A New Instance-Based Label Ranking Approach Using the Mallows Model

Weiwei Cheng and Eyke Hüllermeier

Mathematics and Computer Science
University of Marburg, Germany
{cheng,eyke}@mathematik.uni-marburg.de

Abstract. In this paper, we introduce a new instance-based approach to the label ranking problem. This approach is based on a probability model on rankings which is known as the Mallows model in statistics. Probabilistic modeling provides the basis for a theoretically sound prediction procedure in the form of maximum likelihood estimation. Moreover, it allows for complementing predictions by diverse types of statistical information, for example regarding the reliability of an estimation. Empirical experiments show that our approach is competitive to start-of-the-art methods for label ranking and performs quite well even in the case of incomplete ranking information.

Key words: Instance-based learning, Label ranking, Classification, Maximum likelihood estimation

1 Introduction

The topic of learning preferences has attracted increasing attention in the recent machine learning literature [1]. Label ranking, a particular preference learning scenario, studies the problem of learning a mapping from instances to rankings over a finite number of predefined labels. It can be considered as a natural generalization of the conventional classification problem, where only a single label is requested instead of a ranking of all labels.

Various approaches for label ranking have been proposed in recent years. Typically, these are extensions of learning algorithms used in binary classification problems. Ranking by pairwise comparison (RPC) is a natural extension of pairwise classification, in which binary preference models are learned for each pair of labels, and the predictions of these models are combined into a ranking of all labels [1]. Two other approaches, constraint classification (CC) and log-linear models for label ranking (LL), seek to learn linear utility functions for each individual label instead of preference predicates for pairs of labels [2, 3].

In this paper, we are interested in an alternative to model-based approaches, namely the use of an *instance-based* approach. Instance-based or case-based learning algorithms have been applied successfully in various fields, such as machine learning and pattern recognition, for a long time [4]. These algorithms simply store the training data, or at least a selection thereof, and defer the

processing of this data until an estimation for a new instance is requested, a property distinguishing them from typical model-based approaches. Instance-based approaches therefore have a number of potential advantages, especially in the context of the label ranking problem.

As a particular advantage of delayed processing, these learning methods may estimate the target function *locally* instead of inducing a global prediction model for the entire input domain (instance space) $\mathbb{X}$. Predictions are typically obtained using only a small, locally restricted subset of the entire training data, namely those examples that are close to the query $\boldsymbol{x} \in \mathbb{X}$ (hence $\mathbb{X}$ must be endowed with a distance measure). These examples are then *aggregated* in a reasonable way. As aggregating a finite set of objects from an output space Ω is often much simpler than representing a complete $\mathbb{X} \to \Omega$ mapping in an explicit way, instance-based methods are especially appealing if Ω has a complex structure.

In label ranking, Ω corresponds to the set of all rankings of an underlying label set $\mathcal{L}$. To represent an Ω-valued mapping, the aforementioned model-based approaches encode this mapping in terms of conventional binary models, either by a large set of such models in the original label space $\mathcal{L}$ (RPC), or by a single binary model in an expanded, high-dimensional space (CC, LL). Since for instance-based methods, there is no need to represent an $\mathbb{X} \to \Omega$ mapping explicitly, such methods can operate on the original target space Ω directly.

The paper is organized as follows: In Section 2, we introduce the problem of label ranking in a more formal way. The core idea of our instance-based approach to label ranking, namely maximum likelihood estimation based on a special probability model for rankings, is discussed in Section 4. The model itself is introduced beforehand in Section 3. Section 5 is devoted to experimental results. The paper ends with concluding remarks in Section 6.

2 Label Ranking

Label ranking can be seen as an extension of the conventional setting of classification. Roughly speaking, the former is obtained from the latter through replacing single class labels by complete label rankings. So, instead of associating every instance $\boldsymbol{x}$ from an instance space $\mathbb{X}$ with one among a finite set of class labels $\mathcal{L} = \{\lambda_1 \dots \lambda_n\}$, we now associate $\boldsymbol{x}$ with a total order of the class labels, that is, a complete, transitive, and asymmetric relation $\succ_{\boldsymbol{x}}$ on $\mathcal{L}$ where $\lambda_i \succ_{\boldsymbol{x}} \lambda_j$ indicates that λ_i precedes λ_j in the ranking associated with $\boldsymbol{x}$. It follows that a ranking can be considered as a special type of preference relation, and therefore we shall also say that $\lambda_i \succ_{\boldsymbol{x}} \lambda_j$ indicates that λ_i is *preferred* to λ_j given the instance $\boldsymbol{x}$. To illustrate, suppose that instances are students (characterized by attributes such as sex, age, and major subjects in secondary school) and $\succ$ is a preference relation on a fixed set of study fields such as Math, CS, Physics.

Formally, a ranking $\succ_{\boldsymbol{x}}$ can be identified with a permutation $\pi_{\boldsymbol{x}}$ of the set $\{1 \dots n\}$. It is convenient to define $\pi_{\boldsymbol{x}}$ such that $\pi_{\boldsymbol{x}}(i) = \pi_{\boldsymbol{x}}(\lambda_i)$ is the position of λ_i in the ranking. This permutation encodes the (ground truth) ranking:

$$\lambda_{\pi_{\boldsymbol{x}}^{-1}(1)} \succ_{\boldsymbol{x}} \lambda_{\pi_{\boldsymbol{x}}^{-1}(2)} \succ_{\boldsymbol{x}} \dots \succ_{\boldsymbol{x}} \lambda_{\pi_{\boldsymbol{x}}^{-1}(n)} \ ,$$

where $\pi_{\boldsymbol{x}}^{-1}(j)$ is the index of the label at position j in the ranking. The class of permutations of $\{1 \ldots n\}$ (the symmetric group of order n) is denoted by Ω. 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 that there exists a deterministic $\mathbb{X} \to \Omega$ mapping. Instead, every instance is associated with a *probability distribution* over Ω. This means that, for each $\boldsymbol{x} \in \mathbb{X}$, there exists a probability distribution $\Pr(\cdot \,|\, \boldsymbol{x})$ such that, for every $\pi \in \Omega$,

$$\Pr(\pi \,|\, \boldsymbol{x}) \tag{1}$$

is the probability that $\pi_{\boldsymbol{x}} = \pi$.

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

$$\lambda_{\pi_{\boldsymbol{x}}^{-1}(i_1)} \succ_{\boldsymbol{x}} \lambda_{\pi_{\boldsymbol{x}}^{-1}(i_2)} \succ_{\boldsymbol{x}} \ldots \succ_{\boldsymbol{x}} \lambda_{\pi_{\boldsymbol{x}}^{-1}(i_k)} \ ,$$

where $\{i_1, i_2 \ldots i_k\}$ is a subset of the index set $\{1 \ldots n\}$ such that $1 \leq i_1 < i_2 < \ldots < i_k \leq n$. For example, for an instance $\boldsymbol{x}$, it might be known that $\lambda_2 \succ_{\boldsymbol{x}} \lambda_1 \succ_{\boldsymbol{x}} \lambda_5$, while no preference information is given about the labels λ_3 or λ_4.

To evaluate the predictive performance of a label ranker, a suitable loss function on Ω is needed. In the statistical literature, several distance measures for rankings have been proposed. One commonly used measure is the number of discordant pairs,

$$D(\pi, \sigma) \ = \ \{ \, (i,j) \,|\, i < j, \, \pi(i) > \pi(j) \text{ and } \sigma(i) < \sigma(j) \, \} \ , \tag{2}$$

which is closely related to the Kendall's tau coefficient. In fact, the latter is a normalization of (2) to the interval $[-1, 1]$ that can be interpreted as a correlation measure (it assumes the value 1 if $\sigma = \pi$ and the value -1 if σ is the reversal of π). Kendall's tau is a natural, intuitive, and easily interpretable measure [5]. We shall focus on (2) throughout the paper, even though other distance measures could of course be used. A desirable property of any distance $D(\cdot)$ is its invariance toward a renumbering of the elements (renaming of labels). This property is equivalent to the *right invariance* of $D(\cdot)$, namely $D(\sigma\nu, \pi\nu) = D(\sigma, \pi)$ for all $\sigma, \pi, \nu \in \Omega$, where $\sigma\nu = \sigma \circ \nu$ denotes the permutation $i \mapsto \sigma(\nu(i))$. The distance (2) is right-invariant, and so are most other commonly used metrics on Ω.

3 The Mallows Model

So far, we did not make any assumptions about the probability measure (1) despite its existence. To become more concrete, we resort to a distance-based

probability model introduced by Mallows [5]. The standard Mallows model is a two-parameter model that belongs to the exponential family:

$$\Pr(\sigma \,|\, \theta, \pi) = \frac{\exp(\theta D(\pi, \sigma))}{\phi(\theta, \pi)}, \tag{3}$$

where the two parameters are the location parameter (modal ranking, center ranking) $\pi \in \Omega$ and the spread parameter $\theta \leq 0$. For right-invariant metrics, it can be shown that the normalization constant does not depend on π and, therefore, can be written as a function $\phi(\theta)$ of θ alone. This is due to

$$\begin{aligned}\phi(\theta, \pi) = \sum_{\sigma \in \Omega} \exp(\theta D(\sigma, \pi)) &= \sum_{\sigma \in \Omega} \exp(\theta D(\sigma \pi^{-1}, e)) \\ &= \sum_{\sigma' \in \Omega} \exp(\theta D(\sigma', e)) = \phi(\theta) \ ,\end{aligned}$$

where $e = (1 \ldots n)$ is the identity ranking. More specifically, it can be shown that the normalization constant is given by [6]

$$\phi(\theta) = \prod_{j=1}^{n} \frac{1 - \exp(j\theta)}{1 - \exp(\theta)}, \tag{4}$$

and that the expected distance from the center is

$$\mathbb{E}\left[D(\sigma, \pi) \,|\, \theta, \pi\right] = \frac{n \exp(\theta)}{1 - \exp(\theta)} - \sum_{j=1}^{n} \frac{j \exp(j\theta)}{1 - \exp(j\theta)} \ . \tag{5}$$

Obviously, the Mallows model assigns the maximum probability to the center ranking π. The larger the distance $D(\sigma, \pi)$, the smaller the probability of σ becomes. The spread parameter θ determines how quickly the probability decreases, i.e., how peaked the distribution is around π. 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 π and 0 to all other rankings.

4 Learning and Inference

Coming back to the label ranking problem and the idea of instance-based learning, 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. Moreover, let $\sigma_1 \ldots \sigma_k \in \Omega$ denote the rankings associated, respectively, with $\boldsymbol{x}_1 \ldots \boldsymbol{x}_k$.

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 $\Pr(\cdot \,|\, \boldsymbol{x})$ on Ω is (at least approximately) *locally constant* around the query $\boldsymbol{x}$. By furthermore assuming

independence of the observations, the probability to observe $\boldsymbol{\sigma} = \{\sigma_1 \dots \sigma_k\}$ given the parameters (θ, π) becomes

$$\begin{aligned} \Pr(\boldsymbol{\sigma} \,|\, \theta, \pi) &= \prod_{i=1}^{k} \Pr(\sigma_i \,|\, \theta, \pi) = \prod_{i=1}^{k} \frac{\exp\left(\theta D(\sigma_i, \pi)\right)}{\phi(\theta)} \\ &= \frac{\exp\left(\theta \sum_{i=1}^{k} D(\sigma_i, \pi)\right)}{\left(\prod_{j=1}^{n} \frac{1-\exp(j\theta)}{1-\exp(\theta)}\right)^k}. \end{aligned} \tag{6}$$

The maximum likelihood estimation (MLE) of (θ, π) is then given by those parameters that maximize this probability. It is easily verified that the MLE of π is given by

$$\hat{\pi} \;=\; \arg\min_{\pi} \sum_{i=1}^{k} D(\sigma_i, \pi), \tag{7}$$

i.e., by the (generalized) median of the rankings $\sigma_1 \dots \sigma_k$. Moreover, the MLE of θ is derived from the average observed distance from $\hat{\pi}$, which is an estimation of the expected distance $\mathbb{E}\left[D(\sigma, \pi)|\theta, \pi\right]$:

$$\frac{1}{k} \sum_{i=1}^{k} D(\sigma_i, \hat{\pi}) = \frac{n \exp(\theta)}{1 - \exp(\theta)} - \sum_{j=1}^{n} \frac{j \exp(j\theta)}{1 - \exp(j\theta)}. \tag{8}$$

Since the right-hand side of (8) is monotone increasing, a standard line search quickly converges to the MLE [6].

Now, consider the more general case of incomplete preference information, which means that a ranking σ_i does not necessarily contain all labels. The probability of σ_i is then given by

$$\Pr(E(\sigma_i)) \;=\; \sum_{\sigma \in E(\sigma_i)} \Pr(\sigma \,|\, \theta, \pi) \;\;,$$

where $E(\sigma_i)$ denotes the set of all *consistent extensions* of σ_i: A permutation $\sigma \in \Omega$ is a consistent extension of σ if it ranks all labels that also occur in σ_i in the same order.

The probability of observing the neighbor rankings $\boldsymbol{\sigma} = (\sigma_1 \dots \sigma_k)$ then becomes

$$\begin{aligned} \Pr(\boldsymbol{\sigma} \,|\, \theta, \pi) &= \prod_{i=1}^{k} \Pr(E(\sigma_i) \,|\, \theta, \pi) = \prod_{i=1}^{k} \sum_{\sigma \in E(\sigma_i)} \Pr(\sigma \,|\, \theta, \pi) \\ &= \frac{\prod_{i=1}^{k} \sum_{\sigma \in E(\sigma_i)} \exp\left(\theta D(\sigma, \pi)\right)}{\left(\prod_{j=1}^{n} \frac{1-\exp(j\theta)}{1-\exp(\theta)}\right)^k} \;\;. \end{aligned} \tag{9}$$

Computing the MLE of (θ, π) by maximizing this probability now becomes more difficult. For label sets of small to moderate size, say up to 7, one can afford a

simple brute force approach, namely an exhaustive search over Ω to find the center ranking π, combined with a numerical procedure to optimize the spread θ. For larger label sets, this procedure becomes too inefficient. Here, we propose an approximation algorithm that can be seen as an instance of the EM (Expectation-Maximization) family of algorithms.

The algorithm works as follows. Starting from an initial (complete) center ranking $\hat{\pi}$, each incomplete neighbor ranking σ_i is replaced by the most probable consistent extension given $\hat{\pi}$. Regardless of θ, this extension is obviously given by a ranking in $\arg\min_{\sigma \in E(\sigma_i)} D(\sigma, \hat{\pi})$. It can be found by (minimally) re-ranking the center $\hat{\pi}$ so as to make it consistent with the incomplete ranking σ_i. Having replaced all neighbor rankings by their most probable extensions, an MLE (θ, π) can be derived as described for the case of complete information above. The center ranking $\hat{\pi}$ is then replaced by π, and the whole procedure is iterated until the center does not change any more. In the following, we discuss two subproblems of the algorithm in more detail, namely the solution of the median problem (7), which needs to be solved to find an MLE π, and the choice of an initial center ranking.

Solving the (generalized) median problem (7) is known to be NP-complete for Kendall's tau, i.e., if the distance D is given by the number of rank inversions [7]. To solve this problem approximately, we make use of the fact that Kendall's tau is well approximated by Spearman's rank correlation [8], and that the median can be computed for this measure (i.e., for D given by the sum of squared rank differences) by a procedure called *Borda count* [9]: Given a (complete) ranking σ_i of n labels, the top-label receives n votes, the second-ranked $n-1$ votes, and so on. Given k rankings $\sigma_1 \ldots \sigma_k$, the sum of the k votes are computed for each label, and the labels are then ranked according to their total votes.

The choice of the initial center ranking in the above algorithm is of course critical. To find a good initialization, we again resort to the idea of solving the problem (7) approximately using the Borda count principle. At the beginning, however, the neighbor rankings σ_k are still incomplete. To handle this situation, we make the simplifying assumption that the completions are uniformly distributed in $E(\sigma_i)$. Again, this is an approximation, since we actually proceed from the Mallows and not from the uniform model. On the basis of this assumption, we can show the following result (proof omitted due to space restrictions).

Theorem 1. *Let a set of incomplete rankings $\sigma_1 \ldots \sigma_k$ be given, and suppose the associated complete rankings $S_1 \ldots S_k$ to be distributed, respectively, uniformly in $E(\sigma_1) \ldots E(\sigma_k)$. The expected sum of distances $D(\pi, S_1) + \ldots + D(\pi, S_k)$, with D the sum of squared rank distances, becomes minimal for the ranking π which is obtained by a generalized Borda count, namely a Borda count with a generalized distribution of votes from incomplete rankings: If σ_i is an incomplete ranking of $m \leq n$ labels, then the label on rank $i \in \{1 \ldots m\}$ receives $(m-i+1)(n+1)/(m+1)$ votes, while each missing label receives a vote of $(n+1)/2$.*

Table 1. Statistics for the semi-synthetic and real datasets

dataset	#examples	#classes	#features
iris	150	3	4
wine	178	3	13
glass	214	6	9
vehicle	846	4	18
dtt	2465	4	24
cold	2465	4	24

5 Experimental Results

5.1 Methods

In this section, we compare our instance-based (nearest neighbor, NN) approach with existing methods for label ranking, namely ranking by pairwise comparison (RPC), constraint classification (CC), and log-linear models for label ranking (LL). Since space restrictions prevent from a detailed review, we refer to the original literature and [1] for a short review of these methods. Regarding the concrete implementation and parameterization of these methods, we also follow [1].

To fit the Mallows model, we test the two previously discussed variants, namely the exhaustive search which guarantees an optimal solution (NNE) and the approximation algorithm outlined in Section 4 (NNH). The parameter k (neighborhood size) was selected through cross validation on the training set. As a distance measure on the instance space we used the Euclidean distance (after normalizing the attributes).

5.2 Data

We used two real-world data sets, dtt and cold, from the bioinformatics field. These data sets contain two types of genetic data, namely phylogenetic profiles and DNA microarray expression data for the Yeast genome.[1] The genome consists of 2465 genes, and each gene is represented by an associated phylogenetic profile of length 24. Using these profiles as input features, we investigated the task of predicting a "qualitative" representation of an expression profile; see [1] for a detailed description and motivation of this task.

In addition to the real-world data sets, the following multiclass datasets from the UCI repository of machine learning databases and the Statlog collection were included in the experimental evaluation: iris, wine, glass, vehicle. For each of these datasets, a corresponding ranking dataset was generated in the following manner: We trained a naive Bayes classifier on the respective dataset. Then, for each example, *all* the labels present in the dataset were ordered with respect to decreasing predicted class probabilities (in the case of ties, labels with lower

[1] This data is publicly available at http://www1.cs.columbia.edu/compbio/ .

Table 2. Experimental results in terms of Kendall's tau (mean and standard deviation) for different missing label rates (parameter p).

iris	*0%*	*10%*	*20%*	*30%*	*40%*	*50%*	*60%*	*70%*
RPC	.885±.068	.888±.064	.886±.060	.871±.074	.854±.082	.837±.089	.779±.110	.674±.139
CC	.836±.089	.825±.095	.815±.088	.807±.099	.788±.105	.766±.115	.743±.131	.708±.105
LL	.818±.088	.811±.089	.805±.087	.806±.087	.800±.091	.788±.087	.778±.096	.739±.186
NNE	.960±.036	**.956±.041**	**.941±.044**	**.934±.049**	**.915±.056**	**.882±.085**	**.859±.082**	**.812±.107**
NNH	**.966±.034**	.948±.036	.917±.051	.863±.072	.822±.088	.802±.084	.767±.122	.733±.104
wine								
RPC	.921±.053	.900±.067	.886±.073	.902±.063	.910±.065	.882±.082	.864±.097	.822±.118
CC	.933±.043	.918±.057	.929±.058	.911±.059	.922±.057	.885±.074	.853±.078	.802±.123
LL	.942±.043	.944±.046	.939±.051	**.944±.042**	.933±.062	.918±.065	.906±.072	**.864±.094**
NNE	.952±.048	.945±.051	.943±.055	.940±.054	**.941±.050**	**.930±.058**	**.910±.061**	.677±.173
NNH	**.953±.042**	**.949±.041**	**.949±.041**	.933±.048	.899±.075	.709±.186	.591±.210	.587±.180
glass								
RPC	**.882±.042**	**.875±.046**	**.867±.044**	**.851±.052**	**.840±.053**	**.813±.062**	.799±.054	.754±.076
CC	.846±.045	.848±.053	.838±.059	.835±.054	.833±.051	.807±.066	.789±.052	.747±.061
LL	.817±.060	.815±.061	.813±.063	.819±.062	.819±.060	.809±.066	**.806±.065**	**.807±.063**
NNE	.875±.063	.866±.059	.840±.059	.803±.062	.750±.071	.677±.066	.598±.082	.500±.078
NNH	.865±.059	.847±.062	.810±.056	.754±.069	.691±.063	.633±.061	.550±.069	.484±.079
vehicle								
RPC	.854±.025	.848±.025	.847±.024	.834±.026	.823±.032	.803±.033	.786±.036	.752±.041
CC	.855±.022	.848±.026	**.849±.026**	**.839±.025**	**.834±.026**	**.827±.026**	**.810±.026**	**.791±.030**
LL	.770±.037	.769±.035	.769±.033	.766±.040	.770±.038	.764±.031	.757±.038	.756±.036
NNE	**.863±.030**	**.859±.031**	.847±.029	.834±.031	.822±.030	.795±.033	.766±.034	.723±.036
NNH	.862±.025	.852±.024	.845±.030	.828±.029	.798±.031	.776±.033	.748±.032	.701±.047
dtt								
RPC	.174±.034	.172±.034	.168±.036	.166±.036	.164±.034	.153±.035	.144±.028	.125±.030
CC	.180±.037	.178±.034	.176±.033	**.172±.032**	.165±.033	.158±.033	.149±.031	.136±.033
LL	.167±.034	.168±.033	.168±.034	.168±.034	**.167±.033**	**.167±.036**	**.162±.032**	**.156±.034**
NNE	.182±.036	.179±.036	.173±.036	.169±.036	.162±.036	.161±.037	.154±.036	.136±.035
NNH	**.191±.034**	**.183±.037**	**.176±.036**	.168±.038	.163±.034	.146±.036	.145±.033	.128±.035
cold								
RPC	.221±.028	.217±.028	.213±.030	.212±.030	.208±.030	.201±.030	.188±.030	.174±.031
CC	.220±.029	.219±.030	.212±.030	.212±.028	.205±.024	.197±.030	.185±.031	.162±.035
LL	.209±.028	.210±.031	.206±.030	.210±.030	.203±.031	.203±.031	**.202±.032**	**.192±.031**
NNE	.230±.028	.226±.029	.220±.030	.213±.031	.199±.029	.195±.033	.190±.035	.188±.035
NNH	**.244±.026**	**.237±.028**	**.235±.031**	**.226±.024**	**.220±.029**	**.214±.029**	.199±.030	.192±.032

index are ranked first). Thus, by substituting the single labels contained in the original multiclass datasets with the complete rankings, we obtain the label ranking datasets required for our experiments. A summary of the data sets and their properties is given in Table 1.

5.3 Experiments and Results

Results were derived in terms of the Kendall's tau correlation coefficient from five repetitions of a ten-fold cross-validation. To model incomplete preferences, we modified the training data as follows: A biased coin was flipped for every label in a ranking in order to decide whether to keep or delete that label; the probability for a deletion is specified by a parameter p.

The results are summarized in Table 2. As can be seen, NN is quite competitive to the model-based approaches and often outperforms these methods. In any case, it is always close to the best result. It is also remarkable that NN seems to be quite robust toward missing preferences and compares comparably well in this regard. This was not necessarily expected, since NN uses only local

information, in contrast to the other approaches that induce global models. Our approximation algorithm NNH gives very good approximations of NNE throughout and is especially appealing for large label sets: It dramatically reduces the runtime (not shown due to space restrictions) without any significant decrease of the performance.

A nice feature of our approach, not shared by the model-based methods, is that it comes with a natural measure of the reliability of a prediction. In fact, the smaller the parameter θ, the more peaked the distribution around the center ranking and, therefore, the more reliable this ranking becomes as a prediction. To test whether (the estimation of) θ is indeed a good measure of uncertainty of a prediction, we used it to compute a kind of *accuracy-rejection* curve: By averaging over five 10-fold cross validations (with NNE), we computed an accuracy degree $\tau_{\boldsymbol{x}}$ (the average Kendall's tau) and a reliability degree $\theta_{\boldsymbol{x}}$ for each instance $\boldsymbol{x}$. The instances are then sorted in decreasing order of reliability. Our curve plots a value p against the mean τ-value of the first p percent of the instances. Given that θ is indeed a good indicator of reliability, this curve should be decreasing, because the higher p, the more instances with a less strong θ-value are taken into consideration. As can be seen in Fig. 1, the curves obtained for our data sets are indeed decreasing and thus provide evidence for our claim that θ may serve as a reasonable indicator of the reliability of a prediction.

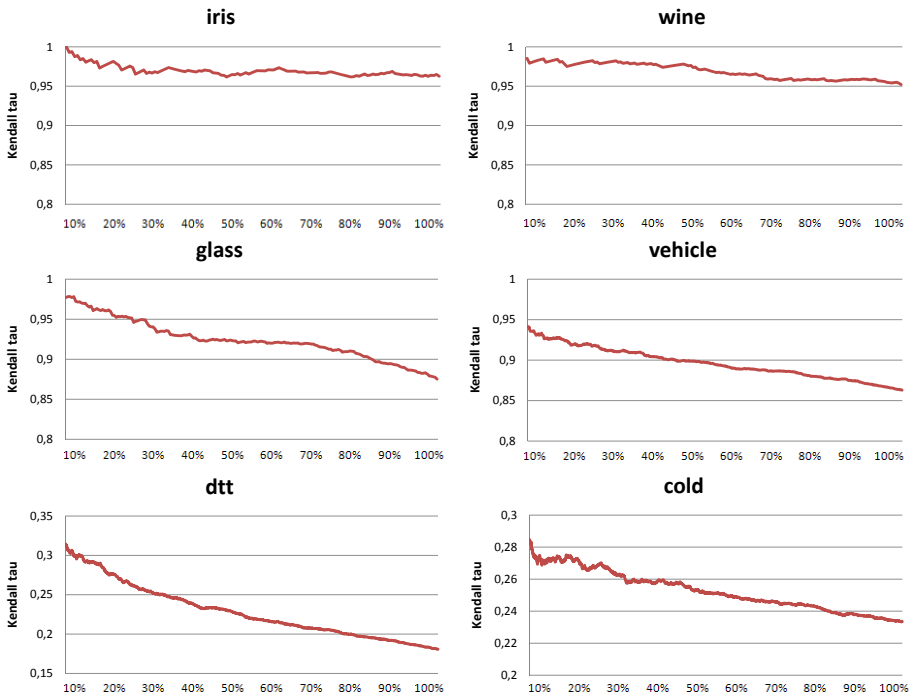

Fig. 1. Accuracy-rejection curves computed on the basis of the parameter θ.

6 Conclusions and Future Work

In this paper, we have introduced an instance-based (nearest neighbor) approach to the label ranking problem that has recently attracted attention in the field of machine learning. Our basic inference principle is a consistent extension of the nearest neighbor estimation principle, as used previously for well-known learning problems such as classification and regression: Assuming that the conditional (probability) distribution of the output given the query is locally constant, we derive a maximum likelihood estimation based on the Mallows model, a special type of probability model for rankings. Our first empirical results are quite promising and suggest that this approach is fully competitive, in terms of predictive accuracy, to (model-based) state-of-the-art methods for label ranking. Besides, it has some further advantages, as it does not only produce a single ranking as an estimation but instead delivers a probability distribution over all rankings. This distribution can be used, for example, to quantify the reliability of the predicted ranking.

Currently, we are working on extensions and variants of the label ranking problem, such as calibrated label ranking and multi-label classification [10]. In fact, we believe that the approach proposed in this paper can be extended to a solid framework that not only allows for solving the label ranking problem itself but also variants thereof.

References

1. Hüllermeier, E., Fürnkranz, J., Cheng, W., Brinker, K.: Label Ranking by Learning Pairwise Preferences. Artificial Intelligence 172(16-17), 1897–1916 (2008)
2. Har-Peled, S., Roth, D., Zimak, D.: Constraint Classification for Multiclass Classification and Ranking. In Becker, S., Thrun, S., Obermayer, K., eds.: Advances in Neural Information Processing Systems 15, 785–792 (2003)
3. Dekel, O., Manning, C., Singer, Y.: Log-Linear Models for Label Ranking. In Touretzky, D.S., Thrun, S., Saul, L.K., Schölkopf, B., eds.: Advances in Neural Information Processing Systems 16, 497–504 (2004)
4. Aha, D., Kibler, D., Albert, M.: Instance-Based Learning Algorithms. Machine Learning 6(1), 37–66 (1991)
5. Mallows, C.: Non-Null Ranking Models. Biometrika 44(1), 114–130 (1957)
6. Fligner, M., Verducci, J.: Distance Based Ranking Models. Journal of the Royal Statistical Society 48(3), 359–369 (1986)
7. Alon, N.: Ranking Tournaments. SIAM Journal on Discrete Mathematics 20(1), 134–142 (2006)
8. Diaconis, P., Graham, R.: Spearman's Footrule as a Measure of Disarray. Journal of the Royal Statistical Society 39(2), 262–268 (1977)
9. Saari, D.: Chaotic Elections!: A Mathematician Looks at Voting. American Mathematical Society (2001)
10. Brinker, K., Hüllermeier, E.: Case-based Multilabel Ranking. In: Proc. IJCAI–07, 20th International Joint Conference on Artificial Intelligence, 701–707, Hyderabad, India (January 2007)