A Simple Instance-Based Approach to Multilabel Classification Using the Mallows Model

Weiwei Cheng and Eyke Hüllermeier

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

Abstract. Multilabel classification is an extension of conventional classification in which a single instance can be associated with multiple labels. Recent research has shown that, just like for standard classification, instance-based learning algorithms relying on the nearest neighbor estimation principle can be used quite successfully in this context. In this paper, we propose a new instance-based approach to multilabel classification, which is based on calibrated label ranking, a recently proposed framework that unifies multilabel classification and label ranking. Within this framework, instance-based prediction is realized is the form of MAP estimation, assuming a statistical distribution called the Mallows model.

1 Introduction

In conventional classification, each instance is assumed to belong to exactly one among a finite set of candidate classes. As opposed to this, the setting of multilabel classification allows an instance to belong to several classes simultaneously or, say, to attach more than one label to an instance. Multilabel classification has received increasing attention in machine learning in recent years.

Even though quite a number of sophisticated methods for multilabel classification has been proposed in the literature, the application of *instance-based learning* (IBL) has not been studied very deeply in this context so far. This is a bit surprising, given that IBL algorithms based on the nearest neighbor estimation principle have been applied quite successfully in classification and pattern recognition for a long time [1]. A notable exception is the *multilabel k-nearest neighbor* (MLKNN) method that was recently proposed in [2], where it was shown to be competitive to state-of-the-art machine learning methods.

In this paper, we introduce a new instance-based approach to multilabel classification, which is based on calibrated label ranking, a recently proposed framework that unifies multilabel classification and label ranking (see Section 2). Within this framework, instance-based prediction is realized is the form of MAP estimation, assuming a statistical distribution called the Mallows model (see Section 3). Experimental results, presented in Section 5, provide evidence for the strong performance of this approach in terms of predictive accuracy.

2 Multilabel Classification as Calibrated Label Ranking

Let $\mathbb{X}$ denote an instance space and let $\mathcal{L} = \{\lambda_1, \lambda_2 \dots \lambda_m\}$ be a finite set of class labels. Moreover, suppose that each instance $\boldsymbol{x} \in \mathbb{X}$ can be associated with a subset of labels $L \in 2^{\mathcal{L}}$; this subset is often called the set of *relevant* labels, while the complement $\mathcal{L} \setminus L$ is considered as *irrelevant* for $\boldsymbol{x}$. Given training data in the form of a finite set T of observations in the form of tuples $(\boldsymbol{x}, L_{\boldsymbol{x}}) \in \mathbb{X} \times 2^{\mathcal{L}}$, typically assumed to be drawn independently from an (unknown) probability distribution on $\mathbb{X} \times 2^{\mathcal{L}}$, the goal in multilabel classification is to learn a classifier $h : \mathbb{X} \to 2^{\mathcal{L}}$ that generalizes well beyond these observations in the sense of minimizing the expected prediction loss with respect to a specific loss function.

Note that multilabel classification can be reduced to a conventional classification problem in a straightforward way, namely by considering each label subset $L \in 2^{\mathcal{L}}$ as a distinct (meta-)class. This approach is referred to as *label powerset* in the literature. An obvious drawback of this approach is the potentially large number of classes that one has to deal with in the newly generated problem. Another way of reducing multilabel to conventional classification is offered by the *binary relevance* (BR) approach. Here, a single binary classifier h_i is trained for each label $\lambda_i \in \mathcal{L}$. For a query instance $\boldsymbol{x}$, this classifier is supposed to predict whether λ_i is relevant for $\boldsymbol{x}$ ($h_i(\boldsymbol{x}) = 1$) or not ($h_i(\boldsymbol{x}) = 0$). A multilabel prediction for $\boldsymbol{x}$ is then given by $h(\boldsymbol{x}) = \{\lambda_i \in \mathcal{L} \,|\, h_i(\boldsymbol{x}) = 1\}$. Since binary relevance learning treats every label independently of all other labels, an obvious disadvantage of this approach is that it ignores potential correlations and interdependencies between labels.

Some of the more sophisticated approaches learn a multilabel classifier h in an indirect way via a scoring function $f : \mathbb{X} \times \mathcal{L} \to \mathbb{R}$ that assigns a real number to each instance/label combination. Such a function does not only allow one to make multilabel predictions (via thresholding the scores), but also offers the possibility to produce a ranking of the class labels, simply by ordering them according to their score. Sometimes, this ranking is even more desirable as a prediction, and indeed, there are several evaluation metrics that compare a true label subset with a predicted ranking instead of a predicted label subset.

In the following, we propose a formalization of multilabel classification within the framework of label ranking. More specifically, as will be seen, this framework allows one to combine the concepts of a ranking and a multilabel prediction (label subset) in a convenient way.

2.1 Label Ranking

The problem of *label ranking*, which has recently been introduced in machine learning [3, 4], can be seen as another extension of the conventional classification setting. Instead of associating every instance $\boldsymbol{x} \in \mathbb{X}$ with one among a finite set of class labels $\mathcal{L} = \{\lambda_1, \lambda_2 \dots \lambda_m\}$, we associate $\boldsymbol{x}$ with a total order of all 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. Since a ranking can be considered as a

special type of preference relation, we shall also say that $\lambda_i \succ_{\boldsymbol{x}} \lambda_j$ indicates that λ_i is *preferred* to λ_j given the instance $\boldsymbol{x}$.

Formally, a total order $\succ_{\boldsymbol{x}}$ can be identified with a permutation $\pi_{\boldsymbol{x}}$ of the set $\{1 \dots m\}$. 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 order. This permutation encodes the (ground truth) order relation

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

where $\pi_{\boldsymbol{x}}^{-1}(j)$ is the index of the label put at position j. The class of permutations of $\{1 \dots m\}$ (the symmetric group of order m) 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 $\mathbb{X} \to \Omega$ 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 $\mathbf{P}(\cdot \,|\, \boldsymbol{x})$ such that, for every $\pi \in \Omega$, $\mathbf{P}(\pi \,|\, \boldsymbol{x})$ is the probability that $\pi_{\boldsymbol{x}} = \pi$.

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 $\boldsymbol{x}_k$, $k = 1 \dots n$, together with information about one or more pairwise preferences of the form $\lambda_i \succ_{\boldsymbol{x}_k} \lambda_j$. 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 label pairs,

$$D(\pi, \sigma) \ = \ \#\{(i,j) \,|\, \pi(i) > \pi(j) \ \wedge \sigma(i) < \sigma(j)\} \,, \tag{1}$$

which is closely related to Kendall's tau coefficient. In fact, the latter is a normalization of (1) to the interval $[-1, +1]$. We shall focus on Kendall's tau as a natural, intuitive, and easily interpretable measure [5] 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 (1) is right-invariant, and so are most other commonly used metrics on Ω.

2.2 Calibrated Label Ranking

A label ranking provides information about the *relative* preference for labels, but not about the absolute preference or, say, relevance of a label. To combine the information offered by a label ranking and a multilabel classification (label subset), the concept of a *calibrated label ranking* has been proposed in [6]. A calibrated label ranking is a ranking of the label set Ω extended by a *neutral label* λ_0. The idea is that λ_0 splits a ranking into two parts, the positive (relevant) part consisting of those labels λ_i preceding λ_0 (i.e., $\lambda_i \succ_{\boldsymbol{x}} \lambda_0$), and the negative

(irrelevant) part given by those labels λ_j ranked lower than λ_0 (i.e., $\lambda_0 \succ_{\boldsymbol{x}} \lambda_j$). In this way, a multilabel prediction can be derived from a (predicted) calibrated label ranking.

The other way around, a multilabel set $L_{\boldsymbol{x}}$ translates into the set of pairwise preferences $\{\lambda \succ_{\boldsymbol{x}} \lambda' \,|\, \lambda \in L_{\boldsymbol{x}}, \lambda' \in \mathcal{L} \setminus L_{\boldsymbol{x}_i}\}$, and can hence be considered as *incomplete* information about an underlying calibrated label ranking. More specifically, $L_{\boldsymbol{x}}$ is consistent with the set of label rankings $E(L_{\boldsymbol{x}})$ given by those permutations $\pi \in \Omega$ that rank all labels in $L_{\boldsymbol{x}}$ higher and all labels in $\mathcal{L} \setminus L_{\boldsymbol{x}_i}$ lower than the neutral label λ_0. In the following, when we speak about a ranking, we always mean a calibrated ranking (i.e., Ω contains the neutral label λ_0).

3 Instance-Based Multilabel Classification

So far, no assumptions about the conditional probability measure $\mathbf{P}(\cdot \,|\, \boldsymbol{x})$ on Ω were made, despite its existence. To become more concrete, we resort to a popular and commonly used distance-based probability model introduced by Mallows [5]. The standard Mallows model is a two-parameter model that belongs to the exponential family:

$$\mathbf{P}(\sigma \,|\, \theta, \pi) = \frac{\exp(-\theta D(\pi, \sigma))}{\phi(\theta, \pi)} \tag{2}$$

The ranking $\pi \in \Omega$ is the location parameter (mode, center ranking) and $\theta \geq 0$ is a spread parameter.

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.

Coming back to the label ranking problem and the idea of instance-based learning, i.e., local prediction 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 $\boldsymbol{x}_i$ is associated with a subset $L_{\boldsymbol{x}_i} \subseteq \mathcal{L}$ of labels. 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 Ω is (at least approximately) *locally constant* around the query $\boldsymbol{x}$, so that the neighbors can be considered as a sample on the basis of which $\mathbf{P}(\cdot \,|\, \boldsymbol{x})$ can be estimated.

Thus, assuming an underlying (calibrated) label ranking, the probability to observe $L_{\boldsymbol{x}_i}$ is given by

$$\mathbf{P}(E(L_{\boldsymbol{x}_i})) \;=\; \sum_{\sigma \in E(L_{\boldsymbol{x}_i})} \mathbf{P}(\sigma \,|\, \theta, \pi) \;\;,$$

where $E(L_{\boldsymbol{x}_i})$ denotes the set of all label rankings consistent with $L_{\boldsymbol{x}_i}$. Making a simplifying assumption of independence, the probability of the complete set of observations $\mathbf{L} = \{L_{\boldsymbol{x}_1}, L_{\boldsymbol{x}_2} \dots L_{\boldsymbol{x}_k}\}$ then becomes

$$\begin{aligned}\mathbf{P}(\mathbf{L} \mid \theta, \pi) &= \prod_{i=1}^{k} \mathbf{P}(E(L_{\boldsymbol{x}_i}) \mid \theta, \pi) \\ &= \prod_{i=1}^{k} \sum_{\sigma \in E(L_{\boldsymbol{x}_i})} \mathbf{P}(\sigma \mid \theta, \pi) \\ &= \frac{\prod_{i=1}^{k} \sum_{\sigma \in E(L_{\boldsymbol{x}_i})} \exp\left(-\theta D(\sigma, \pi)\right)}{\left(\prod_{j=1}^{m} \frac{1-\exp(-j\theta)}{1-\exp(-\theta)}\right)^k} \quad .\end{aligned} \tag{3}$$

Instance-based prediction of the (calibrated) label ranking $L_{\boldsymbol{x}}$ can now be posed as a Maximum Likelihood problem, namely as finding the Maximum Likelihood estimation (MLE) of π (and θ) in (3). This problem is extremely difficult in general. Fortunately, in the context of multi-label classification, we are able to exploit the special structure of the observations. More specifically, we can show the following theorem (proof omitted).

Theorem 1: For each label $\lambda_i \in \mathcal{L}$, let $f(\lambda_i)$ denote the frequency of occurrence of this label in the neighborhood of $\boldsymbol{x}$, i.e., $f(\lambda_i) = \#\{j \mid \lambda_i \in L_{\boldsymbol{x}_j}\}/k$. Moreover, let $f(\lambda_0) = 1/2$ by definition. Then, a ranking $\pi \in \Omega$ is a MLE in (3) iff it guarantees that $f(\lambda_i) > f(\lambda_j)$ implies $\pi(i) < \pi(j)$.

According to this result, an optimal ranking and, hence, an optimal multi-label prediction can simply be found by sorting the labels according to their frequency of occurrence in the neighborhood. A disadvantage of this estimation is its ambiguity in the presence of ties: If two labels have the same frequency, they can be ordered in either way. Interestingly, we can remove this ambiguity by replacing the MLE by a Bayes estimation.

Theorem 2: Let $g(\lambda_i)$ denote the frequency of occurrence of the label λ_i in the complete training set. There exists a prior distribution $\mathbf{P}$ on Ω such that, for large enough k, a ranking $\pi \in \Omega$ is a maximum posterior probability (MAP) estimation iff it guarantees the following: If $f(\lambda_i) > f(\lambda_j)$ or $f(\lambda_i) = f(\lambda_j)$ and $g(\lambda_i) > g(\lambda_j)$, then $\pi(i) < \pi(j)$.

This result suggests a very simple prediction procedure: Labels are sorted according to their frequency in the neighborhood of the query, and ties are broken by resorting to global information outside the neighborhood, namely the label frequency in the complete training data (which serve as estimates of the unconditional probability of a label).

4 Related Work

Multilabel classification has received a great deal of attention in machine learning in recent years, and a number of methods has been developed, often motivated by specific types of applications such as text categorization [7–10], computer vision[11], and bioinformatics [12, 13, 10]. Besides, several well-established methods for conventional classification have been extended to the multi-label case, including support vector machines [14, 13, 11], neural networks [10], and decision trees [15].

Our interest in instance-based multilabel classification is largely motivated by the *multilabel k-nearest neighbor* (MLKNN) method that has recently been proposed in [2]. In that paper, the authors show that MLKNN performs quite well in practice. In the concrete experiments presented, MLKNN even outperformed some state-of-the-art model-based approaches to multilabel classification, including RankSVM and AdaBoost.MH [13, 16].

MLKNN is a binary relevance learner, i.e., it learns a single classifier h_i for each label $\lambda_i \in \mathcal{L}$. However, instead of using the standard k-nearest neighbor (KNN) classifier as a base learner, it implements the h_i by means of a combination of KNN and Bayesian inference: Given a query instance $\boldsymbol{x}$ with unknown multilabel classification $L \subseteq \mathcal{L}$, it finds the k nearest neighbors of $\boldsymbol{x}$ in the training data and counts the number of occurrences of λ_i among these neighbors. Considering this number, y, as information in the form of a realization of a random variable Y, the posterior probability of $\lambda_i \in L$ is given by

$$\mathbf{P}(\lambda_i \in L \,|\, Y = y) = \frac{\mathbf{P}(Y = y \,|\, \lambda_i \in L) \cdot \mathbf{P}(\lambda_i \in L)}{\mathbf{P}(Y = y)} \;, \tag{4}$$

which leads to the decision rule

$$h_i(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \mathbf{P}(Y = y \,|\, \lambda_i \in L)\mathbf{P}(\lambda_i \in L) \geq \mathbf{P}(Y = y \,|\, \lambda_i \notin L)\mathbf{P}(\lambda_i \notin L) \\ 0 & \text{otherwise} \end{cases}$$

The prior probabilities $\mathbf{P}(\lambda_i \in L)$ and $\mathbf{P}(\lambda_i \notin L)$ as well as the conditional probabilities $\mathbf{P}(Y = y \,|\, \lambda_i \in L)$ and $\mathbf{P}(Y = y \,|\, \lambda_i \notin L)$ are estimated from the training data in terms of corresponding relative frequencies. While the estimation of the former probabilities is uncritical from a computational point of view, the estimation of the conditional probabilities can become quite expensive. Essentially, it requires the consideration of all k-neighborhoods of all training instances, and the counting of the number of occurrences of each label within these neighborhoods. Implementing nearest neighbor search in a naive way, namely by linear search, this would mean a complexity of $O(kn^2)$, where n is the size of the training data. Of course, this complexity can be reduced by using more efficient algorithms and data structures for nearest neighbor search; for example, the *all nearest neighbors* problem, i.e., the problem to find the (first) nearest neighbor for each element of a data set, can be solved in time $O(n \log(n))$ [17]. Nevertheless, the computational overhead produced by this kind of preprocessing on the training data will remain a dominating factor for the overall runtime of the method.

Table 1. Statistics for the multilabel data sets used in the experiments. The symbol * indicates that the data set contains binary features; *cardinality* is the average number of labels per instance.

DATA SET	DOMAIN	#INSTANCES	#ATTRIBUTES	#LABELS	CARDINALITY
emotions	music	593	72	6	1.87
image	vision	2000	135	5	1.24
genbase	biology	662	1186*	27	1.25
mediamill	multimedia	5000	120	101	4.27
reuters	text	7119	243	7	1.24
scene	vision	2407	294	6	1.07
yeast	biology	2417	103	14	4.24

5 Experimental Results

This section is devoted to experimental studies that we conducted to get a concrete idea of the performance of our method. Before presenting results, we give some information about the learning algorithms and data sets included in the study, as well as the criteria used for evaluation.

5.1 Learning Algorithms

For the reasons mentioned previously, our main interest is focused on MLKNN, which is arguably the state-of-the-art in instance-based multilabel ranking; we used its implementation in the MULAN package.[1] MLKNN is parameterized by the size of the neighborhood, for which we adopted the value $k = 10$. This value is recommended in [2], where it was found to yield the best performance. For the sake of fairness, we use the same neighborhood size for our method (Mallows). In both cases, the simple Euclidean metric (on the complete attribute space) was used as a distance function. As an additional baseline we used binary relevance learning (BR) with C4.5 (the WEKA [18] implementation J48 in its default setting) as a base learner.

5.2 Data Sets

Benchmark data for multi-label classification is not as abundant as for conventional classification, and indeed, experiments in this field are often restricted to a very few or even only a single data set. For our experimental study, we have collected a comparatively large number of seven data sets from different domains; an overview is given in Table 1.[2]

The *emotions* data was created from a selection of songs from 233 musical albums [19]. From each song, a sequence of 30 seconds after the initial 30 seconds

[1] `http://mlkd.csd.auth.gr/multilabel.html`

[2] All data sets are public available at `http://mlkd.csd.auth.gr/multilabel.html` and `http://lamda.nju.edu.cn/data.htm`.

was extracted. The resulting sound clips were stored and converted into wave files of 22050 Hz sampling rate, 16-bit per sample and mono. From each wave file, 72 features have been extracted, falling into two categories: rhythmic and timbre. Then, in the emotion labeling process, 6 main emotional clusters are retained corresponding to the Tellegen-Watson-Clark model of mood: amazed-surprised, happy-pleased, relaxing-clam, quiet-still, sad-lonely and angry-aggressive.

Image and *scene* are semantic scene classification data sets proposed, respectively, by [20] and [11], in which a picture can be categorized into one or more classes. In the scene data, for example, pictures can have the following classes: beach, sunset, foliage, field, mountain, and urban. Features of this data set correspond to spatial color moments in the LUV space. Color as well as spatial information have been shown to be fairly effective in distinguishing between certain types of outdoor scenes: bright and warm colors at the top of a picture may correspond to a sunset, while those at the bottom may correspond to a desert rock. Features of the image data set are generated by the SBN method [21] and essentially correspond to attributes in an RGB color space.

From the biological field, we have chosen the two data sets *yeast* and *genbase*. The yeast data set is about predicting the functional classes of genes in the Yeast Saccharomyces cerevisiae. Each gene is described by the concatenation of microarray expression data and a phylogenetic profile, and is associated with a set of 14 functional classes. The data set contains 2417 genes in total, and each gene is represented by a 103-dimensional feature vector. In the *genbase* data, 27 important protein families are considered, including, for example, PDOC00064 (a class of oxydoreductases) and PDOC00154 (a class of isomerases). During the preprocessing, a training set was exported, consisting of 662 proteins that belong to one or more of these 27 classes.

From the text processing field, we have chosen a subset of the widely studied *Reuters-21578* collection [22]. The seven most frequent categories are considered. After removing documents whose label sets or main texts are empty, 8866 documents are retained where only 3.37% of them are associated with more than one class label. After randomly removing documents with only one label, a text categorization data set containing 2,000 documents is obtained. Each document is represented as a bag of instances using the standard sliding window techniques, where each instance corresponds to a text segment enclosed in one sliding window of size 50 (overlapped with 25 words). "Function words" are removed from the vocabulary and the remaining words are stemmed. Instances in the bags adopt the "bag-of-words" representation based on term frequency. Without loss of effectiveness, dimensionality reduction is performed by retaining the top 2% words with highest document frequency. Thereafter, each instance is represented as a 243-dimensional feature vector.

The *mediamill* data set is from the field of multimedia indexing and originates from the well-known TREC Video Retrieval Evaluation data (TRECVID 2005/2006) initiated by American National Institute of Standards and Technology (NIST), which contains 85 hours of international broadcast news data. The task in this data set is the automated detection of a lexicon of 101 semantic

Table 2. Experimental results in terms of Hamming loss (left) and rank loss (right).

DATA SET	MLKNN	Mallows	BR	MLKNN	Mallows	BR
emotions	0.261	0.197	0.253	0.262	0.163	0.352
genbase	0.005	0.003	0.001	0.006	0.006	0.006
image	0.193	0.192	0.243	0.214	0.208	0.398
mediamill	0.027	0.027	0.032	0.037	0.036	0.189
reuters	0.073	0.085	0.057	0.068	0.087	0.089
scene	0.087	0.094	0.131	0.077	0.088	0.300
yeast	0.194	0.197	0.249	0.168	0.165	0.360

concepts in videos. Every instance of this data set has 120 numeric features including visual, textual, as well as fusion information. The trained classifier should be able to categorize an unseen instance to some of these 101 labels, e.g., face, car, male, soccer, and so on. More details about this data set can be found at [23].

5.3 Evaluation Measures

To evaluate the performance of multilabel classification methods, a number of criteria and metrics have been proposed in the literature. For a classifier h, let $h(\boldsymbol{x}) \subseteq \mathcal{L}$ denote its multilabel prediction for an instance $\boldsymbol{x}$, and let $L_{\boldsymbol{x}}$ denote the true set of relevant labels. The *Hamming loss* computes the percentage of labels whose relevance is predicted incorrectly:

$$\textsc{HamLoss}(h) = \frac{1}{|\mathcal{L}|}\big|h(\boldsymbol{x}) \,\Delta\, L_{\boldsymbol{x}}\big|, \tag{5}$$

where Δ is the symmetric difference between two sets.

To measure the ranking performance, we used the *rank loss*, which computes the average fraction of label pairs that are not correctly ordered:

$$\textsc{RankLoss}(f) = \frac{\#\{(\lambda,\lambda') \,|\, \pi_{\boldsymbol{x}}(\lambda) \leq \pi_{\boldsymbol{x}}(\lambda'), (\lambda,\lambda') \in L_{\boldsymbol{x}} \times \overline{L_{\boldsymbol{x}}}\}}{|L_{\boldsymbol{x}}||\overline{L_{\boldsymbol{x}}}|}, \tag{6}$$

where $\pi_{\boldsymbol{x}}(\lambda)$ denotes the position assigned to label λ for instance $\boldsymbol{x}$, and $\overline{L_{\boldsymbol{x}}} = \mathcal{L} \setminus L_{\boldsymbol{x}}$ is the set of irrelevant labels.

5.4 Results

The results of a cross validation study (10-fold, 5 repeats) are summarized in Table 2. As can be seen, both instance-based approaches perform quite strongly in comparison to the baseline, which is apparently not competitive. The instance-based approaches themselves are more or less en par, with a slight though statistically non-significant advantage for our method.

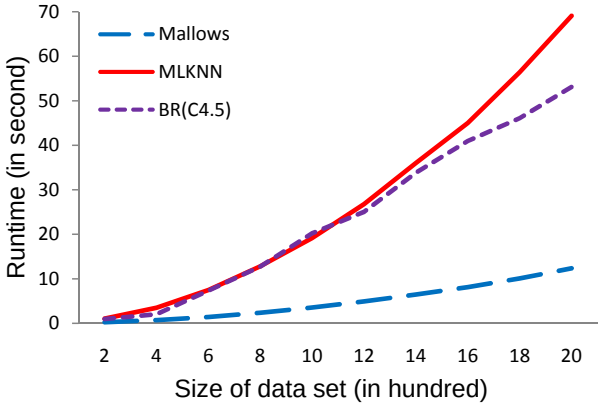

Fig. 1. Runtime of the methods on the *image* data.

As discussed in the previous section, MLKNN is expected to be less efficient from a computational point of view, and this expectation was confirmed by our experiments. Indeed, our approach scales much better than MLKNN. A typical example is shown in Fig. 1, where the runtime (total time needed to conduct a 10-fold cross validation) is plotted as a function of the size of the data; to obtain data sets of different size, we sampled from the *image* data.

6 Summary and Conclusions

According to the literature, MLKNN can be considered as the state-of-the-art in instance-based multilabel classification. In this paper, we have presented an alternative instance-based multilabel classifier, which is (at least) competitive in terms of predictive accuracy, while being computationally more efficient. In fact, our approach comes down to a very simple prediction procedure, in which labels are sorted according to their local frequency in the neighborhood of the query, and ties are broken by global frequencies. Despite its simplicity, this approach is well justified in terms of an underlying theoretical model.

References

1. Aha, D., Kibler, D., Alber, M.: Instance-based learning algorithms. Machine Learning **6**(1) (1991) 37–66
2. Zhang, M.L., Zhou, Z.H.: ML-kNN: A lazy learning approach to multi-label learning. Pattern Recognition **40**(7) (2007) 2038–2048
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. Volume 15. (2003) 785–792
4. Hüllermeier, E., Fürnkranz, J., Cheng, W., Brinker, K.: Label ranking by learning pairwise preferences. Artificial Intelligence **172**(16-17) (2008) 1897–1916

5. Mallows, C.: Non-null ranking models. In: Biometrika. Volume 44., Biometrika Trust (1957) 114–130
6. Fürnkranz, J., Hüllermeier, E., Mencia, E., Brinker, K.: Multilabel classification via calibrated label ranking. Machine Learning **73**(2) (2008) 133–153
7. Schapire, R.E., Singer, Y.: Boostexter: a boosting-based system for text categorization. Machine Learning **39**(2) (2000) 135–168
8. Ueda, N., Saito, K.: Parametric mixture models for multi-label text. In Becker, S., Thrun, S., Obermayer, K., eds.: Advances in Neural Information Processing. Volume 15., Cambridge MA, MIT Press (2003) 721–728
9. Kazawa, H., Izumitani, T., Taira, H., Maeda, E.: Maximal margin labeling for multi-topic text categorization. In Saul, L.K., Weiss, Y., Bottou, L., eds.: Advances in Neural Inf. Proc. Syst. Volume 17., Cambridge MA, MIT Press (2005)
10. Zhang, M.L., Zhou, Z.H.: Multi-label neural networks with applications to functional genomics and text categorization. In: IEEE Transactions on Knowledge and Data Engineering. Volume 18. (2006) 1338–1351
11. Boutell, M.R., Luo, J., Shen, X., Brown, C.M.: Learning multi-label scene classification. Pattern Recognition **37**(9) (2004) 1757–1771
12. Clare, A., King, R.D.: Knowledge discovery in multi-label phenotype data. In Raedt, L.D., Siebes, A., eds.: Lecture Ntes in Computer Science. Volume 2168., Berlin, Springer (2001) 42–53
13. Elisseeff, A., Weston, J.: A kernel method for multi-labelled classification. In Dietterich, T.G., Becker, S., Z.Ghahramani, eds.: Advances in Neural Information Processing Systems. Volume 14., Cambridge MA, MIT Press (2002) 681–687
14. Godbole, S., Sarawagi, S.: Discriminative methods for multi-labeled classiffication. In: Advances in Knowledge Discovery and Data Mining. Volume 3056 of LNCS., Springer (2004) 20–33
15. Vens, C., Struyf, J., Schietgat, L., Dzeroski, S., Blockeel, H.: Decision trees for hierarchical multi-label classification. Machine Learning **73** (2008) 185–214
16. Comite, F.D., Gilleron, R., Tommasi, M.: Learning multi-label alternating decision tree from texts and data. In Perner, P., Rosenfeld, A., eds.: Lecture Notes in Computer Science. Volume 2734., Berlin, Springer (2003) 35–49
17. Vaidya, P.: An O(n log n) algorithm for the all-nearest-neighbors problem. Discrete and Computational Geometry **4**(1) (1989) 101–115
18. Witten, I., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. 2nd edn. Morgan Kaufmann, San Francisco, CA, USA (2005)
19. Trohidis, K., Tsoumakas, G., Kalliris, G., Vlahavas, I.: Multilabel classification of music into emotions. In: Proc. Int. Conf. Music Information Retrieval. (2008)
20. Zhou, Z.H., Zhang., M.L.: Multi-instance multi-label learning with application to scene classification. In Schölkopf, B., Platt, J., Hofmann, T., eds.: Advances in Neural Inf. Proc. Syst. Volume 19., Cambridge MA, MIT Press (2007) 1609–1616
21. Maron, O., Ratan, A.L.: Multiple-instance learning for natural scene classification. In: Proc. ICML, Madison WI (1998) 341–349
22. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys **34**(1) (2002) 1–47
23. Snoek, C.G.M., Worring, M., van Gemert, J.C., Geusebroek, J.M., Smeulders, A.W.M.: The challenge problem for automated detection of 101 semantic concepts in multimedia. In: Proc. ACM Multimedia, Santa Barbara, USA (2006) 421–430