Probability Estimation for Multi-class Classification Based on Label Ranking

Weiwei Cheng and Eyke Hüllermeier

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

Abstract. We consider the problem of probability estimation in the setting of multi-class classification. While this problem has already been addressed in the literature, we tackle it from a novel perspective. Exploiting the close connection between probability estimation and ranking, our idea is to solve the former on the basis of the latter, taking advantage of recently developed methods for label ranking. More specifically, we argue that the Plackett-Luce ranking model is a very natural choice in this context, especially as it can be seen as a multinomial extension of the Bradley-Terry model. The latter provides the basis of pairwise coupling techniques, which arguably constitute the state-of-the-art in multi-class probability estimation. We explore the relationship between the pairwise and the ranking-based approach to probability estimation, both formally and empirically. Using synthetic and real-world data, we show that our method does not only enjoy nice theoretical properties, but is also competitive in terms of accuracy and efficiency.

1 Introduction

The problem of classification is normally understood as learning a model that maps instances to class labels. While useful for many purposes, there are numerous applications in which the estimation of the probabilities of the different classes is more desirable than just selecting one of them. Application areas of this kind include safety-critical domains such as medical decision making, where probabilities are useful as a measure of the reliability of a classification, or applications like computational advertising, where they allow one to focus on the most promising alternatives. Moreover, given a probability for each class, it is in principle possible to minimize any loss function, that is, to derive (or at least approximate) Bayes-optimal decisions. This is especially useful in cost-sensitive classification, where different types of misclassification may incur different costs [12]. Simply minimizing the standard 0/1 loss will normally not produce desirable results in this setting.

As discussed in more detail in Section 2, the problem of probability estimation is rather challenging and in a sense even more difficult than conventional classification. In the field of machine learning, the problem has been approached from different directions. Specifically interesting in this regard is the idea of exploiting the connection between probability estimation and *ranking*, another type of

P. Flach et al. (Eds.): ECML PKDD 2012, Part II, LNCS 7524, pp. 83–98, 2012.

prediction problem that has attracted increasing attention in recent years. Indeed, ranking is in a sense in-between classification and probability estimation or, stated differently, can be seen as an intermediate step from classification to probability estimation [8]. In particular, the maximization of ranking measures like the AUC requires sorting a given set of alternatives from most probable positive to most probable negative [5]. Thus, although precise probability degrees of all alternatives are not necessarily needed, at least their order relation must be predicted correctly.

For far, most work on the connection between probability estimation and ranking, including AUC maximization, has focused on the *binary* case, distinguishing only between two classes (positive and negative). Essentially, this means that only a single value needs to be estimated for each instance, namely the probability of belonging to the positive class. In this paper, we establish a connection between probability estimation and ranking for the case of *multiple classes.* To this end, we refer to another type of ranking problem, namely *label ranking* [3,4]. While the binary case is intimately connected with *bipartite ranking*, in which the *instances* are ranked themselves, the problem of label ranking consists of ranking the *class labels* given an instance.

The rest of the paper is organized as follows. In the next section, we discuss the problem of multi-class probability estimation and recall the basic ideas of pairwise coupling and classifier calibration. In Section 3, we introduce the problem of label ranking. Then, in Section 4, we establish a tight link between label ranking and probability estimation, taking advantage of a probabilistic ranking model called Plackett-Luce. In Section 5, we show how the label ranking problem can be approached on the basis of this model. Building on the connection established in Section 4 and the PL-based label ranking method introduced in Section 5, we then introduce a method for probability estimation based on label ranking in Section 6. Experimental results are presented in Section 7, before concluding the paper in Section 8.

2 Multi-class Probability Estimation

Consider the standard setting of multi-class classification with an instance space $\mathbb{X}$ and a set of classes $\mathcal{Y} = \{y_1, \ldots, y_K\}$. We are interested in learning a probabilistic classifier, that is, a model that estimates the conditional probabilities of classes given an instance $\boldsymbol{x} \in \mathbb{X}$:

$$(p_1, \ldots, p_K) = (\mathbf{P}_{\mathcal{Y}}(y_1 \,|\, \boldsymbol{x}), \ldots, \mathbf{P}_{\mathcal{Y}}(y_K \,|\, \boldsymbol{x})) \tag{1}$$

Since true probability degrees are rarely available for training, probabilistic classifiers are typically trained on standard classification data, that is, observations of the form $(\boldsymbol{x}, y) \in \mathbb{X} \times \mathcal{Y}$, where the class label y is assumed to be generated according to $\mathbf{P}_{\mathcal{Y}}(\cdot \,|\, \boldsymbol{x})$.

Probability estimation is known to be a quite hard problem, especially in comparison to standard classification. This comes at no surprise, noting that proper probability estimation is a sufficient but not necessary condition for proper classification: If the conditional class probabilities (1) are predicted accurately, an optimal classification can simply be made by picking the class with highest probability:

$$\hat{y} = \operatorname*{arg\,max}_{y_i \in \mathcal{Y}} \hat{\mathbf{P}}(y_i \,|\, \boldsymbol{x}) \tag{2}$$

More generally, the Bayes decision can be taken so as to minimize any loss in expectation. On the other hand, a correct classification can also be obtained based on less accurate probability estimates. In fact, the classification will remain correct as long as the estimated probability is highest for the true class. Or, stated differently, an estimation error will remain ineffective unless it changes the result of the arg max operation in (2). This is also the reason for why methods like naive Bayes show competitive performance in classification despite producing relatively inaccurate probability estimates [7].

Methods like naive Bayes and decision trees are multi-class classifiers and can in principle be used to produce probability estimates in this setting. In practice, however, one often prefers to estimate probabilities in the two-class setting, especially because estimating a single probability (of the positive class) is much simpler than estimating $K - 1$ probabilities simultaneously. Moreover, the binary case is amenable to a broader spectrum of classifiers, including logistic regression, which is a proven method for probability estimation. On the other hand, the reduction of multinomial to binomial probability estimation obviously involves an aggregation problem, namely the need to combine probabilities on pairs of classes into probabilities on the label set $\mathcal{Y}$. This is the idea of “pairwise coupling” techniques.

2.1 Pairwise Coupling

As a special type of binary decomposition technique, pairwise coupling allows one to tackle multi-class problems with binary classifiers. The key idea is to transform a K-class problem into $K(K-1)/2$ binary problems, one for each pair of classes. More specifically, a separate model $M_{i,j}$ is trained for each pair of labels $(y_i, y_j) \in \mathcal{Y} \times \mathcal{Y}$, $1 \leq i < j \leq K$, using the examples from these two classes as their training set; thus, a total number of $K(K-1)/2$ models is needed. $M_{i,j}$ is intended to separate the objects with label y_i from those having label y_j.

At prediction time, a query instance $\boldsymbol{x} \in \mathbb{X}$ is submitted to all models $M_{i,j}$. The predictions $p_{i,j} = M_{i,j}(\boldsymbol{x})$ are typically interpreted by means of the Bradley-Terry model [1], a probabilistic choice model expressing the probability that “y_i wins against y_j” as follows:

$$p_{i,j} = \mathbf{P}(y_i \succ y_j) = \mathbf{P}_{\mathcal{Y}}(y_i \,|\, \{y_i, y_j\}) = \frac{p_i}{p_i + p_j} \tag{3}$$

Based on the relationship (3), the unconditional probabilities p_i can be derived from the conditional (pairwise) probabilities $p_{i,j}$. Obviously, however, it will not always be possible to find a distribution $(p_1, \ldots, p_K)$ such that the equality

$p_{i,j} = p_i/(p_i + p_j)$ holds for all $1 \leq i < j \leq K$, simply because this system of equations is over-constrained: K variables have to satisfy $K(K-1)/2$ equations (plus the constraint $p_1 + \ldots + p_K = 1$). In fact, one should notice that the models $M_{i,j}$ are learnt independently of each other, so that the predictions $p_{i,j}$ are not necessarily coherent.

Pairwise coupling techniques therefore seek to solve the above reconstruction problem approximately. Different methods for putting this idea into practice have been proposed and compared in [21]. For example, the following system of linear equations can be derived by "averaging" over the identities $\mathbf{P}_{\mathcal{Y}}(y_i) = \mathbf{P}_{\mathcal{Y}}(y_i \,|\, \{y_i, y_j\}) \cdot \mathbf{P}_{\mathcal{Y}}(\{y_i, y_j\})$:

$$\begin{aligned}\mathbf{P}_{\mathcal{Y}}(y_i) &= \frac{1}{K-1} \sum_{j \neq i} \mathbf{P}_{\mathcal{Y}}(y_i \,|\, \{y_i, y_j\}) \cdot \mathbf{P}_{\mathcal{Y}}(\{y_i, y_j\}) \\ &= \frac{1}{K-1} \sum_{j \neq i} \mathbf{P}_{\mathcal{Y}}(y_i \,|\, \{y_i, y_j\}) \cdot (\mathbf{P}_{\mathcal{Y}}(y_i) + \mathbf{P}_{\mathcal{Y}}(y_j))\end{aligned}$$

Or, in terms of the p_i an $p_{i,j}$:

$$(K-1)p_i \;=\; \sum_{j \neq i} p_{i,j} \cdot (p_i + p_j)$$

In conjunction with the constraint $p_1 + \ldots + p_K = 1$ and the non-negativity of the p_i, this system has a unique solution provided that $p_{i,j} > 0$ for all $1 \leq i, j \leq K$. Among the methods compared in [21], this approach turned out to perform specifically well.

2.2 Classifier Calibration

As mentioned earlier, the scores produced by conventional classification methods are typically quite biased: Although they might be good enough for correct classification, they do not provide accurate probability estimates. This is true even for methods with an inherently probabilistic interpretation, such as logistic regression [24]. Among a number of possible reasons, let us mention that all such methods are based on rather strong model assumptions that will commonly be violated in practice. In naive Bayes, for example, this is the assumption of conditional independence of the attributes given the class. Likewise, logistic regression assumes that the log of the odds ratio is a linear function of the attributes. Another reason is the fact that for commonly used loss functions such as 0/1 loss or hinge loss, the true probability is *not* a risk minimizer [2].

To overcome this problem, several methods for "classifier calibration" have been proposed in the literature [18,22]. These are post-processing methods whose idea is to find a mapping that turns classifier scores into meaningful probability estimates. As an example, we mention the method of isotonic regression [23], which, due to its nature as a nonparametric approach, is less susceptible to the aforesaid problem of model misspecification. Besides, it has been shown to perform quite well in practice [16].

Isotonic regression finds a monotone mapping from scores to probabilities or, say, from poorly estimated probabilities to hopefully better ones. Monotonicity assures that the order of classes can never be reversed: If class y_i receives a higher score by the classifier than y_j, then the calibrated probability estimate for the former cannot be smaller than the estimate for the latter. Against the background of our discussion about the relationship between ranking and probability estimation, this is clearly a desirable property.

3 Label Ranking

In the setting of label ranking, each instance $\boldsymbol{x}$ from the 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 $\boldsymbol{x}$.

Formally, a total order $\succ_{\boldsymbol{x}}$ can be identified with a permutation $\pi_{\boldsymbol{x}}$ of the set $[K] = \{1, \ldots, K\}$. 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

$$y_{\pi_{\boldsymbol{x}}(1)} \succ_{\boldsymbol{x}} y_{\pi_{\boldsymbol{x}}(2)} \succ_{\boldsymbol{x}} \ldots \succ_{\boldsymbol{x}} y_{\pi_{\boldsymbol{x}}(K)} \enspace .$$

The class of permutations of $[K]$ (the symmetric group of order K) 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 the existence of a deterministic $\mathbb{X} \longrightarrow \Omega$ 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 $\mathbf{P}_\Omega(\cdot \,|\, \boldsymbol{x})$ such that, for every $\pi \in \Omega$, $\mathbf{P}_\Omega(\pi \,|\, \boldsymbol{x})$ is the probability that $\pi_{\boldsymbol{x}} = \pi$.

The goal in label ranking is to learn a "label ranker" in the form of an $\mathbb{X} \longrightarrow \Omega$ mapping. As training data, a label ranker uses a set of instances $\boldsymbol{x}_n$, $n \in [N]$, 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

$$y_{\pi_{\boldsymbol{x}}(1)} \succ_{\boldsymbol{x}} y_{\pi_{\boldsymbol{x}}(2)} \succ_{\boldsymbol{x}} \ldots \succ_{\boldsymbol{x}} y_{\pi_{\boldsymbol{x}}(k)} \enspace , \tag{4}$$

where $k < K$ and $\{\pi(1), \ldots, \pi(k)\} \subset [K]$. 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 π; thus, the presence of a class y_i is equivalent to $\pi^{-1}(i) > 0$.

Table 1. A distribution of rankings with three labels

π	$\mathbf{P}_\Omega(\pi \,\|\, \boldsymbol{x})$
$y_1 \succ y_2 \succ y_3$	0.10
$y_1 \succ y_3 \succ y_2$	0.25
$y_2 \succ y_1 \succ y_3$	0.20
$y_2 \succ y_3 \succ y_1$	0.20
$y_3 \succ y_1 \succ y_2$	0.25
$y_3 \succ y_2 \succ y_1$	0

4 Label Ranking vs Classification: A Probabilistic Link

In contrast to conventional classification, the setting of label ranking does not assume the existence of a "true class label" of an instance. In fact, while the output space in classification is given by the set $\mathcal{Y}$ of class labels, and a probability vector of conditional class probabilities (1) can be associated with every instance $\boldsymbol{x} \in \mathcal{X}$, the output space in label ranking is the class of permutations Ω. Yet, as will be explained in the following, label ranking can be interpreted as a generalization of conventional classification or, the other way around, classification can be seen as a special case of label ranking. Most naturally, this connection is obtained by associating the "true class" in classification with the top-ranked label in label ranking. For the ease of exposition, we shall subsequently drop the conditioning on the instance $\boldsymbol{x}$.

4.1 From Probabilities on Rankings to Class Probabilities

Formally, the connection between label ranking and classification is established by means of a mapping between the spaces $\mathfrak{P}(\mathcal{Y})$ and $\mathfrak{P}(\Omega)$, that is, the space of probability distributions on $\mathcal{Y}$ and the space of probability distributions on Ω. Associating the observed class in classification with the top-ranked label in label ranking then comes down to mapping a measure $\mathbf{P}_\Omega \in \mathfrak{P}(\Omega)$ to a measure $\mathbf{P}_\mathcal{Y} \in \mathfrak{P}(\mathcal{Y})$ such that

$$p_j = \mathbf{P}_\mathcal{Y}(y_j) = \sum_{\pi \in \Omega \colon \pi(1) = j} \mathbf{P}_\Omega(\pi) \; . \tag{5}$$

For example, the probability distribution $\mathbf{P}_\Omega$ in Table 1 is mapped to the distribution $\mathbf{P}_\mathcal{Y} = (p_1, p_2, p_3) = (0.35, 0.4, 0.25)$. Note that the most probable class (y_2) differs from the top-label in the most probable ranking (y_1).

The other way around, there are several ways of embedding $\mathfrak{P}(\mathcal{Y})$ in $\mathfrak{P}(\Omega)$ (indeed, note that $|\Omega|$ is in general much larger than $|\mathcal{Y}|$); we will come back to this issue when discussing the so-called Plackett-Luce model below.

4.2 The Plackett-Luce Model

So far, no specific assumptions about the probability measure $\mathbf{P}_\Omega$ on Ω were made. Needless to say, due to the large cardinality of the space Ω, it is practically impossible to work with the full class of distributions $\mathfrak{P}(\Omega)$. For that reason, different types of *parametrized* classes of probability distributions on rankings have been proposed in statistics [15].

A prominent example is the Mallows model [14], a *distance-based* probability model belonging to the family of exponential distributions. The standard Mallows model is determined by two parameters:

$$\mathbf{P}_{\Omega}(\pi \,|\, \theta, \pi_0) = \frac{\exp(-\theta D(\pi, \pi_0))}{\phi(\theta)} \tag{6}$$

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 π_0).

In the following, we shall focus on another model that was first studied by Luce [13] and subsequently by Plackett [17]. The Plackett-Luce (PL) model appears to be especially appealing for our purpose, as it establishes a natural bridge between label ranking and classification. The PL model is specified by a parameter vector $\boldsymbol{v} = (v_1, v_2, \ldots, v_K) \in \mathbb{R}_+^K$:

$$\mathbf{P}_{\Omega}(\pi \,|\, \boldsymbol{v}) \;=\; \prod_{i=1}^{K} \frac{v_{\pi(i)}}{v_{\pi(i)} + v_{\pi(i+1)} + \ldots + v_{\pi(K)}} \tag{7}$$

Obviously, this model can be seen as a generalization of the above-mentioned Bradley-Terry model (3) for the pairwise comparison of alternatives. Indeed, a natural interpretation of the PL model is a *stage-wise* construction of a ranking: A ranking is produced by a sequence of *choices*, where each choice problem consists of selecting one among the labels that have not been picked so far, and the probability of a label y_i being selected is always proportional to its “skill” parameter v_i. First, the top label is chosen, and the probability of each label y_i to be selected is given by $v_i/(v_1+v_2+\ldots+v_K)$. Then, the second label is chosen among those still available, using the same selection principle, and so on and so forth. In other words, with probabilities p_i defined as “normalized skills”

$$p_i = \mathbf{P}_{\mathcal{Y}}(y_i) = \frac{v_i}{v_1 + v_2 + \ldots + v_K} \;\;, \tag{8}$$

the probability of y_i to be chosen among a set $C \subseteq \mathcal{Y}$ of remaining candidates exactly equals the conditional probability $\mathbf{P}_{\mathcal{Y}}(y_i \,|\, C)$. Consequently, the probability (7) can also be written as follows:

$$\mathbf{P}_{\Omega}(\pi \,|\, \boldsymbol{v}) = \mathbf{P}_{\Omega}(\pi \,|\, \boldsymbol{p}) \;=\; \prod_{i=1}^{K} \mathbf{P}_{\mathcal{Y}}\left(y_{\pi(i)} \,|\, C_i\right) \;\;, \tag{9}$$

where $C_i = \{y_{\pi(i)}, \ldots, y_{\pi(K)}\}$ is the set of remaining candidates and $\boldsymbol{p} = \boldsymbol{v}/||\boldsymbol{v}||$ is the probability vector obtained by normalizing the parameter vector $\boldsymbol{v}$.

Thus, with a PL model $\boldsymbol{v} = (v_1, \ldots, v_K)$, one can simultaneously associate a distribution $\mathbf{P}_{\mathcal{Y}}$ on $\mathcal{Y}$ and a distribution $\mathbf{P}_{\Omega}$ on Ω that are closely connected with each other. In particular, this model is coherent with the mapping (5) in the sense that $\mathbf{P}_{\mathcal{Y}}(y_i) = \mathbf{P}_{\Omega}(\pi(1) = i)$. Moreover, the PL model defines a specific though natural embedding of $\mathfrak{P}(\mathcal{Y})$ in $\mathfrak{P}(\Omega)$ via (9). Last but not least, it allows for computing the probability of incomplete rankings (which normally requires an expensive marginalization, i.e., summation over all linear extensions) in a quite convenient way: The probability of an incomplete ranking (4) is given by

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

As an aside, we mention that the appealing properties of the PL model as outlined above are closely connected with the "choice axioms" of Luce [13]. In fact, it is known that the PL model is the only ranking model satisfying these axioms.

5 Label Ranking Based on the PL Model

A label ranking method based on the PL model has been proposed in [3]. The key idea of this method is to define the PL parameters $\boldsymbol{v}$ as a function of the input attributes specifying an instance: $\boldsymbol{v} = (v_1, \ldots, v_K) = f(\boldsymbol{x})$. More specifically, log-linear models are used to guarantee non-negativity, that is, the logarithm of each parameter v_i is modeled as a linear function:

$$v_i \;=\; \exp\left(\langle \boldsymbol{w}^{(i)}, \boldsymbol{x} \rangle\right) \;=\; \exp\left(\sum_{j=1}^{d} w_j^{(i)} \cdot x_j\right) \quad , \tag{10}$$

where an instance is assumed to be represented in terms of a feature vector $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathcal{X} \subseteq \mathbb{R}^d$.

5.1 Parameter Estimation

Learning a label ranking model then comes down to estimating the parameters $w_j^{(i)}$ $(i \in [K], j \in [d])$ in (10). This can be accomplished by means of maximum likelihood estimation. More precisely, given a training data set

$$\mathcal{T} \;=\; \left\{ \left(\boldsymbol{x}^{(q)}, \pi^{(q)} \right) \right\}_{q=1}^{N} \tag{11}$$

with $\boldsymbol{x}^{(q)} = \left(x_1^{(q)}, \ldots, x_d^{(q)} \right)$, the parameters are determined by maximizing the log-likelihood function

$$L = \sum_{q=1}^{N} \sum_{i=1}^{n_q} \left[\log v\left(\pi^{(q)}(i), q\right) - \log \sum_{j=i}^{n_q} v\left(\pi^{(q)}(j), q\right) \right] , \tag{12}$$

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

$$v(i,q) \;=\; \exp\left(\langle \boldsymbol{w}^{(i)}, \boldsymbol{x}^{(q)} \rangle \right) \quad . \tag{13}$$

For algorithmic details, we refer to [3].

5.2 From Label Ranking to Logistic Regression

Interestingly, we can show that the standard multinomial logistic regression approach for classification can be seen as a special case of the PL-based label ranking method introduced above. To this end, consider the case of classification, where a single class label is observed for each training instance. An observation of this kind can be interpreted as a label ranking, of which only the top-position is known. Or, stated differently, the probability of this observation corresponds to the probability of selecting y_i in the first step of the choice process:

$$\mathbf{P}_{\mathcal{Y}}(y_i \,|\, \boldsymbol{x}; \boldsymbol{w}) = \mathbf{P}_{\Omega}(\pi(1) = i \,|\, \boldsymbol{x}; \boldsymbol{w}) = \frac{\exp\left(\langle \boldsymbol{w}^{(i)}, \boldsymbol{x} \rangle\right)}{\sum_{j=1}^{K} \exp\left(\langle \boldsymbol{w}^{(j)}, \boldsymbol{x} \rangle\right)} \tag{14}$$

The log-likelihood function of the data is then given by

$$L = \sum_{q=1}^{N} \sum_{i=1}^{K} t_{qi} \left[\langle \boldsymbol{w}^{(i)}, \boldsymbol{x} \rangle - \log \sum_{j=1}^{K} \exp\left(\langle \boldsymbol{w}^{(j)}, \boldsymbol{x} \rangle \right) \right] , \tag{15}$$

where $\boldsymbol{t}$ is a coding matrix with $t_{qi} = 1$ if the class of the q-th instance is y_i and $t_{qi} = 0$ otherwise. This model exactly corresponds to the standard model of multinomial logistic regression.

6 A Ranking Approach to Probability Estimation

In our discussion so far, we have established a close connection between (label) ranking and classification. In terms of *modeling*, this connection mainly rests on the interpretation of a classification (an observed class label) as a ranking with the top-label observed. This connection is ideally supported by the PL model, notably because the ranking parameters of this model are in direct correspondence with class probabilities; besides, probabilities of incomplete rankings (i.e., rankings of a subset of the labels) are obtained through simple conditioning. As a consequence, the PL model is also consistent with our monotonicity assumption: The higher the class probability, the higher the (expected) position of the corresponding label in the ranking.

In terms of *methods*, we have noticed that label ranking based on the PL model can be seen as an extension of conventional multinomial logistic regression; or, vice versa, logistic regression corresponds to a special case of PL-based label ranking, namely the case where only top-1 rankings (classes) are observed. The obvious advantage of the label ranking framework is an increased flexibility with regard to the exploitation of training information: While standard logistic regression can only learn from observed class labels, label ranking is also able to exploit comparative preference information of more general type. This includes, for example, pairwise comparisons of the kind “for the instance $\boldsymbol{x}$, class label y_i is more probable than y_j”, even if one cannot assure that y_i is the most likely label. This could be useful in many practical situations, for example if the correct class label cannot be determined precisely although some candidate classes can certainly be excluded [6].

More generally, a ranking can be interpreted as a special type of *qualitative probability* on $\mathcal{Y}$ [20]. The order relation $y_i \succ_{\boldsymbol{x}} y_j$ indicates that the conditional probability of y_i given $\boldsymbol{x}$ is higher than the probability of y_j given $\boldsymbol{x}$, though without specifying any concrete numerical values. By learning a PL-based label ranking model, these qualitative probabilities are then turned into quantitative probabilities $p_i \propto v_i(\boldsymbol{x})$. Thus, label ranking can indeed be seen as a natural bridge between classification and probability estimation.

6.1 PELARA: Probability Estimation via Label Ranking

Our method of Probability Estimation via LAbel RAnking (PELARA) can be summarized as follows:

- The method assumes as training information a set of data (11) consisting of instances $\boldsymbol{x} \in \mathbb{X}$ together with label rankings (4) of varying length $k \in [K]$ (including $k = 1$ for the special case of a class observation).
- On this data, a label ranker is trained using the method outlined in Section 5 (and explained in more detail in [3]).
- As a result, we obtain a model M' that assigns a vector of PL parameters to each query instance $\boldsymbol{x}$:

$$M' : \boldsymbol{x} \mapsto \boldsymbol{v} = \boldsymbol{v}(\boldsymbol{x}) \in \mathbb{R}_+^K$$

- To obtain an (uncalibrated) probability estimate, these vectors are normalized, i.e., $\boldsymbol{v}(\boldsymbol{x})$ is turned into

$$M(\boldsymbol{x}) = \boldsymbol{p}(\boldsymbol{x}) = (p_1(\boldsymbol{x}), \ldots, p_K(\boldsymbol{x})) \propto \boldsymbol{v}(\boldsymbol{x}) \tag{16}$$

  such that $||\boldsymbol{p}(\boldsymbol{x})|| = 1$.

The model M thus obtained defines a probability estimator.

6.2 Comparison with Decomposition Schemes

PELARA offers an appealing alternative to conventional methods such as pairwise coupling. Instead of decomposing the problem into a quadratic number of binary problems first, and combining the predictions of the pairwise models afterward, our label ranking method solves the original problem in one go. As a potential advantage, apart from simplicity, let us mention that the scores (probabilities) thus produced should be well-balanced right away, without the need to couple them in an approximate manner.

Indeed, one should notice that a pairwise decomposition will normally come with a loss of information, and the underlying assumptions justifying the reduction are not entirely clear. For example, while in our approach, the observation of class y_i for an instance $\boldsymbol{x}$ is modeled in terms of the probability $v_i/(v_1+\ldots+v_K)$, it is split into $K-1$ binary training examples $y_i \succ y_j$, $j \in [K]\backslash\{i\}$, in the pairwise approach. However, selecting y_i among the set of candidates $\mathcal{Y}$ is obviously not the same as (independently) selecting y_i in the pairwise comparisons between y_i and y_j:

$$\frac{v_i}{v_1+\ldots+v_K} \neq \prod_{j\neq i} \frac{v_i}{v_i+v_j}$$

In the case of a uniform distribution $\boldsymbol{v} \equiv 1$, for instance, the left-hand side is $1/K$ while the right-hand side is $(1/2)^{K-1}$. Similar arguments apply to the decomposition of an observed ranking into pairwise preferences.

The most common alternative to the pairwise (all pairs) decomposition scheme is one-vs-rest (OVR) decomposition [19]: One model is trained for each class label y_i, using this label as positive and all others as negative examples; for probability estimation, the predictions of these models are simply normalized. Thus, OVR trains a smaller number of models. The individual models, however, are typically more complex: Separating a class from all other class simultaneously is normally more difficult than only separating it from each class individually, and consequently may call for more complex decision boundaries. Besides, the individual problems may become quite imbalanced.

Our approach is in a sense in-between pairwise and OVR learning: Like OVR, it trains a linear number of models, one for each label. Yet, since these models are all trained simultaneously, without building negative meta-classes, the aforesaid disadvantage of OVR is avoided.

7 Experiments

This section presents experimental results, starting with a simplified analysis that is meant to help understand some key differences between the pairwise coupling and the ranking-based approach to probability estimation. Next, we compare our method with state-of-the-art approaches to probability estimation on a set of classification benchmarks.

7.1 On the Reconstruction Error of Pairwise Coupling

In comparison to the pairwise approach to probability estimation, which consists of decomposing the original multi-class problem into binary problems first and "coupling" the solutions (probability estimates) of these problems afterward, our ranking-based method allows for solving the original problem in a single step: Since all labels are treated simultaneously, there is no need for any type of aggregation. In principle, this should be seen as an advantage, especially since the decomposition step in pairwise learning is supposed to come along with a loss of information.

More concretely, one may wonder to what extent pairwise coupling is able to reconstruct a probability vector $\boldsymbol{p} = (p_1, \ldots, p_K)$ from its pairwise components

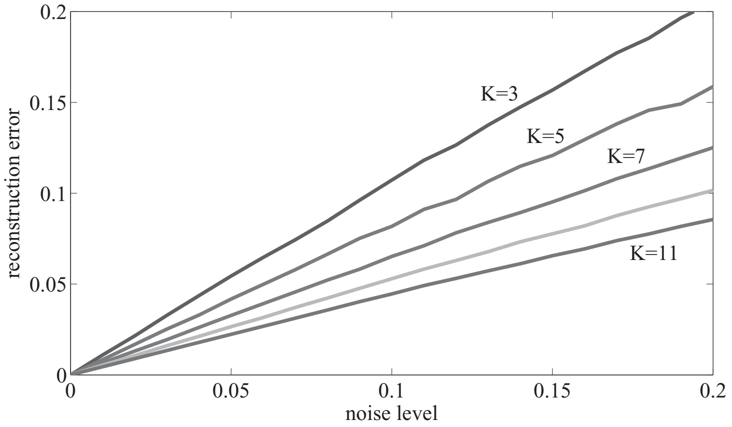

Fig. 1. Reconstruction error (measured in terms of RMSE) of pairwise coupling as a function of the level of noise (standard deviation) in the pairwise predictions; K corresponds to the number of class labels

$p_{i,j} = p_i/(p_i + p_j)$ if these are corrupted with noise; this ability is in fact crucial, since these pairwise components correspond to the predictions of the binary classifiers, which are never perfect.

Fig. 1 shows the expected RMSE between the true probability vector $\boldsymbol{p}$ and its coupled reconstruction (based on the method described in Section 2.1) when the pairwise probability estimates are given by the $p_{i,j}$ independently corrupted with additive Gaussian noise (truncated if necessary, so as to assure values in $[0, 1]$). More specifically, the figure shows the expected RMSE as a function of the noise level, measured in terms of the standard deviation. As can be seen, the reconstruction error does indeed increase almost linearly with the noise level. What is also interesting to observe, however, is that the error becomes smaller if the number of classes increases. This effect, which has also been observed for other types of pairwise learning methods, can be explained by the level of redundancy produced by the pairwise approach: Since the number of models (and hence the number of prediction errors) increases quadratically, there is a good chance to “average out” the individual prediction errors.

On the other side, the ranking-based approach will of course be affected by prediction errors, too. These errors are not easily comparable to the pairwise ones, but suppose that we add the same Gaussian noise to the individual components p_i of the vector $\boldsymbol{p}$. The “reconstruction” in this case simply comes down to renormalization. Fig. 2 plots the corresponding reconstruction error r against the reconstruction error r' of the pairwise coupling approach; the circles in this picture are centered at the points (r, r'), where both r and r' refer to the same underlying (true) probability vector. Interestingly, while the ranking-based approach seems to have an advantage in the case of a low number of labels (the cloud of circles for four labels is above the diagonal), this advantage turns into a disadvantage if the number of labels increases. Indeed, the larger the number of

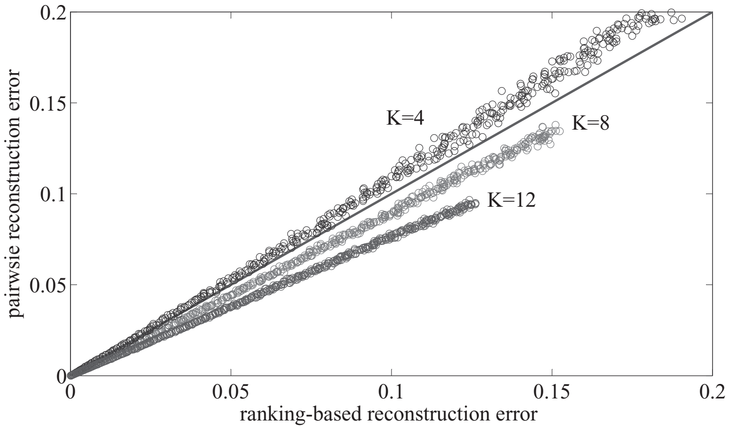

Fig. 2. Reconstruction error of the ranking-based approach (x-axis) versus reconstruction error of pairwise coupling (y-axis) for $K = 4, 8$ and 12 labels

labels, the more advantageous the pairwise coupling approach becomes; in the figure, the cases of eight and twelve labels are shown for illustration.

Needless to say, these results have to be interpreted with caution, since they are based on very idealized assumptions (e.g., independence of errors). Yet, they confirm an observation that was already made in previous studies of pairwise learning (albeit related to classification, not probability estimation): Due to the large number of binary models constructed, coming along with a high level of redundancy, the pairwise decomposition technique exhibits a kind of error-correction mechanism, and the larger the number of classes, the better this mechanism works [10].

7.2 Multi-class Classification

In the absence of benchmark data with given probabilities as ground-truth, we test our approach on standard classification benchmarks using the Brier score as a performance measure. The Brier score, which is commonly used for this purpose, compares a predicted probability vector $\boldsymbol{p} = (p_1, \ldots, p_K)$ with a true class $y \in \mathcal{Y}$ in terms of the following loss:

$$L(\boldsymbol{p}, y) = \sum_{i=1}^{K} \big(p_i - [\![y = y_i]\!]\big)^2$$

For comparison, we use pairwise coupling (PC) as described in Section 2.1 with logistic regression as base learner. Additionally, we used the pairwise coupling technique proposed by Hastie and Tibshirani [11], which is also quite commonly used for this purpose (PC-HT). Finally, we include one-vs-rest logistic regression (OVR) as a common approach to multi-class classification.

Table 2. Results in terms of average Brier score (± standard deviation)

data set	#ins.	#att.	#cls.	PC	PC-HT	OVR	PELARA
iris	150	4	3	0.044±0.045	0.044±0.045	0.087±0.042	0.043±0.050
glass	214	9	6	0.439±0.013	0.434±0.018	0.442±0.050	0.432±0.043
wine	178	13	3	0.044±0.028	0.044±0.028	0.037±0.023	0.044±0.025
vowel	528	10	11	0.246±0.054	0.241±0.044	0.555±0.047	0.389±0.063
vehicle	846	18	4	0.241±0.021	0.240±0.020	0.270±0.021	0.240±0.023
segment	2310	19	7	0.060±0.018	0.070±0.015	0.134±0.016	0.068±0.012
dna	2000	180	3	0.140±0.025	0.141±0.021	0.124±0.029	0.157±0.041
pendigits	7494	16	10	0.028±0.002	0.043±0.002	0.094±0.004	0.053±0.003
poker	25010	10	10	0.566±0.002	0.566±0.002	0.567±0.002	0.565±0.002
satimage	4435	36	6	0.189±0.012	0.190±0.012	0.246±0.009	0.198±0.011
svmguide4	300	10	6	0.642±0.015	0.716±0.005	0.715±0.008	0.737±0.006
svmguide2	391	20	3	0.275±0.034	0.259±0.032	0.277±0.032	0.266±0.034
letter	15000	16	26	0.228±0.009	0.291±0.006	0.473±0.005	0.336±0.008
shuttle	43500	9	7	0.068±0.003	0.067±0.002	0.135±0.003	0.061±0.002

Table 3. Runtimes in seconds for training each fold of the data; the relative runtimes are summarized in the brackets

data set	PC	PC-HT	OVR	PELARA
iris	0.19(1.63)	0.23(2.00)	0.13(1.14)	0.12(1)
glass	2.37(1.73)	2.18(1.59)	1.75(1.28)	1.37(1)
wine	0.24(1.88)	0.35(2.70)	0.33(2.51)	0.13(1)
vowel	6.08(1.04)	6.99(1.19)	0.74(0.13)	5.86(1)
vehicle	7.37(2.45)	5.51(1.83)	6.14(2.04)	3.01(1)
segment	18.80(1.77)	14.73(1.39)	17.84(1.68)	10.63(1)
dna	161.57(0.96)	166.18(0.99)	336.30(2.00)	168.54(1)
pendigits	25.87(1.30)	39.52(1.99)	46.09(2.32)	19.91(1)
poker	10.98(0.32)	62.70(1.83)	7.30(0.21)	34.29(1)
satimage	38.52(2.24)	44.13(2.57)	10.08(0.59)	17.16(1)
svmguide4	11.23(6.72)	5.42(3.25)	2.69(1.62)	1.67(1)
svmguide2	8.58(5.55)	3.07(1.98)	3.54(2.29)	1.55(1)
letter	179.76(0.33)	264.75(0.49)	25.13(0.05)	538.56(1)
shuttle	39.16(0.63)	90.00(1.44)	22.93(0.37)	62.32(1)

The results for various data sets from the UCI repository [9] are shown in Table 2, together with some statistics of the data. These results are averages over 5 repeats of 10-fold cross validation. As can be seen, OVR is clearly outperformed by the other methods. This is confirmed by a two-tailed sign test, which reports significance at the level $\alpha = 0.05$. PC, PC-HT and PELARA are almost perfectly en par (with similar numbers of wins and losses in each pairwise comparison).

The average runtimes are shown in Tables 3. Here, PELARA performs rather well and seems to be the most efficient on average. In particular, our ranking-based approach shows clear advantages over the pairwise coupling methods (while OVR is often quite fast, too).

8 Conclusions

While the problem of multi-class probability estimation is commonly tackled by means of reduction techniques, which decompose the original problem into a set of binary problems, we have proposed an alternative method that exploits the intimate connection between probability estimation and ranking. More specifically, we take advantage of recent work on *label ranking*, which provides a natural bridge between classification and probability estimation. This connection becomes especially apparent when making use of the Plackett-Luce model, a probabilistic ranking model that links classification and ranking in a seamless manner (by modeling ranking as a sequence of classifications).

Compared to the pairwise approach, our ranking-based method appears to be more solid from a theoretical point of view, especially as it does not require any ad-hoc aggregation mechanism. The corresponding reduction of complexity also comes with improvements in terms of runtime. Regarding predictive accuracy, however, the best approaches to pairwise coupling are indeed difficult to beat, especially if the number of classes is large. A plausible explanation for this observation, which is also coherent with similar findings for pairwise classification, is the redundancy produced by the quadratic number of pairwise models. Nevertheless, our results have shown that the ranking-based alternative put forward in this paper is at least competitive to state-of-the-art pairwise coupling methods.

Due to the lack of proper benchmark data, we could not yet explore what we suppose to be the main strength of our method, namely the learning of probability models from incomplete rankings, including pairwise comparisons of the form "y_i is more likely than y_j as a class label for $\boldsymbol{x}$, but also (qualitative) comparisons involving more than two labels, such as $y_3 \succ_{\boldsymbol{x}} y_5 \succ_{\boldsymbol{x}} y_1$. Currently, we are looking for data of that kind, which, despite not having been collected systematically so far, should in principle occur quite naturally in many domains.

Acknowledgments. The authors are supported by German Research Foundation (DFG).

References

1. Bradley, R., Terry, M.: Rank analysis of incomplete block designs I. the method of paired comparisons. Biometrika 39, 324–345 (1952)
2. Buja, A., Stuetzle, W., Shen, Y.: Loss functions for binary class probability estimation: Structure and applications. Technical report, University of Pennsylvania (2005)
3. Cheng, W., Dembczyński, K., Hüllermeier, E.: Label ranking methods based on the Plackett-Luce model. In: Proc. ICML 2010, pp. 215–222 (2010)
4. Cheng, W., Hühn, J., Hüllermeier, E.: Decision tree and instance-based learning for label ranking. In: Proc. ICML 2009, pp. 161–168 (2009)
5. Clemencon, S., Lugosi, G., Vayatis, N.: Ranking and empirical minimization of U-statistics. The Annals of Statistics 36(2), 844–874 (2008)
6. Cour, T., Sapp, B., Taskar, B.: Learning from partial labels. Journal of Machine Learning Research 12, 1225–1261 (2011)

7. Domingos, P., Pazzani, M.: On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning 29, 103–137 (1997)
8. Flach, P.A.: Putting Things in Order: On the Fundamental Role of Ranking in Classification and Probability Estimation. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) ECML 2007. LNCS (LNAI), vol. 4701, pp. 2–3. Springer, Heidelberg (2007)
9. Frank, A., Asuncion, A.: UCI machine learning repository (2010)
10. Fürnkranz, J.: Round robin classification. Journal of Machine Learning Research 2, 721–747 (2003)
11. Hastie, T., Tibshirani, R.: Classification by pairwise coupling. The Annals of Statistics 26(1), 451–471 (1998)
12. Herbei, R., Wegkamp, M.: Classification with reject option. Canadian Journal of Statistics 34(4), 709–721 (2006)
13. Luce, R.: Individual Choice Behavior: A Theoretical Analysis. Wiley (1959)
14. Mallows, C.: Non-null ranking models. Biometrika 44(1), 114–130 (1957)
15. Marden, J.: Analyzing and Modeling Rank Data. CRC Press (1995)
16. Niculescu-Mizil, A., Caruana, R.: Predicting good probabilities with supervised learning. In: Proc. ICML, pp. 625–632 (2005)
17. Plackett, R.: The analysis of permutations. Applied Statistics 24(2), 193–202 (1975)
18. Platt, J.C.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Advances in Large Margin Classifiers, pp. 61–74. MIT Press (1999)
19. Rifkin, R., Klautau, A.: In defense of one-vs-all classification. The Journal of Machine Learning Research 5, 101–141 (2004)
20. Wellman, M.P.: Some varieties of qualitative probability. In: Proc. IPMU 1994, Paris, pp. 437–442 (1994)
21. Wu, T., Lin, C., Weng, R.: Probability estimates for multi-class classification by pairwise coupling. Journal of Machine Learning Research 5, 975–1005 (2004)
22. Zadrozny, B., Elkan, C.: Learning and making decisions when costs and probabilities are both unknown. In: Proc. KDD, pp. 204–213 (2001)
23. Zadrozny, B., Elkan, C.: Transforming classifier scores into accurate multiclass probability estimates. In: Proc. KDD, pp. 694–699 (2002)
24. Zhang, T.: Statistical behavior and consistency of classification methods based on convex risk minimization. Annals of Statistics 32(1), 5–85 (2004)