Multilabel Prediction with Probability Sets: The Hamming Loss Case.

Sébastien Destercke

▶ **To cite this version:**

Sébastien Destercke. Multilabel Prediction with Probability Sets: The Hamming Loss Case.. International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2014), Jul 2014, france, France. pp.496-505. hal-01044994

HAL Id: hal-01044994
https://hal.science/hal-01044994

Submitted on 24 Jul 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Multilabel prediction with probability sets: the Hamming loss case

Sebastien Destercke

Heudiasyc, UMR 7253, Centre de recherche de royallieu, 60203 Compiegne, France
sebastien.destercke@hds.utc.fr

Abstract. In this paper, we study how multilabel predictions can be obtained when our uncertainty is described by a convex set of probabilities. Such predictions, typically consisting of a set of potentially optimal decisions, are hard to make in large decision spaces such as the one considered in multilabel problems. However, we show that when considering the Hamming loss, an approximate prediction can be efficiently computed from label-wise information, as in the precise case. We also perform some first experiments showing the interest of performing partial predictions in the multilabel case.

Keywords: Credal sets, multilabel, indeterminate classification, k-nn, binary relevance

1 Introduction

The problem of multi-label classification, which generalizes the traditional (single label) classification setting by allowing multiple labels to belong simultaneously to an instance, has recently attracted a lot of attention. Such a setting indeed appears in a lot of cases: a film can belong to multiple categories, a music can stir multiple emotions [12], proteins can possess multiple functions [14], etc. In such problems, obtaining a complete ground truth (sets of relevant labels) for the training data and making accurate predictions is more complex than in traditional classification, in which the aim is to predict a single label.

In such a setting the appearance of incomplete observations, i.e., instances for which we do not know whether some labels are relevant, is much more likely. For example, a user may be able to tag a movie as a comedy and not as a science-fiction movie, but may hesitate about whether it should be tagged as a drama. Other examples include cases with high number of labels and where an expert cannot be expected to provide all relevant ones. Such partial labels are commonly called weak labels [9] and are common in problems such as image annotation [10] or protein function prediction [14].

Even when considering weak labels, all multilabel methods we are aware of still produce complete predictions as outputs. However, given the complexity of the predictions to make and the likely presence of missing data, it may be sensible to look for means to do cautious yet more trustful predictions. That is it may be interesting for the learner to abstain to make a prediction about a label whose relevance is too uncertain, so that the final prediction is partial but more accurate. Such an approach can be seen

as an extension of the reject option implemented in learning problems [1] or of the fact of making partial predictions [4], and has been recently investigated for the related problem of label ranking [3,2].

In this paper, we consider the problem of making partial predictions in the multilabel setting using convex sets of probabilities, or credal sets [8], as our predictive model. Indeed, making partial predictions is one central feature of approaches using credal sets [4], and these approaches are also well-designed to cope with the problem of missing or incomplete data [15]. However, making partial predictions with credal sets in large decision space such as the one considered in mutlilabel is usually difficult.

In Section 3, we nevertheless demonstrate that when focusing on the Hamming loss, obtaining approximate partial predictions can be done in a quite efficient way by focusing on label-wise information. We then perform (Section 4) some experiments to demonstrate the interest of making partial predictions in the multilabel setting. Section 2 presents necessary background material.

2 Preliminary material

In this section, we introduce the multilabel setting as well as basic notions needed to deal with sets of probabilities.

2.1 Multilabel problem setting

The usual goal of classification problems is to associate an instance $\mathbf{x}$ coming from an instance space $\mathscr{X}$ to a single (preferred) label of the space $\Lambda = \{\lambda_1, \ldots, \lambda_m\}$ of possible classes. In a multilabel setting, an instance $\mathbf{x}$ is associated to a subset $L_{\mathbf{x}} \subset \Lambda$ of labels, often called the subset of relevant labels while its complement $\Lambda \setminus L_{\mathbf{x}}$ is considered as irrelevant. We denote by $\mathscr{Y} = \{0,1\}^m$ the set of m-dimensional binary vectors, and identify a set L of relevant labels with a binary vector $y = (y_1, \ldots, y_m) \in \mathscr{Y}$ such that $y_i = 1$ if and only if $\lambda_i \in L$. As there is a one-to-one mapping between subsets L of Λ and $\mathscr{Y}$, we will indifferently work with one or the other.

The task in a multilabel problem is the same as in usual classification: to use the training instances $(\mathbf{x}^j, y^j)$, $j = 1, \ldots, n$ to estimate the theoretical conditional probability measure $P_{\mathbf{x}} : 2^{\mathscr{Y}} \to [0,1]$ associated to an instance $\mathbf{x} \in \mathscr{X}$. Ideally, observed outputs y^j should be completely specified vectors , however it may be the case that the value for some component y_i^j is unknown, which will be denoted by $y_i^j = *$. We will denote incomplete vectors by capital Y. Alternatively, an incomplete vector Y can be characterized by two sets $\underline{L} \subseteq \overline{L} \subseteq \Lambda$ of necessarily and possible relevant labels, defined as $\underline{L} := \{\lambda_i | y_i = 1\}$ and $\overline{L} := \{\lambda_i | y_i = 1 \vee y_i = *\}$ respectively. An incomplete vector Y describes a corresponding set of complete vectors, obtained by replacing each $y_i = *$ either by 1 or 0, or equivalently by considering any subset L such that $\underline{L} \subseteq L \subseteq \overline{L}$. To simplify notations, in the sequel we will use the same notation for an incomplete vector and its associated set of complete vectors.

Example 1. Table 1 provides an example of a multilabel data set with $\Lambda = \{\lambda_1, \lambda_2, \lambda_3\}$. $Y^3 = [* \; 1 \; 0]$ is an incomplete observed instance with $\underline{L}^3 = \{\lambda_2\}$ and $\overline{L}^3 = \{\lambda_1, \lambda_2\}$. Its corresponding set of complete vectors is $\{[0 \; 1 \; 0], [1 \; 1 \; 0]\}$

X_1	X_2	X_3	X_4	y_1	y_2	y_3
107.1	25	*Blue*	60	1	0	0
−50	10	*Red*	40	1	0	1
200.6	30	*Blue*	58	$*$	1	0
107.1	5	*Green*	33	0	1	$*$
...	...	...	...	...	...	...

Table 1. Multilabel data set example

In multilabel problems the size of the prediction space increases exponentially with m ($|\mathscr{Y}| = 32768$ for $m = 15$), meaning that estimating directly $P_{\mathbf{x}}$ will be intractable even for limited sizes of Λ. As a means to solve this issue, different authors have proposed so-called transformation techniques [13] that reduce the initial problem (in which 2^m parameters have to be estimated) into a set of simpler problems. For instance Binary Relevance (BR) consists in predicting relevance label-wise, solving an independent binary problem for each label. It therefore comes down to estimate m parameters $P_{\mathbf{x}}(y_i)$, $i = 1,\ldots,m$ and to predict $\hat{y}_i = 1$ if $P_{\mathbf{x}}(y_i = 1) \geq 1/2$. A common critic of the BR approach is that it does not take account of label dependencies, however it has been shown that this approach is theoretically optimal for the Hamming loss, on which this paper focuses [5]. Other reduction approaches include, for instance, Calibrated Ranking (CR) [7] that focuses on pairwise comparisons.

Another issue is that making a precise and accurate estimation of $P_{\mathbf{x}}$ is an extremely difficult problem given the number 2^m of alternatives and the possible presence of missing data. This problem is even more severe if little data are available, and this is why making cautious inferences (i.e., partial predictions) using as model a (convex) set $\mathscr{P}_{\mathbf{x}}$ of probability distributions may be interesting in the multilabel setting.

2.2 Notions about probability sets

We assume that our uncertainty is described by a convex set of probabilities $\mathscr{P}_{\mathbf{x}}$ defined over $\mathscr{Y}$ rather than by a precise probability measure $P_{\mathbf{x}}$. Such a set is usually defined either by a collection of linear constraints on the probability masses or by a set of extreme probabilities. Given such a set, we can define for any event $A \subseteq \mathscr{Y}$ the notions of lower and upper probabilities $\underline{P}_{\mathbf{x}}(A)$ and $\overline{P}_{\mathbf{x}}(A)$, respectively defined as

$$\underline{P}_{\mathbf{x}}(A) = \inf_{P_{\mathbf{x}} \in \mathscr{P}_{\mathbf{x}}} P_{\mathbf{x}}(A) \text{ and } \overline{P}_{\mathbf{x}}(A) = \sup_{P_{\mathbf{x}} \in \mathscr{P}_{\mathbf{x}}} P_{\mathbf{x}}(A).$$

Lower and upper probabilities are dual, in the sense that $\underline{P}(A) = 1 - \overline{P}(A^c)$. Similarly, if we consider a real-valued bounded function $f : \mathscr{Y} \to \mathbb{R}$, the lower and upper expectations $\underline{\mathbb{E}}_{\mathbf{x}}(f)$ and $\overline{\mathbb{E}}_{\mathbf{x}}(f)$ are defined as

$$\underline{\mathbb{E}}_{\mathbf{x}}(f) = \inf_{P_{\mathbf{x}} \in \mathscr{P}_{\mathbf{x}}} \mathbb{E}_{\mathbf{x}}(f) \text{ and } \overline{\mathbb{E}}_{\mathbf{x}}(f) = \sup_{P_{\mathbf{x}} \in \mathscr{P}_{\mathbf{x}}} \mathbb{E}_{\mathbf{x}}(f),$$

where $\mathbb{E}_{\mathbf{x}}(f)$ is the expectation of f w.r.t. P. Lower and upper expectations are also dual, in the sense that $\underline{\mathbb{E}}(f) = -\overline{\mathbb{E}}(-f)$. They are also scale and translation invariant in the sense that given two numbers $\alpha \in \mathbb{R}^+, \beta \in \mathbb{R}$, we have $\underline{\mathbb{E}}(\alpha f + \beta) = \alpha \underline{\mathbb{E}}(f) + \beta$

In the next sections, we explore how the multilabel problem can be solved with such credal estimates. We discuss the problem, usually computationally intensive, of making partial decision and show that it can be simplified when considering the Hamming loss as our loss function. Using these results, we then perform some experiment based on label-wise decomposition and k-nn algorithms to assess the interest of making partial predictions based on credal sets.

3 Credal multilabel predictions with Hamming loss

In this section, we first recall the principle of credal predictions, before proceeding to show that in the case of Hamming loss, such predictions can be efficiently approximated by an outer-approximation.

3.1 Credal predictions

Once a space $\mathscr{Y}$ of possible observations is defined, selecting a prediction, or equivalently making a decision, requires to define:

- a space $\mathscr{A} = \{a_1, \ldots, a_d\}$ of possible actions (sometimes equal to $\mathscr{Y}$, but not necessarily);
- a loss function $\ell : \mathscr{A} \times \mathscr{Y} \to \mathbb{R}$ such that $\ell(a,y)$ is the loss associated to action a when y is the ground-truth.

Given an instance $\mathbf{x}$ and a precise estimate $\hat{P}_\mathbf{x}$, a decision a will be preferred to a decision a' under loss function ℓ, denote $a \succ_\ell a'$, if

$$\mathbb{E}_\mathbf{x}\left(\ell(a',\cdot) - \ell(a,\cdot)\right) = \sum_{y\in\mathscr{Y}} \hat{P}_\mathbf{x}(y)\left(\ell(a',y) - \ell(a,y)\right) > 0, \tag{1}$$

where $\mathbb{E}_\mathbf{x}$ is the expectation w.r.t. $\hat{P}_\mathbf{x}$. This equation means that exchanging a' for a would incur a positive expected loss, therefore a should be preferred to a'. In the case of a precise estimate $\hat{P}_\mathbf{x}$, $\succ_\ell$ is a complete pre-order and the prediction comes down to take the maximal element of this pre-order, i.e.,

$$\hat{a}_\ell = \arg\min_{a\in\mathscr{A}} \mathbb{E}_\mathbf{x}\left(\ell(a,\cdot)\right) = \arg\min_{a\in\mathscr{A}} \sum_{y\in\mathscr{Y}} \hat{P}_\mathbf{x}(y)\ell(a,y) \tag{2}$$

that is to minimize the expected loss (ties can be broken arbitrarily, as they will lead to the same expected loss). This means that finding the best action (or prediction) will therefore requires d computations of expectations.

When considering a set $\mathscr{P}_\mathbf{x}$ as cautious estimate, there are many ways [11] to extend Eq. (1), but the most well-founded is the maximality criterion, which states that $a \succ_\ell a'$, if

$$\underline{\mathbb{E}}_\mathbf{x}\left(\ell(a',\cdot) - \ell(a,\cdot)\right) > 0, \tag{3}$$

that is if exchanging a' for a is guaranteed to give a positive expected loss. In such a case, the relation $\succ_\ell$ will be a partial order, and the maximal set $\hat{A}_\ell$ of alternatives will be chosen as a prediction, that is

$$\hat{A}_\ell = \{a \in \mathscr{A} |\ \nexists a' \in \mathscr{A} \text{ s.t. } a' \succ_\ell a\}. \tag{4}$$

Computing $\hat{A}_\ell$ requires at worst $d(d-1)$ computations, a quadratic number of comparisons with respect to the number of alternatives. Also notes that evaluating Eq. (3) usually requires solving a linear programming problem, a computationally more intensive task than evaluating Eq. (1). $\hat{A}_\ell$ is a cautious prediction, since it considers a set of potential optimal solutions.

Multilabel loss functions usually considers the set $\mathscr{A} = \mathscr{Y}$ as possible actions, or even bigger sets (for example the ranking loss considers as actions the set of complete orders over Λ). This means that getting $\hat{a}_\ell$ is already quite hard in the general case, hence computing $\hat{A}_\ell$ will be intractable in most cases, as the worst number of computation will then be 2^{2m} ($m = 15$ labels means at worst $\sim 10^9$ comparisons).

In the next subsection, we show that for the Hamming loss ℓ_H, we can get an outer approximation of $\hat{A}_\ell$ at an affordable computational cost. Offering such efficient way to make cautious predictions based on $\mathscr{P}_{\mathbf{x}}$ is essential to be able to use such kind of models in complex problems.

3.2 The Hamming loss

Let the set of alternatives be $\mathscr{A} = \mathscr{Y}$. Given an observation y and a prediction $\hat{y}$, Hamming loss ℓ_H reads

$$\ell_H(\hat{y}, y) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{(\hat{y}_i \neq y_i)} \,. \tag{5}$$

It counts the number of labels for which our prediction is wrong, and normalizes it. When the estimate $\hat{P}_{\mathbf{x}}$ is precise, it is known [5] that the optimal decision is the vector $\hat{y}$ such that $\hat{y}_j = 1$ if $\hat{P}_{\mathbf{x}}(y_j = 1) \geq 1/2$ and $\hat{y}_j = 0$ else. In particular, this means that the optimal decision can be derived from the sole knowledge of the marginals $\hat{P}_{\mathbf{x}}(y_j = 1)$, $j = 1, \ldots, n$, provided they are good estimates of $P_{\mathbf{x}}$.

Given a probability set $\mathscr{P}_{\mathbf{x}}$, let $\hat{Y}_{\ell_H}$ be the maximal set of vectors that would be obtained using Eq. (4). The next proposition shows that $\hat{Y}_{\ell_H}$ can be outer-approximated using the marginals of the cautious estimate $\mathscr{P}_{\mathbf{x}}$, in contrast with the precise case.

Proposition 1. *Let $\mathscr{P}_{\mathbf{x}}$ be our estimate, then the imprecise vector $\hat{Y}^*$ such that*

$$\hat{Y}^*_j = \begin{cases} 1 & \text{if } \underline{P}(y_j = 1) > 1/2 \\ 0 & \text{if } \underline{P}(y_j = 0) > 1/2 \\ * & \text{else, i.e. } \underline{P}(y_j = 1) \leq 1/2 \leq \overline{P}(y_j = 1) \end{cases} \qquad \text{for } j = 1, \ldots, m$$

is an outer approximation of $\hat{Y}_{\ell_H}$, in the sense that $\hat{Y}_{\ell_H} \subseteq \hat{Y}^*$.

Proof. Consider a given $j \in \{1, \ldots, m\}$ and two alternatives $\hat{y}$ and $\hat{y}'$ such that $\hat{y}_j = 1 \neq \hat{y}'_j$ and $\hat{y}_i = \hat{y}'_i$ for any $i \neq j$. Then for any y such that $y_j = 1$ we have

$$\begin{aligned} \ell_H(\hat{y}', y) - \ell_H(\hat{y}, y) &= \left(\sum_{k \neq j} \mathbf{1}_{(\hat{y}'_k \neq y_k)} + \mathbf{1}_{(\hat{y}'_j \neq y_j)} \right) - \left(\sum_{k \neq j} \mathbf{1}_{(\hat{y}_k \neq y_k)} + \mathbf{1}_{(\hat{y}_j \neq y_j)} \right) \\ &= \mathbf{1}_{(\hat{y}'_j = 0)} - \mathbf{1}_{(\hat{y}_j = 0)} = 1, \end{aligned}$$

and for any y such that $y_j = 0$ we have

$$\begin{aligned}\ell_H(\hat{y}', y) - \ell_H(\hat{y}, y) &= \left(\sum_{k\neq j} \mathbf{1}_{(\hat{y}'_k \neq y_k)} + \mathbf{1}_{(\hat{y}'_j \neq y_j)}\right) - \left(\sum_{k\neq j} \mathbf{1}_{(\hat{y}_k \neq y_k)} + \mathbf{1}_{(\hat{y}_j \neq y_j)}\right) \\ &= \mathbf{1}_{(\hat{y}'_j = 1)} - \mathbf{1}_{(\hat{y}_j = 0)} = -1.\end{aligned}$$

We therefore have $(\ell_H(\hat{y}', \cdot) - \ell_H(\hat{y}, \cdot) + 1)/2 = \mathbf{1}_{(y_j = 1)}$, hence

$$\begin{aligned}\underline{P}(y_j = 1) &= \underline{\mathbb{E}}\left(\frac{\ell_H(\hat{y}', \cdot) - \ell_H(\hat{y}, \cdot) + 1}{2}\right) \\ &= \frac{1}{2}\underline{\mathbb{E}}\left(\ell_H(\hat{y}', \cdot) - \ell_H(\hat{y}, \cdot)\right) + \frac{1}{2}\end{aligned}$$

the last equality coming from scale and translation invariance. Hence $\underline{\mathbb{E}}(\ell_H(\hat{y}', \cdot) - \ell_H(\hat{y}, \cdot)) > 0$ if and only if $\underline{P}(y_j = 1) > 1/2$. This means that, if $\underline{P}(y_j = 1) > 1/2$, any vector $\hat{y}'$ with $\hat{y}'_j = 0$ is dominated (in the sense of Eq. (3)) by the vector $\hat{y}$ where only the j-th element is modified, hence no vector with $\hat{y}'_j = 0$ is in the maximal set $\hat{Y}_{\ell_H}$. The proof showing that if $\underline{P}(y_j = 0) > 1/2$, then no vector with $\hat{y}'_j = 1$ is in the maximal set is similar. ■

We now provide an example showing that the inclusion can be strict in general.

Example 2. Consider the 2 label case $\Lambda = \{\lambda_1, \lambda_2\}$ with the following constraints:

$$\begin{gathered}0.4 \leq P(y_1 = 1) = P(\{[1\,0]\}) + P(\{[1\,1]\}) \leq 0.6 \\ 0.9\,(P(\{[1\,0]\}) + P(\{[1\,1]\})) = P(\{[1\,0]\}) \\ 0.84\,(P(\{[0\,1]\}) + P(\{[0\,0]\})) = P(\{[0\,1]\})\end{gathered}$$

These constraints describe a convex set $\mathscr{P}$, whose extreme points (obtained by saturating the first inequality one way or another) are summarized in Table 2. The first constraints induces that $\underline{P}(y_1 = 1) = 0.4$ and $\overline{P}(y_1 = 0) = 0.6$, while the bounds $\underline{P}(y_2 = 1) = 0.396, \overline{P}(y_2 = 1) = 0.544$, are reached by the extreme distributions $P([1\,1]) = 0.06$, $P([0\,1]) = 0.336$ and $P([1\,1]) = 0.04$, $P([0\,1]) = 0.504$, respectively. Given these bounds, we have that $\hat{Y}^* = [* *]$ corresponds to the complete space $\mathscr{Y}$ (i.e., the empty prediction). Yet we have that

$$\underline{\mathbb{E}}(\ell_H([1\,1], \cdot) - \ell_H([0\,0], \cdot)) = 0.0008 \geq 0$$

also obtained with the distribution $P([1\,1]) = 0.06$, $P([0\,0]) = 0.064$. This means that the vector $[0\,0]$ is not in the maximal set $\hat{Y}_{\ell_H}$, while it is included in $\hat{Y}^*$.

Proposition 1 shows that we can rely on marginal information to provide an outer-approximation of $\hat{Y}_{\ell_H}$ that is efficient to compute, as it requires to compute $2m$ values, which are to be compared to the 2^{2m} usually required to assess $\hat{Y}_{\ell_H}$. It also indicates that extensions of the binary relevance approach are well adapted to provide partial predictions from credal sets when considering the Hamming loss, and that in this case global models integrating label dependencies are not necessary, thus saving a lot of heavy computations.

$P(\{[0\,0]\})$	$P(\{[1\,0]\})$	$P(\{[0\,1]\})$	$P(\{[1\,1]\})$
0.096	0.36	0.504	0.04
0.064	0.54	0.336	0.06

Table 2. Extreme points of $\mathscr{P}$ of Example 2

4 First experimentations

In this section, we provide first experimentations illustrating the effect of making partial predictions with a decreasing amount of information. These experiments illustrate that such partial predictions may indeed improve the correctness of predictions.

4.1 Evaluation

Usual loss functions such as Eq. (5) are based on complete predictions. When making partial predictions, such loss functions need to be adapted. This can be done, for instance, by decomposing it into two components [3], one measuring the accuracy or correctness of the made prediction, the other measuring its completeness.

If the partial prediction is an incomplete vector such as $\hat{Y}^*$, then Hamming loss can be easily split into these two components. Given the prediction $\hat{Y}^*$ characterized by subsets $\underline{L},\overline{L}$, let us denote $Q = \Lambda \setminus (\underline{L} \cap \overline{L})$ the set of predicted labels (i.e., labels such that $\hat{Y}^*_j = 1$ or $\hat{Y}^*_j = 0$). Then, if the observed set is y, we define correctness (CR) and completeness (CP) as

$$CR(\hat{Y}^*, y) = \frac{1}{|Q|} \sum_{\lambda_i \in Q} \mathbf{1}_{(\hat{y}_i \neq y_i)} \; ; \tag{6}$$

$$CP(\hat{Y}^*, y) = \frac{|Q|}{m}. \tag{7}$$

when predicting complete vectors, then $CP = 1$ and CR equals the Hamming loss (5). When predicting the empty vector, then $CP = 0$ and by convention $CR = 1$.

4.2 Method

The method we used was to apply, label-wise, the k-nn method using lower probabilities introduced in [6] (in which details can be found). This means that from an initial training data set $\mathscr{D}$, m data sets $\mathscr{D}_j$ corresponding to binary classification problems are built, this decomposition being illustrated in Figure 1. Given an instance $\mathbf{x}$, the result of the k-nn method on data set $\mathscr{D}_j$ provides an estimate of $[\underline{P}(y_j = 1), \overline{P}(y_j = 1)]$ and by duality an estimate of $\underline{P}(y_j = 0) = 1 - \overline{P}(y_j = 1)$ and $\overline{P}(y_j = 0) = 1 - \underline{P}(y_j = 1)$.

The method also automatically takes account of missing label information, and treat such missing data in a conservative way, considering them as completely vacuous information (that is, we treat them as non-MAR variables [16]).

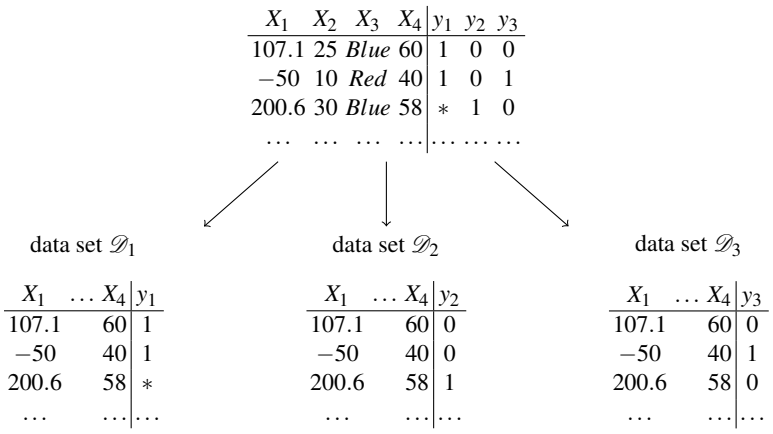

Fig. 1. Label-wise decomposition of data set $\mathscr{D}$

4.3 Results

In the experiments, the parameters of the k-nn algorithm were set to $\beta = 0.75$ and $\varepsilon_0 = 0.99$, so that results obtained when fixing the number k of neighbors to 1 display a sufficient completeness. ε_0 settles the initial imprecision, while β determines how much imprecision increases with distance (details about the role of these parameters can be found in [6]). We ran experiments on well-known multilabel data sets having real-valued features. Their characteristics are summarized in Table 3. For each of them, we ran a 10-fold cross validation with the number k of neighbors varying from 1 to 5, and with various percentages of missing labels in the training data set (0%, 20% and 40%). Varying k in the algorithm allows us to control the completeness of the prediction: the higher k is, the more imprecise become the estimations.

Name	# Features	# Labels	# Instances
emotion	72	6	593
scene	294	6	2407
yeast	103	14	2417
CAL500	68	174	502

Table 3. Multilabel data sets summary

The results of the experiment are displayed in Figure 2. From this figure, two main conclusions can be drawn: on the used data sets, allowing for partial predictions (here, by increasing the number k of neighbours) systematically improve the correctness, and missing labels only influence the completeness of the predictions, not the correctness of the results. This latter fact, however, may be due to the learning method. How fast

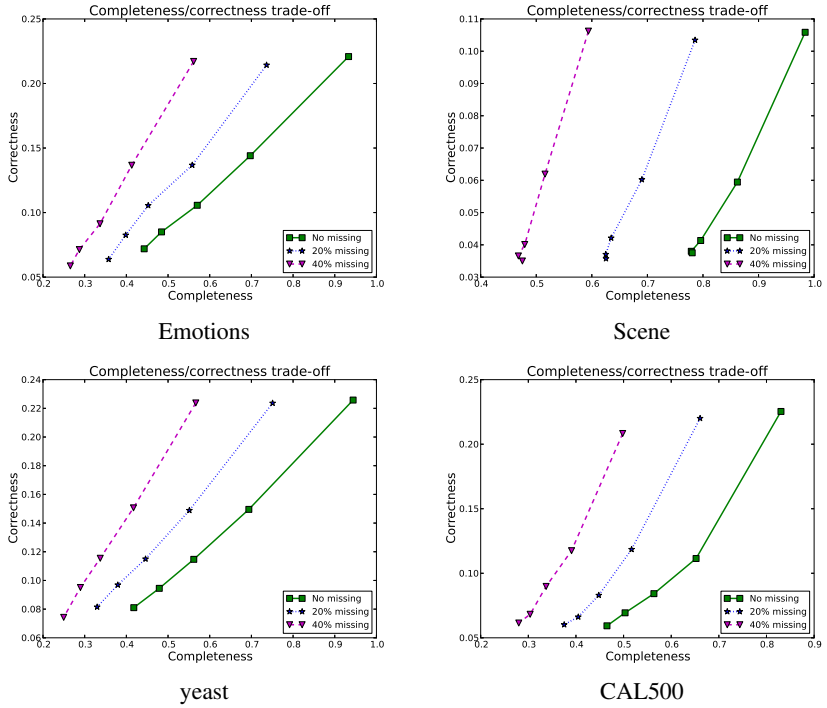

Fig. 2. Experimental results

completeness decreases with the number of neighbors, however, clearly depends on the data set.

5 Conclusions

Producing sets of optimal predictions in the multilabel setting when uncertainty is modeled by convex probability sets is computationally hard. The main contribution of this paper was to show that when using the Hamming loss, such sets can be easily outer-approximated by focusing only on the marginal probability bounds of each label being relevant. This makes both computation and learning issues easier, as one can focus on estimating such marginals (instead of the whole joint model). We can consider that as an important result, as it shows that imprecise probabilistic approaches are computationally affordable (at least under some conditions).

We also made some first preliminary experiments indicating the interest of producing such partial predictions, showing that making more cautious predictions lead to more correct predictions. In the future, we intend to make similar studies for other well-known loss functions, such as the ranking loss. We also intend to make further the experiments, i.e., to compare this approach with other methods, or to empirically assess (for small values of m) the quality of the made approximation.

Acknowledgements

Work carried out in the framework of the Labex MS2T, funded by the French Government, through the National Agency for Research (Reference ANR-11-IDEX-0004-02)

References

1. P. Bartlett and M. Wegkamp. Classification with a reject option using a hinge loss. *The Journal of Machine Learning Research*, 9:1823–1840, 2008.
2. W. Cheng, E. Hüllermeier, W. Waegeman, and V. Welker. Label ranking with partial abstention based on thresholded probabilistic models. In *Advances in Neural Information Processing Systems 25 (NIPS-12)*, pages 2510–2518, 2012.
3. W. Cheng, M. Rademaker, B. De Baets, and E. Hüllermeier. Predicting partial orders: ranking with abstention. *Machine Learning and Knowledge Discovery in Databases*, pages 215–230, 2010.
4. G. Corani, A. Antonucci, and M. Zaffalon. Bayesian networks with imprecise probabilities: Theory and application to classification. *Data Mining: Foundations and Intelligent Paradigms*, pages 49–93, 2012.
5. K. Dembczynski, W. Waegeman, W. Cheng, and E. Hüllermeier. On label dependence and loss minimization in multi-label classification. *Machine Learning*, 88(1-2):5–45, 2012.
6. S. Destercke. A k-nearest neighbours method based on imprecise probabilities. *Soft Comput.*, 16(5):833–844, 2012.
7. J. Fürnkranz, E. Hüllermeier, E. Loza Mencía, and K. Brinker. Multilabel classification via calibrated label ranking. *Machine Learning*, 73(2):133–153, 2008.
8. I. Levi. *The Enterprise of Knowledge*. MIT Press, London, 1980.
9. Y.-Y. Sun, Y. Zhang, and Z.-H. Zhou. Multi-label learning with weak label. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.
10. F. Tian and X. Shen. Image annotation with weak labels. In *Web-Age Information Management*, pages 375–380. Springer, 2013.
11. M. Troffaes. Decision making under uncertainty using imprecise probabilities. *Int. J. of Approximate Reasoning*, 45:17–29, 2007.
12. K. Trohidis, G. Tsoumakas, G. Kalliris, and I. P. Vlahavas. Multi-label classification of music into emotions. In *ISMIR*, volume 8, pages 325–330, 2008.
13. G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007.
14. G. Yu, C. Domeniconi, H. Rangwala, and G. Zhang. Protein function prediction using dependence maximization. In *Machine Learning and Knowledge Discovery in Databases*, pages 574–589. Springer, 2013.
15. M. Zaffalon. Exact credal treatment of missing data. *Journal of Statistical Planning and Inference*, 105(1):105–122, 2002.
16. M. Zaffalon and E. Miranda. Conservative inference rule for uncertain reasoning under incompleteness. *Journal of Artificial Intelligence Research*, 34(2):757, 2009.