: no, weak, moderate and strong noise. To obtain the noisy estimates each auxiliary probability value was modified as follows:

- no noise: 0;
- weak noise: Gaussian noise from 0.05*N(0,1);
- moderate noise: Gaussian noise from 0.1*N(0,1);
- strong noise: Gaussian noise from 0.2*N(0,1).

To avoid inconsistent probability values, we assured all auxiliary values always fell in the interval [0,1]. To better understand the true noise influence on the probabilistic information Table II lists the average proportion of the noise to the original signal after different levels of noise were applied to original data.

**Discussion.** When the noise is injected into the probabilistic information the logistic regression model trained with
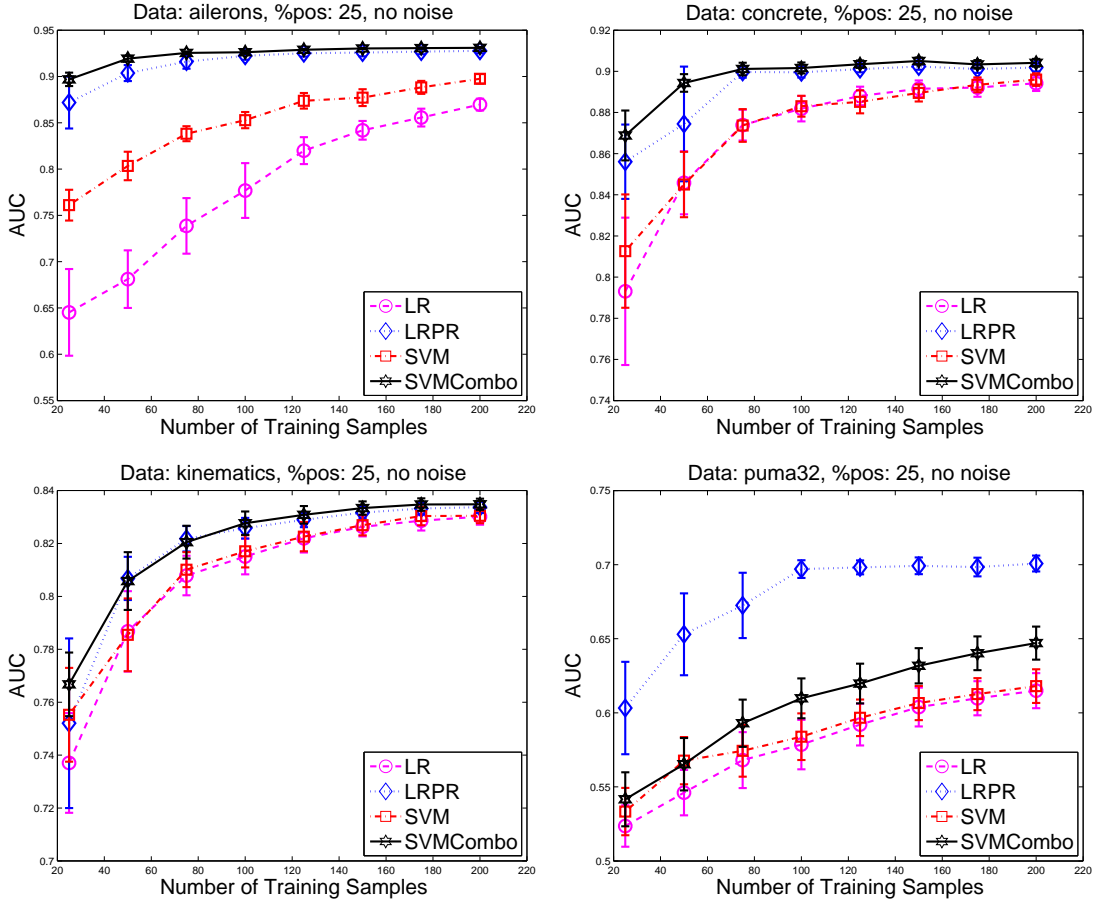
Figure 2. The benefit of learning with auxiliary probabilistic information on four different UCI data sets. The quality of resulting classification models for different training sample sizes is shown in terms of the Area under the ROC curve statistic.

| Data set | Weak noise | Moderate noise | Strong noise |
|----------|-----------|----------------|--------------|
| ailerons | 5.2 % | 10.3 % | 39.8 % |
| concrete | 15.2 % | 29.6 % | 55.1 % |
| kinematics | 10.6 % | 20.8 % | 38.9 % |
| puma32 | 10.3 % | 20.2 % | 39.3 % |

probabilistic information is sensitive and its performance drops. We see the logistic regression model trained on binary labels in some instances outperforms the model with auxiliary information. However, our SVM-Combo approach that uses ranking is more robust and outperforms both the baseline logistic regression and the baseline SVM for all three noise levels. This shows the robustness of our approach to noisy auxiliary information estimates.

*3) Effect of auxiliary information when learning with unbalanced data set:* The binary labels in all previous experiments were generated using the probabilistic information

such that the number of positive and negative examples was 25% and 75% of data respectively. The question we investigate now is how the probabilistic information influences the learning process when different proportions of positive and negative examples are observed. In general, we expect that learning from labelled data when data set is unbalanced is much harder than with a balanced data set.

**Experiment**. Figure 4 compares four learning methods on four UCI data sets where positive labels were restricted to top 10, 25 and 50 percent of examples with the highest observed outcome values respectively. After labels were generated, the weak level of noise was applied to corrupt the probabilistic information.

**Discussion.** The results in Figure 4 clearly show the benefit of learning with the auxiliary information relative to learning binary labels only for more unbalanced data. More specifically, the gap in the predictive performance between models learned with and without the auxiliary information increases when the data set is more unbalanced.
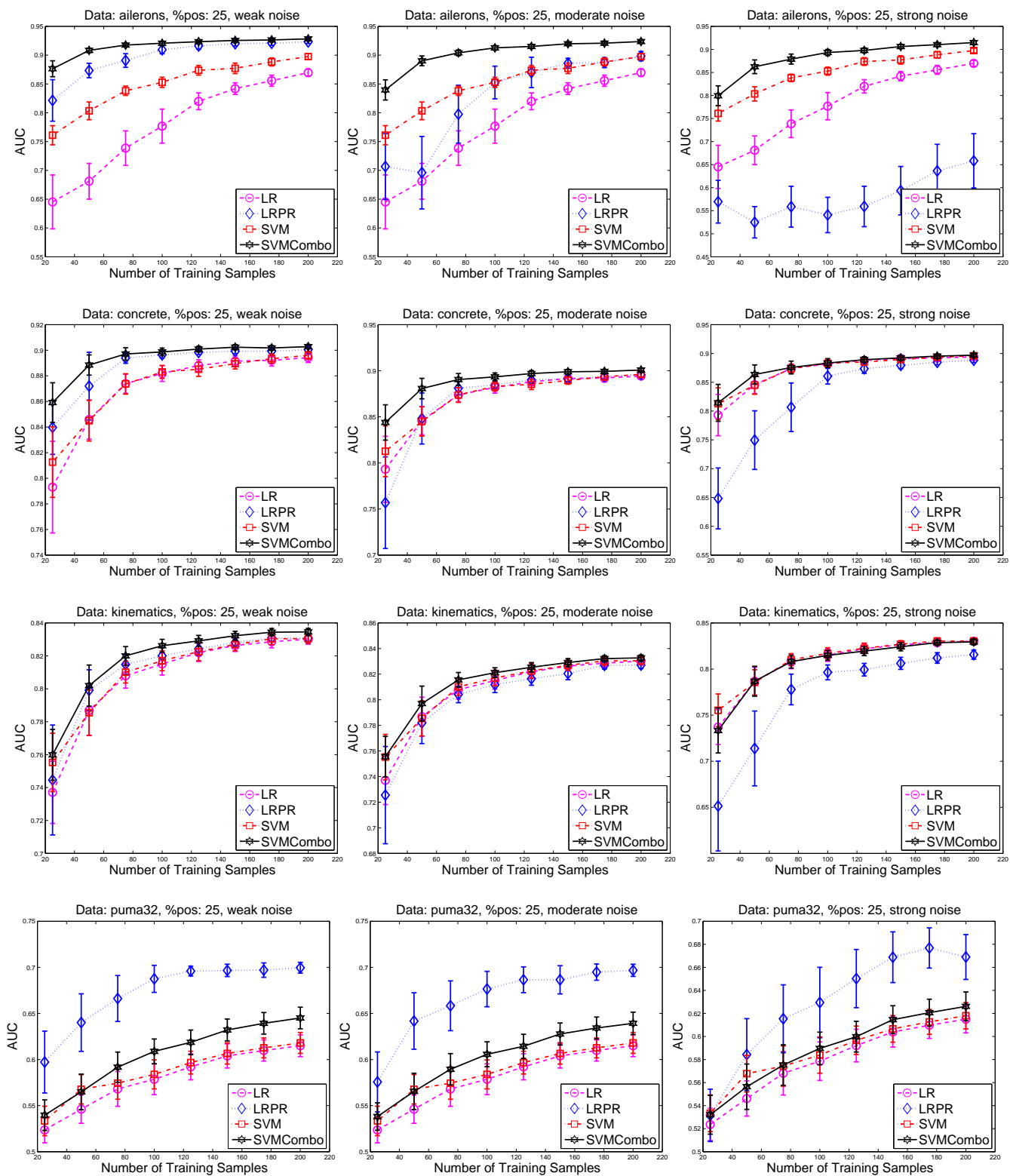
Figure 3. The sensitivity to noise in auxiliary information. Area under the ROC curve vs. sample size for different learning methods trained on data with auxiliary information corrupted by different levels of noise.
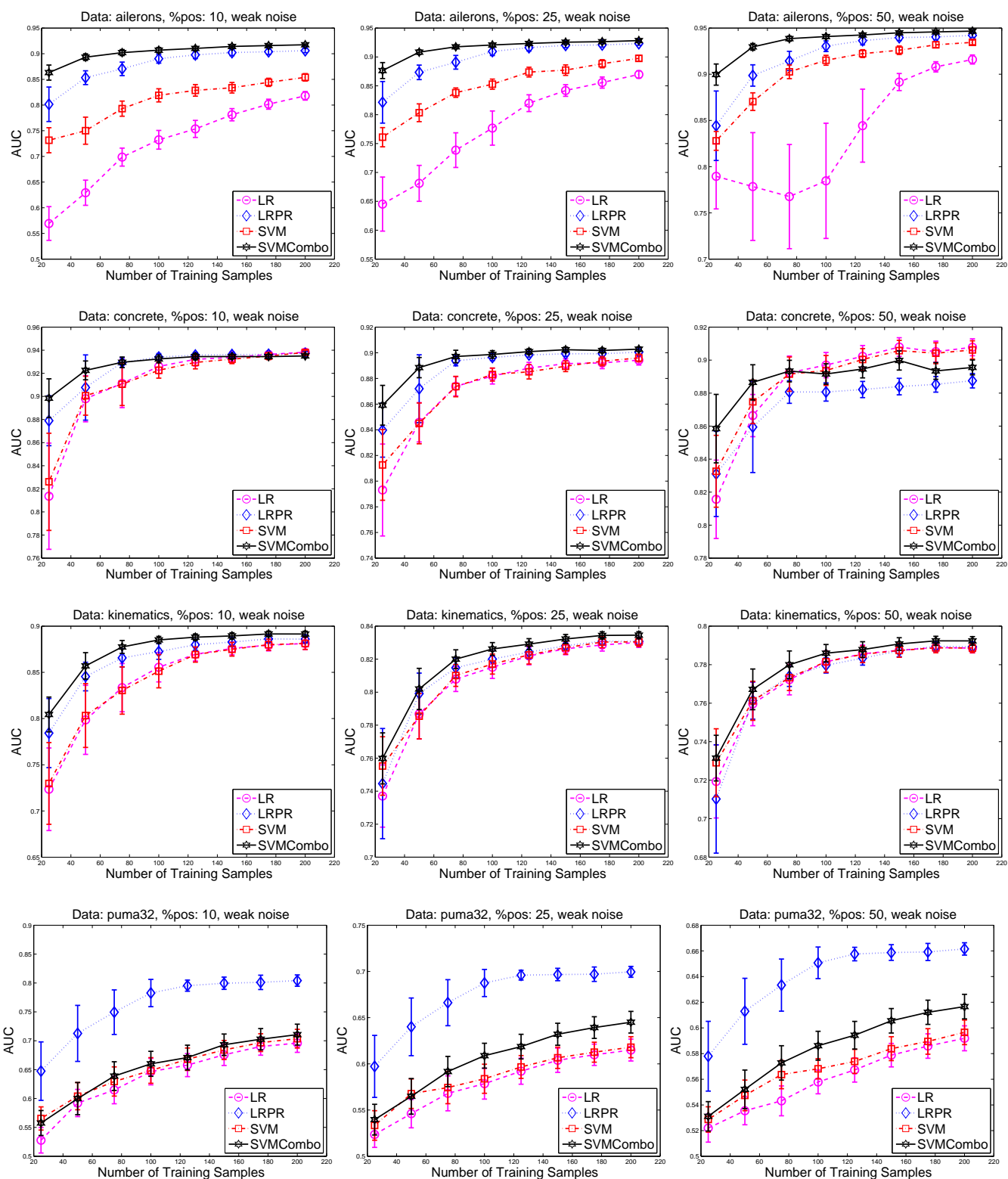
Figure 4. Learning with auxiliary information on data with different percentages of positive examples. Area under the ROC curve statistic vs. sample size is shown.

## B. Experiments with patient data

Our next experiment tests the methodology on the real clinical data and the problem of detection of the risk of the heparin induced thrombocytopenia (HIT) [21]. HIT is an adverse immune reaction that may develop if the patient is treated for a longer time with heparin, which is 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 problem is the monitoring and detection of patients who are at risk of developing the condition. In this work, we investigate the possibility of building a detector from patient data using the assessment of the HIT and its risk by an expert. This corresponds to the problem of learning a classification model from data.

**Data collection.** In this experiment we have started with data extracted from electronic health records that consisted of over 50,000 patient-state instances. Out of these we have selected 199 instances using a special stratified sampling approach where individual strata were defined to increase or decrease the chance the patient-state instance is associated with HIT. We asked an expert to provide us with the following information: (1) assess whether they would consider the instance at the risk of HIT and alert on it, or not, and (2) the confidence in raising the alert. In order to make the qualified judgement, the expert was able to see the complete patient medical record. The review of the case was the most costly part of the process and on average took 247 seconds. The time to enter assessment of the alert and confidence of the case was small and typically was finished under 10 seconds.

The data in medical records are high dimensional. For the purpose of this study, we have selected 50 features derived from patient health record and clinical variables most important for the detection of HIT. These features were used to define the patient state example. The alert decision by the expert was used as a class label. The confidence information collected was the auxiliary information supplementing the class label information. We run the same set of four methods from the previous experiments. The average AUC results over 30 training/testing splits are summarized in Figure 5.

**Discussion.** The results on this experiment confirm the results on UCI data set. They show the auxiliary probabilistic information may help us learning better models with a smaller number of samples. However, the difference from the baseline classifiers is much smaller than for UCI data sets which we attribute to a large amount of noise we observe in human subjective assessments of the alert confidence. This is evident also from the results for the logistic regression model trained with auxiliary information which has been shown above to be sensitive to the noise.

## V. Conclusions

Making use of many real-world data sets often prompts one to fill additional information with subjective human
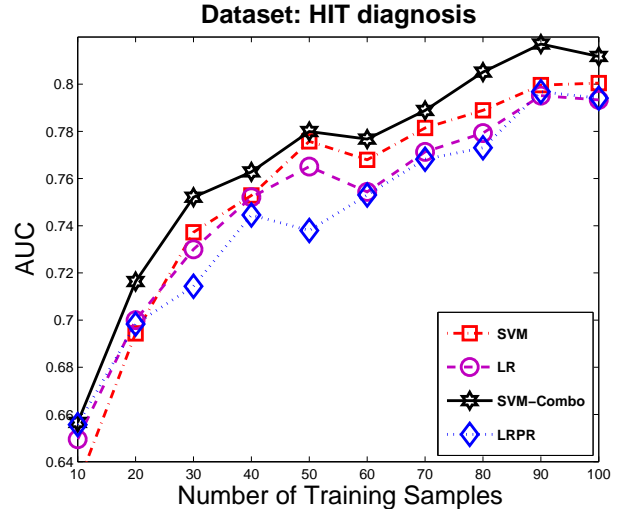


Figure 5. The results of the different methods on the HIT data set

labels. However, this process is often very time consuming and different ways of reducing the labelling 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 UCI and a real-world medical data set.

## VI. Acknowledgement

## References

[1] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *JAIR*, vol. 4, pp. 129–145, 1996.

[2] P. Wu and T. G. Dietterich, "Improving svm accuracy by training on auxiliary data sources," in *ICML*, 2004, pp. 871–878.

[3] X. Liao, Y. Xue, and L. Carin, "Logistic regression with an auxiliary data source," in *Proceedings of the 22nd international conference on Machine learning*, ser. ICML '05. New York, NY, USA: ACM, 2005, pp. 505–512. [Online]. Available: http://doi.acm.org/10.1145/1102351.1102415

[4] S. Kaski, J. Peltonen, U. De, S. Kaski, and J. Peltonen, "Learning from relevant tasks only," in *Proceedings of the 18th European conference on Machine Learning (ECML '07)*, 2007.

[5] L. Duan, I. W. Tsang, D. Xu, and T. seng Chua, "Domain adaptation from multiple sources via auxiliary classifiers," in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009.

[6] S. J. Pan and Q. Yang, "A survey on transfer learning," in *IEEE Transactions on Knowledge and Data Engineering, pp. 1345-1359*, October 2010.

[7] V. S. Sheng, F. Provost, and P. G. Ipeirotis, "Get another label? improving data quality and data mining using multiple, noisy labelers," in *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '08. New York, NY, USA: ACM, 2008, pp. 614–622. [Online]. Available: http://doi.acm.org/10.1145/1401890.1401965

[8] V. C. Raykar, S. Yu, L. H. Zhao, A. Jerebko, C. Florin, G. H. Valadez, L. Bogoni, and L. Moy, "Supervised learning from multiple experts: Whom to trust when everyone lies a bit," in *ICML*, 2009.

[9] P. Welinder, S. Branson, S. Belongie, and P. Perona, "The multidimensional wisdom of crowds," in *Neural Information Processing Systems Conference (NIPS)*, 2010. [Online]. Available: http://vision.ucsd.edu/project/visipedia

[10] Smyth, "Learning with probabilistic supervision," *Computational Learning Theory and Natural Learning System 3*, pp. 163–182, 1995.

[11] Smyth, Fayyad, Burl, Perona, and Baldi, "Inferring ground truth from subjective labeling of venus images," *NIPS 7*, pp. 1085–1092, 1995.

[12] A. E. Hoerl and R. W. Kennard, "Ridge regression—1980. Advances, algorithms, and applications," vol. 1, no. 1, pp. 5–83, 1981.

[13] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.

[14] J. Friedman, "Regularization paths for generalized linear models via coordinate descent," *J. Roy. Statist. Soc. Ser. B*, vol. 33, no. 1, 2010.

[15] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 67, no. 2, pp. 301–320, 2005.

[16] L. Suantak, F. Bolger, and W. R. Ferrell, "The hard-easy effect in subjective probability calibration," *Organizational Behavior and Human Decision Processes*, vol. 67, no. 2, pp. 201 – 221, 1996.

[17] D. Griffin and A. Tversky, "The weighing of evidence and the determinants of confidence," *Cognitive Psychology*, vol. 24, no. 3, pp. 411 – 435, 1992.

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

[19] R. Herbrich, T. Graepel, and K. Obermayer, "Support vector learning for ordinal regression," in *ICANN 1999*, 1999, pp. 97–102.

[20] H. Valizadegan, R. Jin, R. Zhang, and J. Mao, "Learning to rank by optimizing ndcg measure," in *NIPS 22*, Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, Eds., 2009, pp. 1883–1891.

[21] T. Warkentin, J. Sheppard, and P. Horsewood, "Impact of the patient population on the risk for heparin-induced thrombocytopenia," *Blood*, pp. 1703 – 1708, 2000.

[22] T. Warkentin, "Heparin-induced thrombocytopenia: pathogenesis and management," *Br J Haematology*, pp. 535 – 555, 2003.