# Label Ranking Methods based on the Plackett-Luce Model

**Weiwei Cheng**[1] CHENG@INFORMATIK.UNI-MARBURG.DE
**Krzysztof Dembczyński**[1,2] DEMBCZYNSKI@INFORMATIK.UNI-MARBURG.DE
**Eyke Hüllermeier**[1] EYKE@INFORMATIK.UNI-MARBURG.DE
[1]Mathematics and Computer Science, Marburg University, Hans-Meerwein-Str., 35032 Marburg, Germany
[2]Institute of Computing Science, Poznań University of Technology, Piotrowo 2, 60-965 Poznań, Poland

## Abstract

This paper introduces two new methods for label ranking based on a probabilistic model of ranking data, called the Plackett-Luce model. The idea of the first method is to use the PL model to fit locally constant probability models in the context of instance-based learning. As opposed to this, the second method estimates a global model in which the PL parameters are represented as functions of the instance. Comparing our methods with previous approaches to label ranking, we find that they offer a number of advantages. Experimentally, we moreover show that they are highly competitive to start-of-the-art methods in terms of predictive accuracy, especially in the case of training data with incomplete ranking information.

## 1. Introduction

The problem of label ranking can be considered as a generalization of conventional classification, insofar as a complete ranking of all labels is requested as a prediction instead of only a single class label. Thus, as we shall explain in more detail later on, the label ranking problem consists of learning a mapping from instances to rankings over a finite set of predefined labels.

Several methods for label ranking have already been proposed in the literature; we refer to (Vembu & Gärtner, 2010) for a comprehensive survey. Existing methods for label ranking are mostly reduction techniques transforming the original learning problem into one or several binary classification problems. So-called constraint classification, for example, turns the original problem into a single binary classification problem in an expanded space of higher dimension, and constructs a label ranking model from the classifier learned in that space (Har-Peled et al., 2003). Another approach, ranking by pairwise comparison, reduces the original problem to several small instead of a single large binary classification problem. More specifically, one binary model is learned for each pair of labels, and the predictions of these models are combined into a ranking of all labels (Hüllermeier et al., 2008).

Reduction techniques of this kind have shown promising performance in first experimental studies. Moreover, the reduction of the label ranking problem to the simpler problem of binary classification is appealing for several reasons. Notably, it makes the label ranking problem amenable to the large repertoire of (binary) classification methods and existing algorithms in this field. On the other hand, reduction techniques also come with some disadvantages. In particular, theoretical assumptions on the sought "ranking-valued" mapping, which may serve as a proper learning bias, may not be easily translated into corresponding assumptions for the classification problems. Likewise, it is often not clear (and mostly even wrong) that minimizing the classification error, or a related loss function, on the binary problems is equivalent to maximizing the (expected) performance of the label ranking model in terms of the desired *loss function on rankings* (Hüllermeier & Fürnkranz, 2010).

An alternative approach, which avoids these problems to some extent, was recently put forward in (Cheng et al., 2009). Here, the idea is to develop label ranking methods on the basis of statistical models for ranking data, that is, parameterized (conditional) probability distributions on the class of all rankings. Given assumptions of that kind, the learning problem can be posed as a problem of maximum likelihood estimation (or, alternatively, as a problem of Bayesian inference) and thus be solved in a theoretically sound way.

Appearing in *Proceedings of the 27$^{th}$ International Conference on Machine Learning*, Haifa, Israel, 2010. 

In (Cheng et al., 2009), the authors proposed to use the Mallows model and developed an instance-based (nearest neighbor) learning algorithm to estimate this model in a local way. In this paper, we propose the Plackett-Luce model as an alternative, especially since this model is more apt to learning from possibly incomplete label rankings. Moreover, apart from the estimation of locally constant models suitable for instance-based learning, we also develop a method for estimating generalized linear models.

The paper is organized as follows. Section 2 recalls the problem of label ranking in a more formal setting. In Section 3, we introduce a probability model that will be used for estimating predictive models for rankings. Sections 4 and 5 are devoted, respectively, to the instance-based and generalized linear method for label ranking. An experimental evaluation is presented in Section 6, and Section 7 concludes the paper.

## 2. Label Ranking

In the setting of label ranking, each instance $\boldsymbol{x}$ from an instance space $\mathbb{X}$ is associated with a total order of all class labels, that is, a total, transitive, and asymmetric relation $\succ_{\boldsymbol{x}}$ on $\mathcal{Y}$, where $y_i \succ_{\boldsymbol{x}} y_j$ indicates that $y_i$ precedes $y_j$ in the order. Since a ranking can be considered as a special type of preference relation, we shall also say that $y_i \succ_{\boldsymbol{x}} y_j$ indicates that $y_i$ is *preferred* to $y_j$ given the instance $\boldsymbol{x}$.

Note that, in contrast to the classification scenario, there is no such thing as a "true class label" of an instance under this interpretation. However, depending on the type of application, other interpretations of a label ranking are possible. For example, within the setting of conventional classification, a ranking can be interpreted as a special type of *qualitative probability* on $\mathcal{Y}$ (Wellman, 1994). The order relation $y_i \succ_{\boldsymbol{x}} y_j$ then indicates that the conditional probability of $y_i$ given $\boldsymbol{x}$ is higher than the probability of $y_j$ given $\boldsymbol{x}$, without specifying any concrete numerical values. Given that good numerical estimates are hard to obtain, and sometimes not even needed for decision making, a qualitative representation of this kind is an interesting alternative to numerical distributions.

Formally, a total order $\succ_{\boldsymbol{x}}$ can be identified with a permutation $\pi_{\boldsymbol{x}}$ of the set $\{1,\ldots,M\}$. We define $\pi_{\boldsymbol{x}}$ such that $\pi_{\boldsymbol{x}}(i)$ is the index $j$ of the class label $y_j$ put on the $i$-th position in the order (and hence $\pi_{\boldsymbol{x}}^{-1}(j) = i$ the position of the $j$-th label). This permutation thus encodes the (ground truth) order relation

$$y_{\pi_{\boldsymbol{x}}(1)} \succ_{\boldsymbol{x}} y_{\pi_{\boldsymbol{x}}(2)} \succ_{\boldsymbol{x}} \cdots \succ_{\boldsymbol{x}} y_{\pi_{\boldsymbol{x}}(M)} \ .$$

The class of permutations of $\{1,\ldots,M\}$ (the symmetric group of order $M$) is denoted by $\Omega$. By abuse of terminology, though justified in light of the above one-to-one correspondence, we refer to elements $\pi \in \Omega$ as both permutations and rankings.

In analogy with the classification setting, we do not assume the existence of a deterministic $\mathbb{X} \longrightarrow \Omega$ mapping. Instead, every instance is associated with a *probability distribution* over $\Omega$. This means that, for each $\boldsymbol{x} \in \mathbb{X}$, there exists a probability distribution $\mathbf{P}(\cdot \,|\, \boldsymbol{x})$ such that, for every $\pi \in \Omega$, $\mathbf{P}(\pi \,|\, \boldsymbol{x})$ is the probability that $\pi_{\boldsymbol{x}} = \pi$. (Note that, if rankings are interpreted as qualitative probabilities, then $\mathbf{P}(\cdot \,|\, \boldsymbol{x})$ is a probability over probability distributions, i.e., a second-order probability.)

The goal in label ranking is to learn a "label ranker" in the form of an $\mathbb{X} \longrightarrow \Omega$ mapping. As training data, a label ranker uses a set of instances $\boldsymbol{x}_n$ $(n = 1,\ldots,N)$, together with information about the associated rankings $\pi_{\boldsymbol{x}_n}$. Ideally, complete rankings are given as training information. From a practical point of view, however, it is important to allow for incomplete information in the form of a ranking

$$y_{\pi_{\boldsymbol{x}}(1)} \succ_{\boldsymbol{x}} y_{\pi_{\boldsymbol{x}}(2)} \succ_{\boldsymbol{x}} \cdots \succ_{\boldsymbol{x}} y_{\pi_{\boldsymbol{x}}(k)} \ , \tag{1}$$

where $k < M$ and $\{\pi(1),\ldots,\pi(k)\} \subset \{1,\ldots,M\}$. For example, for an instance $\boldsymbol{x}$, it might be known that $y_2 \succ_{\boldsymbol{x}} y_1 \succ_{\boldsymbol{x}} y_5$, while no preference information is given about the labels $y_3$ or $y_4$. By definition, we let $\pi^{-1}(y_i) = \pi^{-1}(i) = 0$ if $y_i$ is not present in the ranking $\pi$; thus, the presence of a class $y_i$ is equivalent to $\pi^{-1}(i) > 0$.

To evaluate the predictive performance of a label ranker, a suitable loss function on $\Omega$ is needed. In the statistical literature, several distance measures for rankings have been proposed. For example, a commonly used measure is Kendall's tau coefficient, defined as

$$\frac{C(\pi,\sigma) - D(\pi,\sigma)}{M(M-1)/2} \tag{2}$$

with $C(\pi,\sigma)$ the number of concordant label pairs (i.e., pairs $(i,j) \in \{1,\ldots,M\}^2$ such that $(\pi(i) - \pi(j)) \cdot (\sigma(i) - \sigma(j)) > 0$) and $D(\pi,\sigma)$ the number of discordant pairs ($(i,j)$ with $(\pi(i) - \pi(j)) \cdot (\sigma(i) - \sigma(j)) < 0$). Actually, (2) is not a loss function but a correlation measure with values in $[-1,+1]$ (it assumes the value 1 if $\sigma = \pi$ and the value $-1$ if $\sigma$ is the reversal of $\pi$).

## 3. Ranking Models

So far, no assumptions about the conditional probability measure $\mathbf{P}(\cdot \,|\, \boldsymbol{x})$ on $\Omega$ were made, despite its existence. In statistics, different types of probability

distributions on rankings have been proposed (Marden, 1995).

A prominent example is the Mallows model (Mallows, 1957), a *distance-based* probability model belonging to the family of exponential distributions. The standard Mallows model is determined by two parameters:

$$\mathbf{P}(\pi \,|\, \theta, \pi_0) = \frac{\exp(-\theta D(\pi, \pi_0))}{\phi(\theta)} \tag{3}$$

The ranking $\pi_0 \in \Omega$ is the location parameter (mode, center ranking) and $\theta \geq 0$ is a spread parameter. Moreover, $D(\cdot)$ is a distance measure on rankings, and the constant $\phi = \phi(\theta)$ is a normalization factor that depends on the spread (but, provided the right-invariance of $D(\cdot)$, not on $\pi_0$).

Obviously, the Mallows model assigns the maximum probability to the center ranking $\pi_0$. The larger the distance $D(\pi, \pi_0)$, the smaller the probability of $\pi$ becomes. The spread parameter $\theta$ determines how quickly the probability decreases, i.e., how peaked the distribution is around $\pi_0$. For $\theta = 0$, the uniform distribution is obtained, while for $\theta \to \infty$, the distribution converges to the one-point distribution that assigns probability 1 to $\pi_0$ and 0 to all other rankings.

The Mallows model was used in (Cheng et al., 2009) in the context of an instance-based approach to label ranking. Notwithstanding some appealing properties, the Mallows model is arguably not ideal for handling incomplete training data, i.e., observations in the form of incomplete rankings (1). Roughly speaking, this is because the probability of such a ranking cannot be expressed in closed form. Instead, it has to be derived through marginalization:

$$\mathbf{P}(\pi \,|\, \theta, \pi_0) \;=\; \sum_{\pi^* \in E(\pi)} \mathbf{P}(\pi^* \,|\, \theta, \pi_0) \;\;,$$

where $E(\pi)$ denotes the set of linear extensions of $\pi$.[1] Consequently, inference (maximum likelihood estimation of $\pi_0$ and $\theta$) is difficult for incomplete observations and becomes computationally complex. The normalization factor $\phi(\theta)$ in (3) causes additional problems.

Another model that seems to be more appropriate from this point of view is the Plackett-Luce (PL) model, which is specified by a parameter vector $\boldsymbol{v} = (v_1, v_2, \ldots v_M) \in \mathbb{R}_+^M$:

$$\mathbf{P}(\pi \,|\, \boldsymbol{v}) \;=\; \prod_{i=1}^{M} \frac{v_{\pi(i)}}{v_{\pi(i)} + v_{\pi(i+1)} + \ldots + v_{\pi(M)}} \tag{4}$$

[1] A permutation $\pi^* \in \Omega$ is a linear extension of $\pi$ if it ranks all labels occurring in $\pi$ in the same order.

This model is a generalization of the well-known Bradley-Terry model, a model for the pairwise comparison of alternatives, which specifies the probability that "$a$ wins against $b$" in terms of

$$\mathbf{P}(a \succ b) \;=\; \frac{v_a}{v_a + v_b} \;\;.$$

Obviously, the larger $v_a$ in comparison to $v_b$, the higher the probability that $a$ is chosen. Likewise, the larger the parameter $v_i$ in (4) in comparison to the parameters $v_j$, $j \neq i$, the higher the probability that the label $y_i$ appears on a top rank. An intuitively appealing explanation of the PL model can be given in terms of a vase model: If $v_i$ corresponds to the relative frequency of the $i$-th label in a vase filled with labeled balls, then $\mathbf{P}(\pi \,|\, \boldsymbol{v})$ is the probability to produce the ranking $\pi$ by randomly drawing balls from the vase in a sequential way and putting the label drawn in the $k$-th trial on position $k$ (unless the label was already chosen before, in which case the trial is annulled).

For the PL model, one easily verifies that the probability of an incomplete ranking (1) is given by

$$\mathbf{P}(\pi \,|\, \boldsymbol{v}) \;=\; \prod_{i=1}^{k} \frac{v_{\pi(i)}}{v_{\pi(i)} + v_{\pi(i+1)} + \ldots + v_{\pi(k)}} \;\;,$$

i.e., by an expression of exactly the same form, except that the number of factors is $k$ (the number of labels observed) instead of $M$. In a different though related context, the use of the PL model for machine learning was recently motivated in (Guiver & Snelson, 2009).

## 4. Instance-Based Label Ranking

In this section, we propose an instance-based approach to label ranking, i.e., a local prediction method based on the nearest neighbor estimation principle. Consider a query instance $\boldsymbol{x} \in \mathbb{X}$ and let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K$ denote the nearest neighbors of $\boldsymbol{x}$ (according to an underlying distance measure on $\mathbb{X}$) in the training set, where $K \in \mathbb{N}$ is a fixed integer. Each neighbor $\boldsymbol{x}_i$ $(i = 1, \ldots, K)$ is associated with a possibly incomplete ranking $\pi_i$ of the labels $y \in \mathcal{Y}$. We denote by $M_i \in \{2, \ldots, M\}$ the number of labels ranked by $\pi_i$. Moreover, recall that $\pi_i(m)$ denotes the index of the label ranked on position $m$.

In analogy to the conventional settings of classification and regression, in which the nearest neighbor estimation principle has been applied for a long time, we assume that the probability distribution $\mathbf{P}(\cdot \,|\, \boldsymbol{x})$ on $\Omega$ is (at least approximately) *locally constant* around the query $\boldsymbol{x}$. By furthermore assuming that the rankings $\pi_i$ have been produced independently of each other by

the PL model (4), the probability to observe the rankings $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_K\}$ in the neighborhood, given the parameters $\boldsymbol{v} = (v_1, \ldots, v_M)$, becomes

$$\mathbf{P}(\boldsymbol{\pi} \mid \boldsymbol{v}) = \prod_{i=1}^{K} \prod_{m=1}^{M_i} \frac{v_{\pi_i(m)}}{\sum_{j=m}^{M_i} v_{\pi_i(j)}} \ . \tag{5}$$

The maximum likelihood estimation (MLE) of $\boldsymbol{v}$ is then given by those parameters that maximize this probability or, equivalently, the log-likelihood function

$$L(\boldsymbol{v}) = \sum_{i=1}^{K} \sum_{m=1}^{M_i} \left[ \log\left(v_{\pi_i(m)}\right) - \log \sum_{j=m}^{M_i} v_{\pi_i(j)} \right] .$$

### 4.1. Maximum Likelihood Estimation

Finding the MLE parameters of the PL model is a problem that has already been considered in the statistical literature. We resort to an algorithm called MM, which is short for Minorization and Maximization. MM seems to perform especially well for this problem (Hunter, 2004). It is an iterative algorithm whose idea is to maximize, in each iteration, a function that minorizes the original log-likelihood, namely

$$Q_k(\boldsymbol{v}) = \sum_{i=1}^{K} \sum_{m=1}^{M_i} \left[ \log\left(v_{\pi_i(m)}\right) - \frac{\sum_{j=m}^{M_i} v_{\pi_i(j)}}{\sum_{j=m}^{M_i} v^{(k)}_{\pi_i(j)}} \right] .$$

Here, $\boldsymbol{v}^{(k)} = (v_1^{(k)}, \ldots, v_M^{(k)})$ is the estimation of the PL parameters in the $k$-th iteration. Considering these values as fixed, the problem to maximize $Q_k(\cdot)$ as a function of $\boldsymbol{v}$ can be solved analytically. The corresponding solution, i.e., the parameter vector $\boldsymbol{v}^*$ for which $Q_k(\cdot)$ is maximal, is then used as a new solution: $\boldsymbol{v}^{(k+1)} = \boldsymbol{v}^*$. This procedure provably converges to an MLE estimation of the PL parameters.

### 4.2. Prediction

Given the MLE $\boldsymbol{v}^*$, a prediction of the ranking associated with $\boldsymbol{x}$ can be derived from the distribution $\mathbf{P}(\cdot \mid \boldsymbol{v}^*)$ on $\Omega$. In particular, a MAP estimate, i.e., a ranking with the highest posterior probability, is given by

$$\pi^* \ \in \ \arg\max_{\pi \in \Omega} \mathbf{P}(\pi \mid \boldsymbol{v}^*) \ . \tag{6}$$

A ranking of this kind can easily be produced by sorting the labels $y_i$ in decreasing order according to the respective parameters $v_i^*$, i.e., such that

$$v_{\pi^*(i)} \geq v_{\pi^*(j)} \tag{7}$$

for all $1 \leq i < j \leq M$. More generally, given a loss function $\ell(\cdot)$ to be minimized, the best prediction is

$$\pi^* \ = \ \arg\min_{\pi \in \Omega} \sum_{\tau \in \Omega} \ell(\pi, \tau) \cdot \mathbf{P}(\tau \mid \boldsymbol{v}^*) \ . \tag{8}$$

In general, an interesting question concerns the complexity of the minimization problem (8). An explicit computation of the expected loss for each ranking $\pi$ is feasible only for small label sets $\mathcal{Y}$, since the cardinality of $\Omega$, which is given by $|\Omega| = |\mathcal{Y}|! = M!$, grows very fast. However, depending on the loss function $\ell(\cdot)$ and the probability distribution $\mathbf{P}(\cdot \mid \boldsymbol{v}^*)$, an explicit enumeration of this type can often be avoided.

The PL model appears to be especially appealing from this point of view. In fact, due to the special structure of the probability distribution (4), a ranking of the form (7) is not only the most intuitive prediction, but also provably optimal for virtually all common loss functions on rankings. Without going into technical details here, we only mention that, in particular, it is a risk minimizer for the 0/1 loss function (defined by $\ell(\pi^*, \pi) = 0$ if $\pi^* = \pi$ and $= 1$ if $\pi^* \neq \pi$) and, likewise, a maximizer of the expected rank correlation in terms of (2).

In contrast to other methods (including most reduction techniques) that simply produce a prediction in terms of a ranking, a probabilistic approach to label ranking allows one to complement predictions by diverse types of statistical information, for example regarding the reliability of a prediction. Besides, the distribution $\mathbf{P}(\cdot \mid \boldsymbol{v}^*)$ supports various types of generalized predictions, such as credible sets of rankings covering the true one with a high probability.

## 5. Generalized Linear Models

The learning method proposed in the previous section is local (and lazy) in the sense that an individual PL model, i.e., an individual parameter vector $\boldsymbol{v} = (v_1, \ldots, v_M)$, is estimated for each query instance $\boldsymbol{x} \in \mathbb{X}$. In this section, we consider the estimation of a *global* model as an alternative. To this end, we model the parameters $v_m$, quantifying the propensity for the $m$-th label $y_m$, i.e., the tendency to put this label on a high rank, as a linear function of the attributes describing an instance. More precisely, to guarantee the non-negativity of the parameters, we model their logarithm as a linear function:

$$v_m \ = \ \exp\left( \sum_{d=1}^{D} \alpha_d^{(m)} \cdot x_d \right) \ , \tag{9}$$

where we assume an instance to be represented in terms of a feature vector $\boldsymbol{x} = (x_1, \ldots, x_D) \in \mathbb{X} = \mathbb{R}^D$.

The model parameters to be estimated are now the $\alpha_d^{(m)}$ $(1 \leq m \leq M,\ 1 \leq d \leq D)$. Given a training data set

$$\mathcal{T} \ = \ \left\{ \left( \boldsymbol{x}^{(n)}, \pi^{(n)} \right) \right\}_{n=1}^{N}$$

with $\boldsymbol{x}^{(n)} = \left(x_1^{(n)}, \ldots, x_D^{(n)}\right)$, the log-likelihood function is given by

$$L = \sum_{n=1}^{N} \left[ \sum_{m=1}^{M_n} \log\left(v(\pi^{(n)}(m), n)\right) - \log \sum_{j=m}^{M_n} v(\pi^{(n)}(j), n) \right],$$

where $M_n$ is the number of labels in the ranking $\pi^{(n)}$, and

$$v(m, n) = \exp\left(\sum_{d=1}^{D} \alpha_d^{(m)} \cdot x_d^{(n)}\right) .$$

The first derivatives of $L(\cdot)$ are given by

$$\frac{\partial L}{\partial \alpha_k^{(a)}} = \sum_{n=1}^{N} \delta(a, n, 1) \cdot x_k^{(n)} - \sum_{n=1}^{N} \sum_{m=1}^{M_n} \delta(a, n, m) \cdot \frac{v(a, n) \cdot x_k^{(n)}}{\sum_{j=m}^{M_n} v(\pi^{(n)}(j), n)},$$

where

$$\delta(a, n, m) = \begin{cases} 1 & (\pi^{(n)})^{-1}(a) \geq m \\ 0 & \text{otherwise} \end{cases} .$$

Moreover, the second derivatives (for $a \neq b$, $k \neq \ell$) are as follows:

$$\frac{\partial^2 L}{\partial \left(\alpha_k^{(a)}\right)^2} = -\sum_{n=1}^{N} \sum_{m=1}^{M_n} \delta(a, n, m) \cdot v(a, n) \cdot \left(x_k^{(n)}\right)^2 \cdot \frac{\left[\sum_{j=m}^{M_n} v(\pi^{(n)}(j), n) - v(a, n)\right]}{\left(\sum_{j=m}^{M_n} v(\pi^{(n)}(j), n)\right)^2}$$

$$\frac{\partial^2 L}{\partial \alpha_k^{(a)} \partial \alpha_\ell^{(a)}} = -\sum_{n=1}^{N} \sum_{m=1}^{M_n} \delta(a, n, m) \cdot v(a, n) \cdot x_k^{(n)} \cdot x_\ell^{(n)} \cdot \frac{\left[\sum_{j=m}^{M_n} v(\pi^{(n)}(j), n) - v(a, n)\right]}{\left(\sum_{j=m}^{M_n} v(\pi^{(n)}(j), n)\right)^2}$$

$$\frac{\partial^2 L}{\partial \alpha_k^{(a)} \partial \alpha_\ell^{(b)}} = \sum_{n=1}^{N} \sum_{m=1}^{M_n} \delta(a, n, m) \cdot \delta(b, n, m) \cdot \frac{v(a, n) \cdot x_k^{(n)} \cdot v(b, n) \cdot x_\ell^{(n)}}{\left(\sum_{j=1}^{M_n} v(\pi^{(n)}(j), n)\right)^2}$$

Note that $\partial^2 L / \partial (\alpha_k^{(a)})^2 \leq 0$ for all $1 \leq a \leq M$ and $1 \leq k \leq D$. Based on these derivatives, the maximization of the log-likelihood can be accomplished by means of gradient-based optimization methods. In our implementation, we use a standard stochastic gradient descent algorithm (Bottou, 2004) that, in terms of efficiency, compared quite favorably with other gradient-based methods.

## 6. Experimental Evaluation

In this section, we present an empirical evaluation of our instance-based (IB-PL) and generalized linear (Lin-PL) approach to label ranking using the PL model. For comparison, we include two other methods, namely the aforementioned instance-based approach using the Mallows model (IB-Mal) and so-called log-linear models for label ranking (Lin-LL) as a representative of the class of linear models.

To guarantee a fair comparison, we used the Euclidean distance (after normalizing the attributes) as a distance measure on the instance space for both IB-PL and IB-Mal (and disabled distance-weighting in IB-Mal). The neighborhood size $K \in \{5, 10, 15, 20\}$ was selected through cross validation on the training set.

Log-linear models for label ranking have been proposed in (Dekel et al., 2004). In this approach, utility functions $f_i(\cdot)$ are expressed as linear combinations of so-called base ranking functions (which map instance/label pairs to real numbers). As a special case, this includes functions of the form $f_i(\boldsymbol{x}) = \sum_{d=1}^{D} \alpha_d x_d$, which should be specified so as to minimize the number of ranking errors. A ranking error for an instance $\boldsymbol{x}$ occurs if $f_i(\boldsymbol{x}) < f_j(\boldsymbol{x})$ even though the $i$-th label should precede the $j$-th label, so the total number of errors on $\boldsymbol{x}$ is

$$\sum_{1 \leq i \leq j \leq M} \begin{cases} 1 & f_{\pi(i)}(\boldsymbol{x}) < f_{\pi(j)}(\boldsymbol{x}) \\ 0 & f_{\pi(i)}(\boldsymbol{x}) \geq f_{\pi(j)}(\boldsymbol{x}) \end{cases}$$

Since minimizing this error (or, more precisely, the sum of this error over all training instances) directly is intractable, the authors propose to minimize a smooth, convex upper bound:

$$\log\left[1 + \sum_{1 \leq i \leq j \leq M} \exp\left(f_{\pi(j)}(\boldsymbol{x}) - f_{\pi(i)}(\boldsymbol{x})\right)\right]$$

Algorithmically, this optimization problem is approached by means of a boosting-based algorithm that works in an iterative way. In (Hüllermeier et al., 2008), it was shown that this approach is quite comparable, in terms of predictive accuracy, to other state-of-the-art methods for label ranking.

### 6.1. Data

Since benchmark data for the label ranking problem is still not available, we resorted to multi-class and regression data sets from the UCI repository and the Statlog collection and turned them into label ranking data in two different ways. (A) For classification data, we followed the procedure proposed in (Hüllermeier et al., 2008): A naive Bayes classifier is first trained on the complete data set. Then, for each example, all the labels present in the data set are ordered with respect to the predicted class probabilities (in the case of ties, labels with lower index are ranked first). This way of generating label ranking data is in line with the interpretation of rankings in terms of qualitative probabilities, as discussed in Section 2. (B) For regression data, a certain number of (numerical) attributes is removed from the set of predictors, and each one is considered as a label. To obtain a ranking, the attributes are standardized and then ordered by size. Given that the original attributes are correlated, the remaining predictive features will contain information about the ranking thus produced. Yet, as will be confirmed by the experimental results, this second type of data generation leads to more difficult learning problems. A summary of the data sets and their properties is given in Table 1.[2]

*Table 1.* Data sets and their properties (the type refers to the way in which the data has been generated).

| data set | type | # inst. | # attr. | # labels |
|---|---|---|---|---|
| authorship | A | 841 | 70 | 4 |
| bodyfat | B | 252 | 7 | 7 |
| calhousing | B | 20640 | 4 | 4 |
| cpu-small | B | 8192 | 6 | 5 |
| elevators | B | 16599 | 9 | 9 |
| fried | B | 40769 | 9 | 5 |
| glass | A | 214 | 9 | 6 |
| housing | B | 506 | 6 | 6 |
| iris | A | 150 | 4 | 3 |
| pendigits | A | 10992 | 16 | 10 |
| segment | A | 2310 | 18 | 7 |
| stock | B | 950 | 5 | 5 |
| vehicle | A | 846 | 18 | 4 |
| vowel | A | 528 | 10 | 11 |
| wine | A | 178 | 13 | 3 |
| wisconsin | B | 194 | 16 | 16 |

### 6.2. Experiments and Results

Results were derived in terms of Kendall's tau coefficient from five repetitions of a ten-fold cross-validation. To model incomplete observations, we modified the training data as follows: A biased coin was flipped for every label in a ranking to decide whether to keep or delete that label; the probability for a deletion is specified by a parameter $p \in [0, 1]$. Hence, $p \times 100\%$ of the labels will be missing on average.

The summary of the results is shown in Table 2. To analyze these results, we followed the two-step procedure recommended in (Demsar, 2006), consisting of a Friedman test of the null hypothesis that all learners have equal performance and, in case this hypothesis is rejected, a Nemenyi test to compare learners in a pairwise way. Both tests are based on the average ranks (for each problem, the methods are ranked in decreasing order of performance, and the ranks thus obtained are averaged over the problems) as shown in the bottom line in Table 2. At a significance level of 5%, IB-PL and IB-Mal are better than Lin-LL in the case of complete rankings, whereas the Friedman test does not discover significant differences in the case of 30% and 60% missing labels.

*Table 3.* Pairwise comparisons of the methods in terms of win/win/win statistics: Wins in the complete ranking scenario, in the 30% and in the 60% missing label scenario.

| | IB-PL | IB-Mal | Lin-PL | Lin-LL |
|---|---|---|---|---|
| IB-PL | — | 6/11/11 | 12/8/7 | 13/11/9 |
| IB-Mal | 10/5/5 | — | 11/8/7 | 12/9/7 |
| Lin-PL | 4/8/9 | 5/8/9 | — | 14/13/11 |
| Lin-LL | 3/5/7 | 4/7/9 | 2/4/5 | — |

Despite being statistically non-significant most of the time, the results are still quite informative and show some important trends (which are likely to become significant when increasing the number of data sets). This becomes especially obvious from the pairwise comparisons of the methods, summarized in Table 3. From these comparisons, the following conclusions can be drawn:

- Regarding the two instance-based learners, IB-PL performs a bit worse in the complete ranking scenario, but is better in the case of missing label information. This is in perfect agreement with our conjecture that the PL model is better suited for learning from incomplete ranking data.
- Comparing the two generalized linear approaches, our method based on the PL model seems to be consistently better than Lin-LL (winning 14, 13 and 11 of the 16 data sets in the three scenarios, respectively).
- Comparing the instance-based with the linear methods, it can be seen that the former perform

[2]The data sets, along with a description, are available at `http://www.uni-marburg.de/fb12/kebi/research/`.

*Table 2.* Performance of the label ranking methods in terms of Kendall's tau (in brackets the rank).

| | complete ranking | | | | 30% missing labels | | | | 60% missing labels | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | IB-PL | IB-Mal | Lin-PL | Lin-LL | IB-PL | IB-Mal | Lin-PL | Lin-LL | IB-PL | IB-Mal | Lin-PL | Lin-LL |
| authorship | .936(1) | .936(2) | .930(3) | .657(4) | .927(1) | .913(2) | .899(3) | .656(4) | .886(1) | .849(2) | .846(3) | .650(4) |
| bodyfat | .230(3) | .229(4) | .272(1) | .266(2) | .204(3) | .198(4) | .266(1) | .251(2) | .151(4) | .160(3) | .222(2) | .241(1) |
| calhousing | .326(2) | .344(1) | .220(4) | .223(3) | .303(2) | .310(1) | .229(3) | .223(4) | .259(2) | .263(1) | .229(3) | .221(4) |
| cpu-small | .495(2) | .496(1) | .426(3) | .419(4) | .477(1) | .473(2) | .418(4) | .419(3) | .437(1) | .428(2) | .412(4) | .418(3) |
| elevators | .721(2) | .727(1) | .712(3) | .701(4) | .702(2) | .683(4) | .706(1) | .699(3) | .633(3) | .596(4) | .704(1) | .696(2) |
| fried | .894(4) | .900(3) | .996(1) | .989(2) | .861(3) | .850(4) | .993(1) | .989(2) | .797(3) | .777(4) | .990(1) | .987(2) |
| glass | .841(2) | .842(1) | .825(3) | .818(4) | .809(3) | .776(4) | .825(1) | .817(2) | .675(3) | .611(4) | .807(2) | .808(1) |
| housing | .711(2) | .736(1) | .659(3) | .626(4) | .654(3) | .669(1) | .658(2) | .625(4) | .492(4) | .543(3) | .636(1) | .614(2) |
| iris | .960(1) | .925(2) | .832(3) | .818(4) | .926(1) | .867(2) | .823(3) | .804(4) | .868(1) | .799(2) | .778(3) | .768(4) |
| pendigits | .939(2) | .941(1) | .909(3) | .814(4) | .918(1) | .902(3) | .909(2) | .802(4) | .794(2) | .781(4) | .907(1) | .787(3) |
| segment | .950(1) | .802(4) | .902(2) | .810(3) | .874(2) | .735(4) | .895(1) | .806(3) | .674(3) | .612(4) | .888(1) | .801(2) |
| stock | .922(2) | .925(1) | .710(3) | .696(4) | .877(1) | .855(2) | .701(3) | .691(4) | .740(1) | .724(2) | .687(4) | .689(3) |
| vehicle | .859(1) | .855(2) | .838(3) | .770(4) | .838(1) | .822(2) | .817(3) | .769(4) | .765(2) | .736(4) | .804(1) | .764(3) |
| vowel | .851(2) | .882(1) | .586(4) | .601(3) | .785(2) | .810(1) | .581(4) | .598(3) | .588(3) | .638(1) | .575(4) | .591(2) |
| wine | .947(2) | .944(3) | .954(1) | .942(4) | .926(4) | .930(3) | .931(2) | .941(1) | .907(2) | .893(4) | .915(1) | .894(3) |
| wisconsin | .479(4) | .501(3) | .635(1) | .542(2) | .453(4) | .464(3) | .615(1) | .533(2) | .381(4) | .399(3) | .585(1) | .518(2) |
| Avg. Rank | 2.06 | 1.94 | 2.56 | 3.44 | 2.13 | 2.63 | 2.19 | 3.06 | 2.44 | 2.94 | 2.06 | 2.56 |

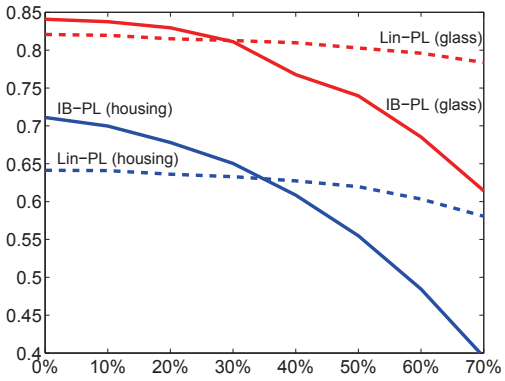


*Figure 1.* Ranking performance (in terms of Kendall's tau) as a function of the missing label rate.

a bit better in the complete ranking scenario, but their performance drops more quickly in the presence of missing label information.

This last observation is plausible, too, and coherent with the complementary nature of global and local methods. Like in the case of conventional classification, instance-based methods are advantageous for problems requiring complex decision boundaries, for which the strong bias of linear methods prevents them from achieving a good separation. On the other hand, if the linearity assumption is (at least approximately) valid, better models can be learned with fewer data. Correspondingly, instance-based learners are more sensitive toward the amount of training data. Some evidence in favor of this hypothesis is indeed provided by the learning curves depicting the performance as a function of the fraction of missing label information. While the learning curves of the linear methods are often rather flat, showing a kind of saturation effect, they are steeper for the instance-based approaches. This suggests that additional label information is still beneficial for these methods even when the linear methods, due to a lack of flexibility, are no longer able to exploit and adapt to extra data. Typical examples for the glass and the housing data are shown in Fig. 1.

As a nice feature of our approach, we already mentioned the possibility to complement a prediction by a measure of reliability. Perhaps the simplest measure of this kind is the probability of the prediction itself, namely $p^* = \mathbf{P}(\pi^* \,|\, \boldsymbol{v}^*)$. To test whether $p^*$ is indeed a good indicator of the uncertainty of a prediction, we used it to compute a kind of *accuracy-rejection* curve: Using IB-PL in a leave-one-out cross validation, we computed the accuracy of the prediction (in terms of Kendall's tau) and its reliability (in terms of $p^*$) for each instance $\boldsymbol{x}$. Subsequent to sorting the instances in decreasing order of reliability, we plot the function $t \mapsto f(t)$, where $f(t)$ is the mean accuracy of the top $t$ percent of the instances. Given that $p^*$ is indeed a good indicator of reliability, this curve should be decreasing, because the higher $t$, the more instances with a low reliability are taken into consideration. This expectation is indeed confirmed for our data sets. Fig. 2 shows two exemplary curves for the glass and the housing data.

## 7. Conclusions and Future Work

Using the Plackett-Luce model as a model of the underlying data generating process, we proposed new methods for the problem of label ranking. The idea of our first approach, an instance-based learning algorithm, is to fit a locally constant model in the neighborhood of a query instance. The same idea was already proposed earlier, using the Mallows model instead of the PL model. As we have seen, the performance of the approaches is quite comparable. However, the PL model seems preferable in the case of incomplete train-

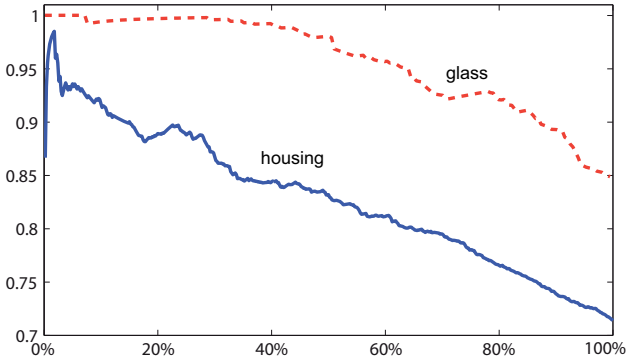


*Figure 2.* Accuracy-rejection curves computed on the basis of $\mathbf{P}(\pi^* \mid \boldsymbol{v}^*)$.

ing data, not only computationally but also regarding performance.

The idea of parameterizing the coefficients of the PL model and expressing them as functions of the input attributes has led to the second approach, namely fitting a global model in the form of log-linear (utility) functions. Empirically, we have shown that it compares favorably with other methods for label ranking, which are closely related in the sense of fitting the same type of model.

Perhaps even more importantly, however, we consider our approach as more solid from a theoretical point of view. In fact, while existing methods are fitting models based on criteria that are to some extent ad-hoc, our probabilistic model provides the basis for a theoretically sound prediction procedure in the form of maximum likelihood estimation. Apart from making model assumptions more explicit, it also has further advantages. For example, it allows for complementing predictions by diverse types of statistical information, for example regarding the reliability of an estimation.

For future work, we plan to combine the two methods presented in this paper, the local and the global one, into a local linear learning method. Similar to methods like local linear regression, the idea is to estimate a (generalized) linear model in a local way, i.e., in the neighborhood of a query instance. From the point of view of our instance-based approach, this means replacing the assumption of a locally constant model by the relaxed assumption of an approximately linear model. Thus, we hope to combine the advantages of both approaches.

## Acknowledgments

The authors gratefully acknowledge financial support by the Germany Research Foundation (DFG).

## References

Bottou, L. Stochastic learning. In Bousquet, O. and von Luxburg, U. (eds.), *Advanced Lectures on Machine Learning*, pp. 146–168. Springer-Verlag, Berlin, 2004.

Cheng, W., Hühn, J., and Hüllermeier, E. Decision tree and instance-based learning for label ranking. In *Proc. ICML–2009*, Montreal, Canada, 2009.

Dekel, O., Manning, CD., and Singer, Y. Log-linear models for label ranking. In Thrun, S., Saul, LK., and Schölkopf, B. (eds.), *Advances in Neural Information Processing Systems (NIPS-2003)*. MIT Press, 2004.

Demsar, J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

Guiver, J. and Snelson, E. Bayesian inference for Plackett-Luce ranking models. In *Proc. ICML–2009*, Montreal, Canada, 2009.

Har-Peled, S., Roth, D., and Zimak, D. Constraint classification for multiclass classification and ranking. In Becker, Suzanna, Thrun, Sebastian, and Obermayer, Klaus (eds.), *Advances in Neural Information Processing Systems 15*, pp. 785–792. MIT Press, 2003.

Hüllermeier, E. and Fürnkranz, J. On loss functions in label ranking and risk minimization by pairwise learning. *Journal of Computer and System Sciences*, 76(1):49–62, 2010.

Hüllermeier, E., Fürnkranz, J., Cheng, W., and Brinker, K. Label ranking by learning pairwise preferences. *Artificial Intelligence*, 172:1897–1917, 2008.

Hunter, D.R. MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics*, 32(1):384–406, 2004.

Mallows, C. Non-null ranking models. *Biometrika*, 44 (1):114–130, 1957.

Marden, J. *Analyzing and Modeling Rank Data*. CRC Press, 1995.

Vembu, S. and Gärtner, T. Label ranking: a survey. In Fürnkranz, J. and Hüllermeier, E. (eds.), *Preference Learning*. Springer-Verlag, 2010.

Wellman, M.P. Some varieties of qualitative probability. In *Proc. IPMU-94*, pp. 437–442, Paris, 1994.