ESANN'2009 proceedings, European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning. Bruges (Belgium), 22-24 April 2009, d-side publi., ISBN 2-930307-09-9.

Supervised Learning as Preference Optimization

Fabio Aiolli and Alessandro Sperduti

Dept. of Pure and Applied Mathematics - University of Padova
Via Trieste 63, 35121, Padova - Italy

Abstract. Learning with preferences is receiving more and more attention in the last few years. The goal in this setting is to learn based on qualitative or quantitative declared preferences between objects of a domain. In this paper we give a survey of a recent framework for supervised learning based on preference optimization. In fact, many of the broad set of supervised tasks can all be seen as particular instances of this preference based framework. They include simple binary classification, (single or multi) label multiclass classification, ranking problems, and (ordinal) regression, just to name a few. We show that the proposed general preference learning model (GPLM), which is based on a large-margin principled approach, gives a flexible way to codify cost functions for all the above problems as sets of linear preferences. Examples of how the proposed framework has been effectively used to address a variety of real-world applications are reported clearly showing the flexibility and effectiveness of the approach.

1 Introduction

Supervised learning is probably the most commonly used learning paradigm and a large spectrum of learning algorithms have been devised for this task in the last decades. The need for such large spectrum of learning algorithms is, in part, due to the many real-world learning problems which are characterized by heterogeneous tasks and problem-specific learning algorithms for their solution. These include classification and regression problems (including multi-label and multi-class classification, and multivariate regression), as well as ranking-based (either label or instance ranking) and ordinal regression problems. The standard approach followed to cope with these complex problems is to map them into a series of simpler, well-known settings and then to combine the resulting predictions. Often, however, these solutions lack a principled theory and/or require too much computational resources to be practical for real-world applications.

In this paper we review a general framework encompassing all these supervised learning settings, and we show that many supervised learning problems can actually be modeled through a set of order preferences over the predictions of the learner. This is done by considering both the different type of predictions and type of supervision involved in the problem to be solved. Specifically, four characterizing problem dimensions are taken into account, namely, the type of prediction which is expected, the feedback that the supervision provides, the hypothesis space used in the learning process, and the evaluation function which determines how to measure the accuracy of a learning device.

ESANN'2009 proceedings, European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning. Bruges (Belgium), 22-24 April 2009, d-side publi., ISBN 2-930307-09-9.

From a practical point of view, we show how all these supervised tasks can be addressed by basically solving a linear binary problem on an augmented space, thus allowing the exploitation of very simple optimization procedures available for the binary case. We also stress the flexibility of the preference model which allows a user to optimize the parameters on the basis of a proper evaluation function. In fact, while in general the goal of a problem in terms of its evaluation function is clear, a crucial issue in the design of a learning algorithm is how to get a theoretical guarantee that the defined learning procedure actually minimizes the target cost function. One advantage of the framework reviewed in this paper is that it defines a very natural and uniform way to devise and code a cost function into a learning algorithm. Examples of real-world applications are then discussed by giving a quick overview of a range of problems already successfully addressed by the framework.

In Section 2, we review the general preference learning model (GPLM). Specifically, we show how the preference model generalizes the supervised learning setting by considering supervision as a partial order of (soft) constraints over the learner predictions. In addition, we show (Section 2.1) how the suggested generalization can be instantiated to well-known supervised learning problems. In the same section, we also discuss a linear model for the learner (Section 2.3) and issues about evaluation (Section 2.2). Quite general optimization procedures for training models within the proposed framework are also presented (Section 2.4). In Section 3, different application scenarios are described and briefly discussed. Related work on ranking and preference optimization are reported in Section 4. Finally, in Section 5 conclusions are drawn.

2 Supervised Learning by preferences

As a very general domain of a learning task we consider a space $\mathcal{X}$ of objects (or instances) and a space $\mathcal{Y}$ of classes (or labels). For example, this can be the domain of a recommender system where instances may correspond to customers and labels to products.

In our general preference learning model (GPLM) we want to learn the parameters of a real valued relevance (or scoring) function defined on instance label pairs

$$u : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$$

which should approximate the actual *target* function. In a recommender system task, for example, this target function would represent the actual rating, a real value, a customer would give to a given product.

We can easily note that, when such a scoring function is given, a predictor is able to rank instances in $\mathcal{X}$ basing on their relevance when fixed any label $y \in \mathcal{Y}$, and similarly, to rank labels in $\mathcal{Y}$ basing on their relevance when fixed any instance $x \in \mathcal{X}$.

ESANN'2009 proceedings, European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning. Bruges (Belgium), 22-24 April 2009, d-side publi., ISBN 2-930307-09-9.

2.1 Prediction and Supervision

In this section, we show that by using the setting given above, it is possible to classify several supervised learning tasks in a taxonomy on the basis of their type of prediction and supervision.

2.1.1 Label and Instance Rankings as Qualitative Preferences

A first important family of supervised learning tasks is related to the ordering of the labels, on the basis of their relevance for an instance, and thus they can be characterized by the fact that predictions are based on a full order over the labels. This family of problems is here referred to as *label rankings*.

Different problems have been studied which can be considered in this class. The interesting thing is that these problems differ only on the type of feedback it is possible to obtain as supervision. For example, in the so called *category ranking* problem, the type of supervision is a full order over the set of labels. Similarly, in a *bipartite category ranking* problem, it is required to predict full orders in which a group of labels is ranked over the remaining labels, and the feedback will correspond to partial rankings of length two. In this family of problems also fall any q-label classification problem, and the feedback will correspond to partial rankings of length two but where the first group of labels has cardinality exactly equal to q. Note in passing that the well known single-label classification problem is a trivial subcase where $q = 1$ and the goal is to select the most relevant label for an instance.

Another interesting family of tasks is *instance rankings* where the goal is to order instances on the basis of the relevance of a given label. The duality with respect to label rankings is self-evident and, for brevity, this subject will not be discussed further. However, in principle, a different problem setting exists for each one of the label ranking problems.

The two families of tasks above can be considered *qualitative tasks* since they are concerned with order relations between instance-class pairs. On the other side, *quantitative tasks* are the ones which are more concerned with the absolute values of the relevance of instance-class pairs.

2.1.2 Prediction of Ratings as Quantitative Preferences

Sometimes there is the need to do quantitative predictions about data at hand. The easier case is in binary classification where one has to decide about the membership of an instance to a class as opposed to rank instances by relevance. These settings are in fact not directly subsumed by the settings presented above. As we will see, this can be overcome by adding a set of threshold values which help on these different types of predictions.

We consider the most general case where instances are ranked according to an ordinal scale which is here referred to as *(multivariate) ordinal regression* task. In a recommender system application, for example, there is the need to

predict people ratings on unseen items. Considering a movie-related application, for example, suitable ranks for movies could be given as 'bad', 'fair', 'good', and 'recommended'. With no loss in generality, we can consider the target space as the integer set $\mathcal{Z} = \{0, \ldots, R-1\}$ of R available ranks. Following an approach similar to the one in [1], ranks can be made corresponding to intervals of the real line. Specifically, a set of thresholds $\mathcal{T} = \{\tau_0 = -\infty, \tau_1, \ldots, \tau_{R-1}, \tau_R = +\infty\}$ can be defined and the prediction based on the rule

$$\hat{z} = \arg_i \{u(\mathbf{x}, y) \in (\tau_{i-1}, \tau_i)\}.$$

Given the target rank $z \in \mathcal{Z}$, a predictor will be correct when the following conditions hold: $u(\mathbf{x}, y) > \tau_i$ when $i < z$ and $u(\mathbf{x}, y) < \tau_i$ when $i \geq z$. Note that, in principle, a different threshold set could be used for different labels. The *multi-label classification*, where it is required to classify instances with a subset (the cardinality of which is not specified) of the available labels, corresponds to the same problem as above with only two ranks, relevant and irrelevant ($\mathcal{Z} = \{0, 1\}$). Finally, we have the well-known *(univariate) ordinal regression* task [2, 3] which is a trivial subcase of the multivariate setting with a single label, and the *binary classification* task which is a univariate ordinal regression with only two ranks. Note that binary classification is considered here conceptually different from the above-mentioned single-label classification with only two classes.

Clearly, the taxonomy presented above is not exhaustive but well highlights how many different kinds of structured predictions can be seen as constraints or preferences in conjunctive form where each basic preference is defined over the scoring values and/or a set of threshold values. In particular, we can differentiate between two types of order preferences: *qualitative* preferences in the form

$$(\mathbf{x}_i, y_r) \rhd (\mathbf{x}_j, y_s)$$

telling that the value of $u(\mathbf{x}_i, y_r)$ should be higher than the value of $u(\mathbf{x}_j, y_s)$, and *quantitative* preferences in the form

$$(\mathbf{x}, y) \rhd \tau \text{ or } \tau \rhd (\mathbf{x}, y), \tau \in \mathcal{T}$$

relating the value of $u(\mathbf{x}, y)$ to a given threshold τ from a set $\mathcal{T}$.

In Table 1, a summary of the set of preferences obtained for the most general supervised learning settings are presented. Particular instantiations to more specific problems are immediate.

2.2 Evaluation and cost functions

We have seen how supervision of typical supervised learning problems can be decomposed in terms of set of qualitative and/or quantitative preferences over the scoring function of a learner. Here, we show that preferences also gives us a flexible way to express cost functions which can be utilized to evaluate a learner. Specifically, we consider preference graphs, i.e. directed graphs where

Setting	Supervision
LR	$\{(\mathbf{x}, y_r) \rhd (\mathbf{x}, y_s)\}_{(\mathbf{x},y_r) \succeq_S (\mathbf{x},y_s)}$
IR	$\{(\mathbf{x}_i, y) \rhd (\mathbf{x}_j, y)\}_{(\mathbf{x}_i,y) \succeq_S (\mathbf{x}_j,y)}$
MOR	$\{(\mathbf{x}, y) \rhd \tau_i\}_{i<z} \cup \{\tau_i \rhd (\mathbf{x}, y)\}_{i \geq z}$

Table 1: Supervision of problems in Section 2.1. Label and instance rankings (LR and IR respectively), have a preference for each order relation induced by the supervision S. In ordinal regression (MOR), a preference is associated to each threshold and $z \in \mathcal{Z}$ is the rank given by the supervision.

nodes take values on the set $\mathcal{H} \equiv (\mathcal{X} \times \mathcal{Y}) \cup \mathcal{T}$ and edges $(h_1, h_2) \in \mathcal{H} \times \mathcal{H}$ represent preferences $h_1 \rhd h_2$. We say that a scoring function is consistent with a preference graph whenever it is consistent with all the preferences in the graph. The evaluation of any scoring function can then be performed by checking for how many graphs the scoring function is not consistent with.

Even more general cost functions can be obtained by associating weights to the edges of a graph. In this case, let g a preference graph, the cost suffered by an hypothesis is defined as the maximum of costs of its unfulfilled preferences. Finally, the total cost suffered by a scoring function u for the supervision S, given as a set of preference graphs G, is defined as the cumulative cost over all the preference graphs. More formally, we have

$$c(G|u) = \sum_{g \in G} c(g|u) \quad \text{and} \quad c(g|u) = \max\{\gamma(\lambda) | \lambda \in g \text{ not fulfilled by } u\}. \qquad (1)$$

and $\gamma(\lambda)$ represents the weight associated to the preference λ.

Using cost mappings as defined above, we are able to reproduce many of the different cost functions used for ranking problems. The reader can see [4, 5] for several examples on how to give a cost map on standard supervised problems.

2.3 A Linear Model for the Scoring Function

In this work, we focus on a simple form of the relevance function, that is

$$u(\mathbf{x}, y) = w \cdot \phi(\mathbf{x}, y)$$

where $\phi(\mathbf{x}, y) \in \mathbb{R}^d$ is a joint representation of instance-class pairs and $w \in \mathbb{R}^d$ is a weight vector [6]. Note that this form generalizes the more standard form $u(\mathbf{x}, y) = w_y \cdot \phi(\mathbf{x})$ where different weight vectors are associated to different labels. In fact, let $|\mathcal{Y}| = m$, we can write:

$$w = (w_1, \ldots, w_m) \text{ and } \phi(\mathbf{x}, y) = (\underbrace{\mathbf{0}, \ldots, \mathbf{0}}_{y-1}, \phi(\mathbf{x}), \mathbf{0}, \ldots, \mathbf{0}).$$

With this assumption, it is possible to conveniently reformulate an order constraint as a linear constraint. Let $\mathcal{T} = \{\tau_1, \ldots, \tau_{R-1}\}$ be the available set of thresholds, in the qualitative case, given $\lambda \equiv (\mathbf{x}_i, y_r) \rhd (\mathbf{x}_j, y_s)$, we obtain

$$u(\mathbf{x}_i, y_r) > u(\mathbf{x}_j, y_s) \Leftrightarrow (w, \tau_1, \ldots, \tau_{R-1}) \cdot \underbrace{(\phi(\mathbf{x}_i, y_r) - \phi(\mathbf{x}_j, y_s), \underbrace{0, \ldots, 0}_{R-1})}_{\psi(\lambda)} > 0$$

Viceversa, in the quantitative case, given $\delta \in \{-1, +1\}$, we have

$$\delta(u(\mathbf{x}, y) - \tau_r) > 0 \Leftrightarrow (w, \tau_1, \ldots, \tau_{R-1}) \cdot \underbrace{(\delta\phi(\mathbf{x}, y), \underbrace{0, \ldots, 0}_{r-1}, -\delta, \underbrace{0, \ldots, 0}_{R-r-1})}_{\psi(\lambda)} > 0.$$

In general, we can see that supervision constraints of all the above-mentioned problems, can be reduced to sets of linear preferences of the form $\mathbf{w} \cdot \psi(\lambda) > 0$ where $\mathbf{w} = (w, \tau_1, \ldots, \tau_{R-1})$ is the vector of weights augmented with the set of available thresholds and $\psi(\lambda)$ is an opportune representation of the preference under consideration. The quantity

$$\rho(\lambda|\mathbf{w}) = \mathbf{w} \cdot \psi(\lambda)$$

will be also referred to as the margin of the hypothesis w.r.t. the preference. Note that this value is greater than zero when the preference is satisfied and less than zero otherwise.

Summarizing, all the problems defined in the taxonomy in Section 2.1 can be reduced to an homogeneous linear problem in a opportune augmented space. Specifically, any algorithm for linear classification (e.g. perceptron or linear programming) can be used to solve it, provided the problem has a solution.

2.4 Learning with Preferences

In earlier sections we have discussed the structure behind the supervision and how it can be modelled using preference graphs. Now, we see how to give learning algorithms which are able to optimize the associated cost functions.

Specifically, the purpose of a GPLM based algorithm will be to minimize costs $c(G|\mathbf{w})$ on a set of preference graphs as in Eq. (1). As these are not continuous w.r.t. the parameter vector $\mathbf{w}$, they can be approximated by introducing a continuous non-increasing loss function $l : \mathbb{R} \to \mathbb{R}^+$ approximating the indicator function. The (approximate) cost will be then defined by

$$\tilde{c}(G|\mathbf{w}) = \sum_{g \in G} \max_{\lambda \in g} \gamma(\lambda) l(\rho(\lambda|\mathbf{w})).$$

Examples of losses one can use are presented in Table 2.

The goal in batch learning is to find the parameters $\mathbf{w}$ such to minimize the expected cost over $\mathcal{D}$, the actual distribution according to which we obtain the

Methods	$l(\rho)$
Perceptron	$\max(0, -\rho)$
β-margin	$\max(0, \beta - \rho)$
Exponential	$e^{-\rho}$
Sigmoidal	$(1 + e^{\lambda(\rho-\theta)})^{-1}$

Table 2: Approximation losses as a function of the margin. $\beta > 0, \lambda > 0, \theta \in \mathbb{R}$ are external parameters.

supervision feedback. Let $\mathcal{G}$ be a function which maps any supervision S into an opportune set of preference graphs, then the expected cost will be

$$R_t[\mathbf{w}] = E_{S\sim\mathcal{D}}[c(\mathcal{G}(S)|\mathbf{w})].$$

Although $\mathcal{D}$ is unknown, we can still try to minimize this expected cost by exploiting the same structure of supervision and as much of the information we can gather from a given training set $\mathcal{S}$. A general learning scheme can be given as in the following:

- Given a set $\mathcal{V}(\mathcal{S}) = \bigcup_{S\in\mathcal{S}} \mathcal{G}(S)$ of preference graphs
- Find a set of parameters $\mathbf{w}$ in such a way to minimize the functional

$$\mathcal{Q}(\mathbf{w}) = \mathcal{R}(\mathbf{w}) + \mu\mathcal{L}(\mathcal{V}(\mathcal{S})|\mathbf{w}) \tag{2}$$

  where $\mathcal{L}(\mathcal{V}(\mathcal{S})|\mathbf{w}) = \sum_{S\in\mathcal{S}} \tilde{c}(\mathcal{G}(S)|\mathbf{w})$ is related to the empirical cost and $\mathcal{R}(\mathbf{w})$ is a regularization term over the set of parameters. Note that, for the solution to be admissible when multiple thresholds are used and there are constraints defined over their values (as in the ordinal regression settings), these constraints should be explicitly enforced.

The use of a regularization term in problems of this type has different motivations, including the theory on regularization networks (see e.g. [7]). Moreover, we can see that by choosing a convex loss function and a convex regularization term (let say the quadratic term $\mathcal{R}(\mathbf{w}) = \frac{1}{2}||\mathbf{w}||^2$) it warranties the convexity of the functional $\mathcal{Q}(\mathbf{w})$ in Eq. 2 and then the uniqueness of the solution. Indeed, current kernel-based approaches defined for basic supervised learning tasks can be seen in this form when using the β-margin with $\beta = 1$. This suggests a new *universal* kernel method which is able to solve many complex learning tasks [8].

3 GPLM Applications

In this section, we briefly discuss recent works we have done which are related to the preference learning framework presented in this paper.

The first application of the preference learning model to a label ranking problem was presented in [9]. The proposed kernel method is able to optimize

ESANN'2009 proceedings, European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning. Bruges (Belgium), 22-24 April 2009, d-side publi., ISBN 2-930307-09-9.

any conjunctive set of label-based preferences, thus being suitable to any label-ranking problem. Moreover, a method is proposed to optimize the embedding of instances and a coding matrix for the set of labels simultaneously.

In [5] there is a comparison of different cost mappings and loss functions for the ordinal regression problem in an on-line setting. The setting on-line and the huge size of data, have motivated a stochastic version of the general algorithm we presented in Section 2.4 for the batch setting. Quite interestingly, experiments have shown that using loss functions which nicely approximate the true evaluation function (0-1 loss function), like the sigmoidal loss for example, gave far better results w.r.t. hinge loss and similar.

In [10] there is an application of the model to an interactive classification context. In such contexts, differently from *autonomous* Text Categorization (TC) systems [11], the system, rather than taking a final categorization decision, is required to support a human expert who is in charge of taking this decision. A real-world application that perfectly fits in this scenario is patent classification [12, 13, 14], where experts at international patent offices are presented with patent applications that they need to classify against a large set of classes of existing patents, in order to check the novelty of the proposed invention. A system that ranks the available classes in terms of their estimated suitability to the document to be classified is extremely useful to these experts since they can thus concentrate on the top-ranked categories, pretty much as a Web searcher concentrates on the top-ranked documents returned by a search engine following a query. Clearly, providing supervision in the form of rankings is more onerous than providing a *crisp* membership value to a document (i.e. being relevant or irrelevant); Then, a mapping which uses membership information to generate subsumed ranking preferences, is used. For example, we can consider a preference like $c_r \rhd_d c_s$ (category c_r is preferred to category c_s for the document d) whenever d belongs to c_r and d does not belong to c_s. Experimental comparison between GPLM and standard SVM has been performed showing a large improvement w.r.t. the precision at recall ranking evaluation measure on the Reuters-21578 benchmark dataset.

In [15] there is an application to a job candidate selection task. In this task, for filling a job role, one or more candidates have to be selected from a pool of candidates. Assume that the $k \geq 1$ most suited candidate for the job are selected. This decision is taken by looking at each candidate professional profile. Moreover, we may assume that the number k of candidates to select is already known from the beginning. This last point is very important to model the problem. In fact, a candidate will be selected on the basis of which other candidates are in the pool. In other words, no decisions can be taken for a candidate without knowing who else is competing for the same position(s). The training set consists of past decisions about promotions to a given role. It follows that this problem can be cast as an k-instance ranking problem.

Finally, in [16] an application to patent classification is proposed. Categories have to be associated to patents according to a three-layered structure (primary, secondary, non-category). A preference model ad hoc for this problem has been

ESANN'2009 proceedings, European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning. Bruges (Belgium), 22-24 April 2009, d-side publi., ISBN 2-930307-09-9.

proposed improving on different baseline strategies.

4 Related Work

Some efforts have been made in the past to give general algorithms for label ranking tasks. In [17] the authors showed that different label ranking problems can be cast as a linear problem which is solvable by a perceptron in an augmented feature space. In [4] the authors proposed a setting in which a label ranking problem was mapped into a set of preference graphs and a convex optimization problem is defined to solve it. More recently, in [6] a large margin method to solve single-label problems with structured output has been proposed. This approach does not seem directly applicable to solve label ranking tasks as it would require an optimization problem with a different constraint for each possible (label) ranking. Far less has been done aiming at generalizing quantitative supervised problems and our framework wanted to fill this gap.

Concerning preference optimization, different paradigms have been studied in the past. Basically, they can be classified in two main families.

In a first family, which is the one our approach belongs, a ranking or *utility model* $f(z)$ is learned which is supposed to respect as much as possible the preferences defined on its values. As another example of an approach in this family is [19], another method to solve general label ranking problems using Gaussian processes. Based on this work, in [21] the same approach is combined with multi-task learning.

In a second family of approaches, namely, the *pairwise preference model*, one tries to directly model the pairwise preferences. In this last approach a function $f(z_1, z_2) \in [0, 1]$ is learned which represents whether z_1 is to prefer to z_2. One drawback of pairwise preference models is that the resulting model could lack of basic properties of a ranking function, for example the model could not respect the transitive property of preferences. A typical method to cope with this kind of ambiguities is to resort to voting strategies. In [18], for example, an interesting framework for label ranking which exploits this idea is given. Using pairwise preference models can require $O(m^2)$ evaluations where m is the number of labels. A nice optimization to drastically reduce this complexity in a multi-label classification task is proposed in [20].

5 Conclusions

We have discussed a general preference model for supervised learning and applications to complex prediction problems, as job selection and patent classification, for example. The interesting aspect of the proposed preference model is that it allows to codify cost functions as preferences and naturally plug them into the same training algorithm. In this view, the role of the cost functions resembles the role of kernels in kernel-machines. Moreover, the proposed method gives a tool for comparing different algorithms and cost functions on a same learning problem.

References

[1] P. McCullagh and J.A. Nelder. *Generalized Linear Models*. Chapman & Hall, 1983.

[2] R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. In *Advances in Large Margin Classifiers*, pages 115–132,. MIT Press, 2000.

[3] H. Wu, H. Lu, and S. Ma. A practical svm-based algorithm for ordinal regression in image retrieval. In *Proceedings of the eleventh ACM international conference on Multimedia*, 2003.

[4] O. Dekel, C.D. Manning, and Y. Singer. Log-linear models for label ranking. In *Advances in Neural Information Processing Systems*, 2003.

[5] F. Aiolli. A preference model for structured supervised learning tasks. In *Proceedings of the 5th IEEE International Conference on Data Mining (ICDM'05)*, pages 557–560, Houston, US, 2005.

[6] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the Twenty-first international conference on Machine learning*, 2004.

[7] T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13:1–50, 2000.

[8] F. Aiolli. *Large Margin Multiclass Learning: Models and Algorithms*. PhD thesis, Dept. of Computer Science, University of Pisa, 2004. http://www.di.unipi.it/˜ aiolli/thesis.ps.

[9] F. Aiolli and A. Sperduti. Preference learning for multiclass problems. In *Advances in Neural Information Processing Systems*. MIT Press, 2004.

[10] F. Aiolli, F. Sebastiani, and A. Sperduti. Preference learning for category-ranking based interactive text categorization. In *Proceedings of the International Joint Conference on Neural Networks*, Orlando, US, 2007.

[11] D. D. Lewis. Evaluating and optmizing autonomous text classification systems. In *Proceedings of 18th ACM International Conference on Research and Development in Information Retrieval (SIGIR'95)*, pages 246–254, Seattle, US, 1995.

[12] C. J. Fall, A. Törcsvári, K. Benzineb, and G. Karetka. Automated categorization in the International Patent Classification. *SIGIR Forum*, 37(1):10–25, 2003.

[13] M. Krier and F. Zaccà. Automatic categorization applications at the european patent office. *World Patent Information*, 24:187–196, 2002.

[14] L. S. Larkey. A patent search and classification system. In *Proceedings of DL-99, 4th ACM Conference on Digital Libraries*, pages 179–187, Berkeley, US, 1999.

[15] F. Aiolli, M. De Filippo, and A. Sperduti. Application of the preference learning model to a human resources selection task. In *Proceedings of the Symposium on Computational Intelligence and Data Mining (CIDM)*.

[16] F. Aiolli, R. Cardin, F. Sebastiani, and A. Sperduti. Preferential text classification: Learning algorithms and evaluation measures. *Information Retrieval*, 2009.

[17] S. Har-Peled, D. Roth, and D. Zimak. Constraint classification for multiclass classification and ranking. In *Advances in Neural Information Processing Systems*, 2002.

[18] E. Hullermeier, J. Furnkranz, W. Cheng, and K. Brinker. Label ranking by learning pairwise preferences. *Artificial Intelligence*, 2008.

[19] W. Chu and Z. Ghaharamani. Preference learning with gaussian processes. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2005.

[20] E.L. Mencia, S.H. Park, and J. Furnkranz. Efficient voting prediction for pair-wise multilabel classification. In *Proceedings of the European Symposium on Artificial Neural Networks (ESANN)*, 2009.

[21] A Birlutiu, P. Groote, and T. Heskes. Multi-task preference learning with gaussian processes. In *Proceedings of the European Symposium on Artificial Neural Networks (ESANN)*, 2009.