Label Ranking with Partial Abstention using Ensemble Learning

Weiwei Cheng and Eyke Hüllermeier

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

Abstract. In label ranking, the problem is to learn a mapping from instances to rankings over a finite set of predefined class labels. In this paper, we consider a generalization of this problem, namely label ranking with a reject option. Just like in conventional classification, where a classifier can refuse a presumably unreliable prediction, the idea is to concede a label ranker the possibility to abstain. More specifically, the label ranker is allowed to make a weaker prediction in the form of a partial instead of a total order. Thus, unlike a conventional classifier which either makes a prediction or not, a label ranker can abstain to a certain degree. To realize label ranking with a reject option, we propose a method based on ensemble learning techniques. First empirical results are presented showing great promise for the usefulness of the approach.

1 Introduction

The problem to learn a mapping from instances to rankings over a finite set of predefined labels, called *label ranking*, is a natural extension of conventional classification where, instead of a ranking of all labels, only a single label is requested as a prediction. Since a ranking is a special type of preference relation, label ranking is of particular interest for preference learning, an emerging subfield of machine learning [1].

Existing methods for label ranking are typically extensions of algorithms for binary classification. 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 [2]. Two other approaches, constraint classification and log-linear models for label ranking, seek to learn a (linear) utility function for each individual label [3, 4].

An important extension of conventional classification is classification with a reject option [5–7]: The classifier is allowed to abstain from a prediction for a query instance in case it is not sure enough. An abstention of this kind is an obvious means to avoid unreliable predictions and, of course, does also make sense in the context of label ranking. One may even argue that the idea of a reject option becomes more interesting here: While a conventional classifier has only two choices, namely to predict a class label or to abstain, a label ranker can abstain *to a certain degree.*

More specifically, for each pair of labels a and b, it can in principle decide whether to make a prediction about the order relation between these labels, namely $a \succ b$ or $b \succ a$, or to abstain from this prediction. However, the pairwise predictions should of course be consistent in the sense of being transitive and acyclic. In other words, a label ranker with a (partial) reject option is expected to make a prediction in the form of a *partial order* on the set of class labels. In this paper, we propose a method that enables a label ranker to make predictions of this kind. Roughly speaking, our idea is to train a committee of label rankers, using ensemble learning techniques, and to derive a prediction in the form of a consensus, namely the intersection between the label rankings predicted by the members of the committee.

The rest of this paper is organized as follows. After a short recapitulation of the label ranking problem in the next section, we propose our method for label ranking with a reject option in Section 3. Experimental results are presented in Section 4, and Section 5 concludes the paper with a summary.

2 Label Ranking

As mentioned earlier, 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 $\mathcal{L} = \{\lambda_1, \lambda_2, \ldots, \lambda_m\}$, 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.

The goal in label ranking is to learn a "label ranker" in the form of an $\mathbb{X} \to \Omega$ mapping, where Ω is the set of all rankings (or, equivalently, permutations) of $\mathcal{L}$. As training data, a label ranker uses a set of instances $\boldsymbol{x}_k$, $k = 1, \ldots, n$, together with information about the associated rankings. 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 most general case, each training instance is simply associated with a set of pairwise preferences of the form $\lambda_i \succ_{\boldsymbol{x}_k} \lambda_j$, suggesting that, for instance $\boldsymbol{x}_k$, label λ_i should be ranked higher than λ_j.

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, that is, the number of label pairs $(\lambda_i, \lambda_j) \in \mathcal{L} \times \mathcal{L}$ which are differently ordered in the two rankings. This measure is closely related to the Kendall tau coefficient [8]. In fact, the latter is a normalization of the number of discordant pairs to the interval $[-1, 1]$ that can be interpreted as a correlation measure (it assumes the value 1 if the two rankings coincide and the value -1 if the first is the reversal of the second).

3 Label Ranking with Reject Option

Our idea to realize label ranking with a partial reject option is to train, instead of only a single model, a complete ensemble of label rankers which is considered as a "committee of experts". Given a query instance $\boldsymbol{x} \in \mathbb{X}$, each member of the committee makes a prediction in the form of a label ranking. Thus, given an ensemble of size k, label rankings $\succ_1, \succ_2, \ldots, \succ_k$ are obtained. The idea, then, is to define the overall prediction by the intersection of these rankings, namely by the partial order $\succapprox$ such that

$$(\lambda_i \succapprox \lambda_j) \overset{df}{\iff} \forall\, l \in \{1, 2, \ldots, k\} : \lambda_i \succ_l \lambda_j \ .$$

In other words, the pairwise order relation $\lambda_i \succ \lambda_j$ is predicted if and only if it is supported by all members of the committee. Obviously, $\succapprox$ thus defined is indeed a proper partial order, i.e., it is transitive and acyclic.

The above prediction could be criticized due to its non-tolerance toward incorrect predictions of single label rankers. An obvious way to make it more tolerant is to ask for the agreement of *most* instead of all members of the committee:

$$(\lambda_i \succapprox \lambda_j) \overset{df}{\iff} \frac{\#\{l \in \{1, 2, \ldots, k\} \,|\, \lambda_i \succ_l \lambda_j\}}{k} \geq t \ , \tag{1}$$

where $t \in]0.5, 1]$ is a tolerance threshold. Unfortunately, $\succapprox$ thus defined is no longer a partial order.

In fact, it is easily verified that $\succapprox$ is not necessarily transitive. This problem can be solved, however, by replacing $\succapprox$ with its transitive closure (which can easily be computed, for example, using Warshall's algorithm [9]). As to the property of being acyclic, the following proposition is of interest:

Proposition: Given a set of total orders on a finite set $\mathcal{L}$, denote by P_{ab} the proportion of orders in which a precedes b. Then, for any triple of elements $a, b, c \in \mathcal{L}$, we have $P_{ca} \leq 2 - P_{ab} - P_{bc}$.
Proof: Looking at the order of the labels a, b, c, there are only three cases in which $c \succ a$, namely (1) $c \succ a \succ b$, (2) $c \succ b \succ a$, (3) $b \succ c \succ a$. Since $c \succ b$ occurs in the first two cases, the relative frequency of these cases is upper-bounded by $1 - P_{bc}$. Moreover, since $b \succ a$ in the third case, the relative frequency of this case is upper-bounded by $1 - P_{ab}$. Thus, in total the proportion of rankings in which $c \succ a$ cannot exceed $2 - P_{ab} - P_{bc}$. $\square$

As a corollary of this proposition, we immediately conclude that $P_{ab} \geq 2/3$ and $P_{bc} \geq 2/3$ implies $P_{ca} \leq 2/3$. Thus, by choosing $t > 2/3$ in (1), we can guarantee that $\succapprox$ is acyclic. In conjunction with the replacement of $\succapprox$ in (1) by its transitive closure, we thus guarantee the prediction of a proper partial order relation.

To realize an ensemble of label rankers, we resort to the bootstrap aggregation (bagging) approach. Given a training set D of size n, bagging generates k new training sets D_i of size $n' \leq n$ by sampling from D with replacement. Thus, it is likely that some examples will occur more than once in a subsample D_i. If $n' = n$ (as will be assumed throughout the paper) and n is large, then D_i is expected to contain 63.2% of the examples of D, the rest being duplicates. This kind of sample is known as a bootstrap sample. The k models are then fitted using the above k bootstrap samples.

The complete approach to learning a label ranker with reject option is summarized in Algorithm 1.

Algorithm 1 Label ranking with partial abstention

Require: query $\boldsymbol{x} \in \mathbb{X}$, training data D, integer k, base learner $\mathcal{B}$, threshold t
Ensure: matrix M encoding the estimation for $\boldsymbol{x}$ ($M_{ij} = 1$ means $\lambda_i \succ \lambda_j$)

```
initialize $M$ as zero matrix
generate $k$ bootstrap samples from $D$
get $k$ label rankings for $\boldsymbol{x}$ using $\mathcal{B}$
for each of $k$ label rankings do
  for every pair of labels $\lambda_i, \lambda_j \in \mathcal{L}$ do
    if $\lambda_i \succ \lambda_j$ then
      set $M_{ij} = M_{ij} + 1$
    end if
  end for
end for
for every entry in $M$ do
  if $M_{ij} \geq t \times k$ then
    set $M_{ij} = 1$
  else
    set $M_{ij} = 0$
  end if
end for
update $M$ with Warshall's algorithm to ensure transitivity
```

4 Experiments and Evaluation

4.1 Learning Method and Data Sets

To implement the approach outlined above, we use ranking by pairwise comparison with logistic regression as a base learner [2, 1].

In light of the lack of benchmark data for label ranking, we used five classification data sets from the UCI repository and the Statlog collection and turned them into label ranking data following the procedure proposed in [1]: 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). A summary of the data sets and their properties is given in Table 1.[1]

Table 1. Data sets and their properties.

data set	# inst.	# attr.	# labels
iris	150	4	3
wine	178	13	3
glass	214	9	6
vowel	528	10	11
vehicle	846	18	4

4.2 Evaluation

To verify the effectiveness of our approach, we produce generalized accuracy-rejection curves. In conventional classification, a curve of this kind plots the accuracy on the subset of instances which are still classified as a function of the level of rejection. If a reject option is used by a learner in a sensible way, this curve should be increasing: The more instances are rejected, the better the performance on the remaining ones should become, since these are the supposedly reliable cases.

To measure the performance on the remaining predictions, we generalize the Kendall measure as follows:

$$\mathrm{C}(\succapprox, \succ) = \frac{\#\text{concordant label pairs} - \#\text{discordant label pairs}}{\#\text{concordant label pairs} + \#\text{discordant label pairs}} \tag{2}$$

This measure compares a predicted partial ranking $\succapprox$ of $\mathcal{L} = \{\lambda_1, \ldots, \lambda_m\}$ with a true ranking $\succ$. A pair of labels λ_i, λ_j is concordant if $\lambda_i \succapprox \lambda_j$ and $\lambda_i \succ \lambda_j$ (or $\lambda_j \succapprox \lambda_i$ and $\lambda_j \succ \lambda_i$) and discordant if $\lambda_i \succapprox \lambda_j$ and $\lambda_j \succ \lambda_i$ (or $\lambda_j \succapprox \lambda_i$ and $\lambda_i \succ \lambda_j$). Note that, if $\succapprox$ abstains on the pair λ_i, λ_j, then this pair is neither concordant nor discordant.[2]

4.3 Results and Discussion

Table 2 shows the results for an ensemble of size 10. The level of rejection is measured in terms of the threshold t in (1): The larger this threshold, the less

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

[2] In the extreme case of complete abstention, the nominator and denominator of (2) both become 0; we ignore such instances in the evaluation.

complete the predicted rankings will become. For comparison, we also show the performance of a conventional label ranker, namely a single RPC model produced on the complete training data. This model always predicts complete rankings and, therefore, has the smallest level of rejection.

As can be seen from the table, our approach is indeed effective in the sense that, throughout all data sets and across all levels, the higher the level of rejection, the better the performance becomes.

Table 2. Results of label ranking with reject option (ensemble size 10) in terms of the mean and standard deviation of measure (2).

threshold	iris	wine	glass	vowel	vehicle
original	0.868±0.093	0.884±0.078	0.793±0.070	0.324±0.028	0.809±0.034
0.7	0.919±0.066	0.918±0.079	0.847±0.055	0.436±0.034	0.851±0.032
0.8	0.921±0.064	0.956±0.057	0.869±0.055	0.478±0.039	0.872±0.031
0.9	0.940±0.050	0.971±0.049	0.892±0.054	0.515±0.045	0.896±0.031
1.0	0.950±0.045	0.995±0.019	0.928±0.046	0.563±0.056	0.926±0.027

5 Summary

In this paper, we have proposed a method for label ranking with a (partial) reject option, that is, for the problem to learn a mapping from instances to partial orders over a fixed set of class labels. The general idea of the method is to train an ensemble of label rankers and to predict a pairwise order relation between two labels only if this relation is supported by the large majority of ensemble members. First experimental results suggest that this approach is effective in the sense of eliminating those parts of a total order that are indeed most unreliable.

References

1. Hüllermeier, E., Fürnkranz, J., Cheng, W., Brinker, K.: Label ranking by learning pairwise preferences. Aritificial Intelligence **172** (2008) 1897–1916 Elsevier.
2. Fürnkranz, J., Hüllermeier, E.: Pairwise preference learning and ranking. In: Proc. ECML–03, 13th European Conference on Machine Learning, Cavtat-Dubrovnik, Croatia (September 2003)
3. 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 (NIPS-02). (2003) 785–792
4. Dekel, O., Manning, C., Singer, Y.: Log-linear models for label ranking. In: Advances in Neural Information Processing Systems. (2003)
5. Chow, C.: On optimum recognition error and reject tradeoff. IEEE Transactions on Information Theory **16**(1) (1970) 41–46
6. Herbei, R., Wegkamp, M.H.: Classification with reject option. Canadian Journal of Statistics **34**(4) (2006) 709–721

7. Bartlett, P.L., Wegkamp, M.H.: Classification with a reject option using a hinge loss. Journal of Machine Learning Research **9** (2008) 1823–1840
8. Kendall, M.: Rank Correlation Methods. Hafner Publishing Co., New York (1955)
9. Warshall, S.: A theorem on boolean matrices. Journal of the ACM **9**(1) (1962) 11–12