Graded Multilabel Classification: The Ordinal Case*

Weiwei Cheng, Krzysztof Dembczynski, and Eyke Hüllermeier
Marburg University
Hans-Meerwein-Str., 35032 Marburg, Germany
{cheng, dembczynski, eyke }@informatik.uni-marburg.de

Abstract

We propose a generalization of multilabel classification that we refer to as *graded multilabel classification*. The key idea is that, instead of requesting a yes-no answer to the question of class membership or, say, relevance of a class label for an instance, we allow for a *graded membership* of an instance, measured on an ordinal scale of membership degrees. This extension is motivated by practical applications in which a graded or partial class membership is natural. Apart from introducing the basic setting, we propose two general strategies for reducing graded multilabel problems to conventional (multilabel) classification problems. Moreover, we address the question of how to extend performance metrics commonly used in multilabel classification to the graded setting, and present first experimental results.

1 Introduction

Problems of *multilabel classification* (MLC), in which an instance may belong to several classes simultaneously or, say, in which more than one label can be attached to a single instance, are ubiquitous in everyday life: At IMDb, a movie can be categorized as *action*, *crime*, and *thriller*, a CNN news report can be tagged as *people* and *political* at the same time, etc. Correspondingly, MLC has received increasing attention in machine learning in recent years.

In this paper, we propose a generalization of MLC that we shall refer to as *graded multilabel classification* (GMLC). The key idea is that, instead of requesting a yes-no answer to the question of class membership or, say, relevance of a class label for an instance, we allow an instance to belong to a class *to a certain degree*. In other words, we allow for graded class membership in the sense of *fuzzy set theory* [Zadeh, 1965]. In fact, there are many applications for which this extension seems to make perfect sense. In the case of movie genres, for example, it is not always easy to say whether or not a movie belongs to the category *action*, and there are definitely examples which can be considered as "almost action" or "somewhat action". Another obvious example comes from one of the benchmark data sets in MLC, namely the *emotions* data [Trohidis *et al.*, 2008]. Here, the problem is to label a song according to the Tellegen-Watson-Clark model of mood: amazed-surprised, happy-pleased, relaxing-clam, quiet-still, sad-lonely, and angry-aggressive.

It is important to emphasize that the relevance of a label is indeed *gradual* in the sense of fuzzy logic and not *uncertain* in the sense of probability theory. The latter would mean that, e.g., a song is either relaxing or it is not—one is only uncertain about which of these two exclusive alternatives is correct. As opposed to this, gradualness is caused by the vagueness of categories like "relaxing song" and "action movie", and means that one does not have to fully agree on one of the alternatives. Instead, one can say that a song is somewhere in-between (and can be certain about this).

As will be explained in more detail later on, our idea is to replace simple "yes" or "no" labels by membership degrees taken from a finite ordered scale such as

$$M = \{\texttt{not at all}, \texttt{somewhat}, \texttt{almost}, \texttt{fully}\}. \quad (1)$$

Admittedly, graded multilabel data sets of that kind are not yet widely available. We believe, however, that this is a kind of hen and egg problem: As long as there are no methods for learning from graded multilabel data, new data sets will be created in the common way, possibly forcing people to give a "yes" or "no" answer even when they are hesitating.

The rest of this paper is organized as follows: The problem of multilabel classification is introduced in a more formal way in Section 2. In Section 3, we propose our graded generalization of MLC and, moreover, outline two different strategies for reducing GMLC problems to conventional (multilabel) classification problems. In Section 4, we address the question of how to extend MLC evaluation metrics from the conventional to the graded setting. Finally, Section 5 presents some first experimental results.

2 Multilabel Classification

Let $\mathbb{X}$ denote an instance space and let $\mathcal{L} = \{\lambda_1, \lambda_2, \ldots, \lambda_m\}$ 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 $H : \mathbb{X} \longrightarrow 2^{\mathcal{L}}$ that generalizes well beyond these observations in the sense of minimizing the expected prediction loss with respect to a specific loss function; examples of commonly used loss functions include the *subset*

*This paper has been presented at International Conference on Machine Learning, Haifa, Israel, 2010.

zero-one loss, which is 0 if $H(\boldsymbol{x}) = L_{\boldsymbol{x}}$ and 1 otherwise, and the *Hamming loss* that computes the percentage of labels whose relevance is incorrectly predicted:

$$\mathrm{E}_H(H(\boldsymbol{x}), L_{\boldsymbol{x}}) = \frac{1}{|\mathcal{L}|}\big|H(\boldsymbol{x}) \,\Delta\, L_{\boldsymbol{x}}\big|, \tag{2}$$

where Δ is the symmetric difference between sets.

An MLC problem 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* (LP) 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; obviously, this number is $2^{|\mathcal{L}|}$ (or $2^{|\mathcal{L}|} - 1$ if the empty set is excluded as a prediction). This is the reason why LP typically works well if the original label set $\mathcal{L}$ is small but quickly deteriorates for larger label sets [Tsoumakas and Vlahavas, 2007].

Another way of reducing multilabel to conventional classification is offered by the *binary relevance* (BR) approach. Here, a separate binary classifier H_i is trained for each label $\lambda_i \in \mathcal{L}$, reducing the supervision to information about the presence or absence of this label while ignoring the other ones. 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 its ignorance of correlations and interdependencies between labels.

Many approaches to MLC learn a multilabel classifier H in an indirect way via a scoring function $f : \mathbb{X} \times \mathcal{L} \longrightarrow \mathbb{R}$ that assigns a real number to each instance/label combination. The idea is that a score $f(\boldsymbol{x}, \lambda)$ is in direct correspondence with the probability that λ is relevant for $\boldsymbol{x}$. Given a scoring function of this type, multilabel prediction can be realized via thresholding:

$$H(\boldsymbol{x}) \;=\; \{\lambda \in \mathcal{L} \,|\, f(\boldsymbol{x}, \lambda) \geq t\} \;\;,$$

where $t \in \mathbb{R}$ is a threshold. As a byproduct, a scoring function offers the possibility to produce a ranking (weak order) $\succeq_{\boldsymbol{x}}$ of the class labels, simply by sorting them according to their score:

$$\lambda_i \succeq_{\boldsymbol{x}} \lambda_j \;\Leftrightarrow\; f(\boldsymbol{x}, \lambda_i) \geq f(\boldsymbol{x}, \lambda_j) \;\;. \tag{3}$$

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; an example is the *rank loss* which computes the average fraction of label pairs that are not correctly ordered:

$$\mathrm{E}_R(f, L_{\boldsymbol{x}}) = \frac{\sum_{(\lambda,\lambda') \in L_{\boldsymbol{x}} \times \overline{L}_{\boldsymbol{x}}} S(f(\boldsymbol{x}, \lambda), f(\boldsymbol{x}, \lambda'))}{|L_{\boldsymbol{x}}| \times |\overline{L}_{\boldsymbol{x}}|},$$

where $\overline{L}_{\boldsymbol{x}} = \mathcal{L} \setminus L_{\boldsymbol{x}}$ is the set of irrelevant labels and $S(u, v) = 1$ if $u < v$, $= 1/2$ if $u = v$, and $= 0$ if $u > v$. The idea to solve both problems simultaneously, ranking and MLC, has recently been addressed in [Fürnkranz *et al.*, 2008]: A *calibrated ranking* is a ranking with a "zero point" separating a positive (relevant) part from a negative (irrelevant) one.

3 Graded Multilabel Classification

Generalizing the above setting of multilabel classification, we now assume that each instance $\boldsymbol{x} \in \mathbb{X}$ can belong to each class $\lambda \in \mathcal{L}$ *to a certain degree*. In other words, the set $L_{\boldsymbol{x}}$ of relevant labels is now a fuzzy subset of $\mathcal{L}$. This fuzzy set is characterized by a membership function, namely an $\mathcal{L} \longrightarrow M$ mapping, where M is the set of graded membership degrees. For notational simplicity, we shall not distinguish between the fuzzy set $L_{\boldsymbol{x}}$ and its membership function, and denote by $L_{\boldsymbol{x}}(\lambda)$ the degree of membership of the label $\lambda \in \mathcal{L}$ in the fuzzy set $L_{\boldsymbol{x}}$.

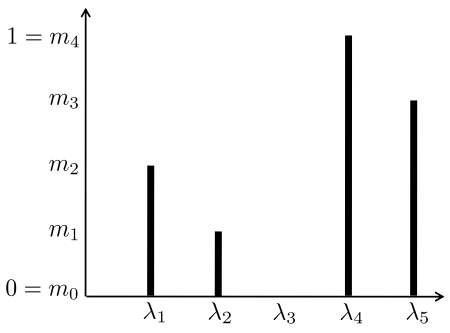

Figure 1: Vertical reduction, viz. prediction of membership degree (ordinate) for each label (abscissa).

In fuzzy set theory, the set of membership degrees is supposed to form a complete lattice and is normally taken as the unit interval (i.e., $M = [0, 1]$ endowed with the standard order). Here, however, we prefer an *ordinal scale* of membership degrees, that is, a finite ordered set of membership degrees such as (1). More generally, we assume that $M = \{m_0, m_1, \ldots, m_k\}$, where $m_0 < m_1 < \ldots < m_k$ (and $m_0 = 0$ and $m_k = 1$ have the special meaning of zero and full membership). In the context of multilabel classification, an ordinal membership scale is arguably more convenient from a practical point of view, especially with regard to data acquisition. In fact, people often prefer to give ratings on an ordinal scale like (1) instead of choosing precise numbers on a cardinal scale.

The goal, now, is to learn a mapping $H : \mathbb{X} \longrightarrow \mathcal{F}(\mathcal{L})$, where $\mathcal{F}(\mathcal{L})$ is the class of fuzzy subsets of $\mathcal{L}$ (with membership degrees in M). Following the general idea of *reduction* [Balcan *et al.*, 2008], we seek to make GMLC problems amenable to conventional multilabel methods via suitable transformations. There are two more or less obvious possibilities to reduce graded multilabel classification to conventional (multilabel) classification. In agreement with the distinction between the "vertical" description of a fuzzy subset F of a set U (through the membership function, i.e., by specifying the degree of membership $F(u)$ for each element $u \in U$) and the "horizontal" description (via level cuts $[F]_\alpha = \{u \in U \,|\, F(u) \geq \alpha\}$), we distinguish between a vertical and a horizontal reduction.

3.1 Vertical Reduction

Recall the *binary relevance* approach to conventional MLC: For each label $\lambda_i \in \mathcal{L}$, a separate binary classifier H_i is trained to predict whether this label is relevant ($H_i(\boldsymbol{x}) = 1$) or not ($H_i(\boldsymbol{x}) = 0$) for a query instance $\boldsymbol{x} \in \mathbb{X}$. Generalizing this approach to GMLC, the idea is to induce a classifier

$$H_i : \mathbb{X} \longrightarrow M \tag{4}$$

for each label λ_i. For each query instance $\boldsymbol{x} \in \mathbb{X}$, this classifier is supposed to predict the degree of membership of λ_i in the fuzzy set of labels $L_{\boldsymbol{x}}$. Instead of a binary classification problem, as in MLC, each classifier H_i is now

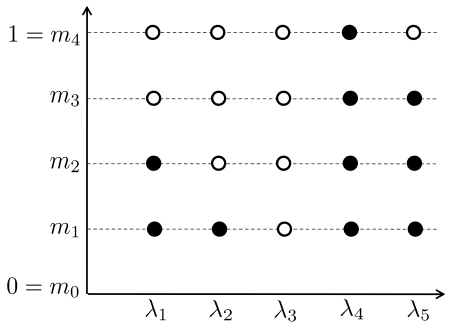

Figure 2: Horizontal reduction, viz. prediction of a subset of labels (indicated by black circles) on each level.

solving a multi-class problem. Since the target space M has an ordinal structure, these problems are *ordinal classification problems*. In other words, the vertical reduction of a GMLC problem eventually leads to solving a set of m (non-independent) ordinal classification problems; see Fig. 1 for an illustration.

Just like simple binary problems, ordinal classification problems are often solved indirectly via "scoring plus thresholding": First, a scoring function $f(\cdot)$ is learned, and k thresholds $t_1, \ldots, t_k$ are determined; then, for an instance $\boldsymbol{x}$, the i-th class is predicted if $f(\boldsymbol{x}, \lambda_i)$ is between t_{i-1} and t_i. Of course, if classifiers (4) are learned in this way, i.e., by inducing a scoring function $f(\cdot, \lambda_i)$ for each label λ_i, then these scoring functions can also be used to predict a ranking (3).

3.2 Horizontal Reduction

From fuzzy set theory, it is well-known that a fuzzy set F can be represented "horizontally" in terms of its level-cuts. This representation suggests another decomposition of a GMLC problem: For each level $\alpha \in \{m_1, m_2, \ldots, m_k\}$, learn the mapping

$$H^{(\alpha)} : \mathbb{X} \longrightarrow 2^M, \boldsymbol{x} \mapsto [L_{\boldsymbol{x}}]_\alpha \ . \tag{5}$$

Obviously, each of these problems is a *standard* MLC problem, since the level-cuts $[L_{\boldsymbol{x}}]_\alpha$ are standard subsets of the label set $\mathcal{L}$. Thus, the horizontal reduction comes down to solving k standard MLC problems; see Fig. 2 for an illustration.

It is worth mentioning that this decomposition comes with a special challenge. In fact, since level-cuts are nested in the sense that $[F]_\alpha \subset [F]_\beta$ for $\beta < \alpha$, the k MLC problems are not independent of each other. Instead, the predictions should be *monotone* in the sense that

$$(H^{(m_j)}(\boldsymbol{x}) = 1) \Rightarrow (H^{(m_{j-1})}(\boldsymbol{x}) = 1) \tag{6}$$

for all $j \in \{2, \ldots, k\}$. Thus, whenever a label λ_i is predicted to be in the m_j-cut of the fuzzy label set $L_{\boldsymbol{x}}$ associated with $\boldsymbol{x}$, it must also be in all lower level-cuts. Satisfying this requirement is a non-trivial problem. In particular, (6) will normally not be guaranteed when solving the k problems independently of each other.

Once an ensemble of k multilabel classifiers $H^{(m_1)}, \ldots, H^{(m_k)}$ has been trained, predictions can be obtained as follows:

$$H(\boldsymbol{x})(\lambda) \ = \ \max\{\, m_i \in M \mid \lambda \in H^{(m_i)}(\boldsymbol{x}) \,\} \tag{7}$$

Thus, the degree of membership of a label $\lambda \in \mathcal{L}$ in the predicted fuzzy set of labels associated with $\boldsymbol{x}$ is given by the maximum degree $m_i \in M$ for which λ is still in the predicted m_i-cut of this set.

The prediction of a ranking (3) is arguably less obvious in the case of the horizontal decomposition. Suppose that $f^{(m_1)}, \ldots, f^{(m_k)}$ are scoring functions trained on the k level cuts, using a conventional MLC method. As a counterpart to the monotonicity condition (6), we should require

$$f^{(m_1)}(\boldsymbol{x}, \lambda) \geq f^{(m_2)}(\boldsymbol{x}, \lambda) \geq \ldots \geq f^{(m_k)}(\boldsymbol{x}, \lambda) \tag{8}$$

for all $\boldsymbol{x} \in \mathbb{X}$ and $\lambda \in \mathcal{L}$. In fact, interpreting $f^{(m_i)}(\boldsymbol{x}, \lambda)$ as a measure of how likely λ is a relevant label on level m_i, this condition follows naturally from $[L_{\boldsymbol{x}}]_{m_1} \supset [L_{\boldsymbol{x}}]_{m_2} \supset \ldots \supset [L_{\boldsymbol{x}}]_{m_k}$. On each level m_i, the function $f^{(m_i)}(\boldsymbol{x}, \cdot)$ induces a ranking $\succeq_{\boldsymbol{x}}^{(m_i)}$ via (3), however, the identity $\succeq_{\boldsymbol{x}}^{(m_i)} \equiv \succeq_{\boldsymbol{x}}^{(m_j)}$ is of course not guaranteed; that is, $\succeq_{\boldsymbol{x}}^{(m_i)}$ may differ from $\succeq_{\boldsymbol{x}}^{(m_j)}$ for $1 \leq i \neq j \leq k$.

To obtain a global ranking, the level-wise rankings $\succeq_{\boldsymbol{x}}^{(m_i)}$ need to be aggregated into a single one. To this end, we propose to score a label λ by

$$f(\boldsymbol{x}, \lambda) \ = \ \sum_{i=1}^{k} f^{(m_i)}(\boldsymbol{x}, \lambda) \ . \tag{9}$$

This aggregation is especially reasonable if the scores $f^{(m_i)}(\boldsymbol{x}, \lambda)$ can be interpreted as probabilities of relevance $\mathbf{P}(\lambda \in [L_{\boldsymbol{x}}]_{m_i})$. Then, $f(\boldsymbol{x}, \lambda)$ simply corresponds to the *expected level* of $\boldsymbol{x}$, since

$$\sum_{i=1}^{k} f^{(m_i)}(\boldsymbol{x}, \lambda) = \sum_{i=1}^{k} \mathbf{P}(\lambda \in [L_{\boldsymbol{x}}]_{m_i}) =$$

$$= \sum_{i=1}^{k} \mathbf{P}(L_{\boldsymbol{x}}(\lambda) \geq m_i) = \sum_{i=1}^{k} i \cdot \mathbf{P}(L_{\boldsymbol{x}}(\lambda) = m_i)$$

Note, however, that we simply equated the levels m_i with the numbers i in this derivation, i.e., the ordinal scale $\mathcal{L}$ was implicitly embedded in a numerical scale by the mapping $m_i \mapsto i$ (on $\mathcal{L}$ itself, an averaging operation of this kind is not even defined). Despite being critical from a theoretical point of view, this embedding is often used in ordinal classification, for example when computing the absolute error $\mathrm{AE}(m_i, m_j) = |i - j|$ as a loss function [Lin and Li, 2007]. Interestingly, the absolute error is minimized (in expectation) by the *median* and, moreover, this estimation is invariant toward rescaling [Berger, 1985]. Thus, it does actually not depend on the concrete embedding chosen. Seen from this point of view, the median appears to be a theoretically more solid score than the mean value (9). However, it produces many ties, which is disadvantageous from a ranking point of view. This problem is avoided by (9), which can be seen an approximation of the median that breaks ties in a reasonable way.

3.3 Combination of Both Reductions

As mentioned above, the binary relevance approach is a standard (meta-)technique for solving MLC problems. Consequently, it can also be applied to each problem (5) produced by the horizontal reduction. Since BR can again be seen as a "vertical" decomposition of a regular MLC problem, one thus obtains a combination of horizontal and vertical decomposition: first horizontal, then vertical.

Likewise, the two types of reduction can be combined the other way around, first vertical and then horizontal. This is done by solving the ordinal classification problems

produced by the vertical reduction by means of a "horizontal" decomposition, namely a meta-technique that has been proposed by [Frank and Hall, 2001]: Given an ordered set of class labels $M = \{m_0, m_1, \ldots, m_k\}$, the idea is to train k binary classifiers. The i-th classifier considers the instances with label $m_0, \ldots, m_{i-1}$ as positive and those with label $m_i, \ldots, m_k$ as negative.

Interestingly, both combinations eventually coincide in the sense of ending up with the same binary classification problems. Roughly speaking, a single binary problem is solved for each label/level combination $(\lambda_i, m_j) \in \mathcal{L} \times M$ (each circle in the picture in Fig. 2), namely the problem to decide whether $L_{\boldsymbol{x}}(\lambda_i) \leq m_j$ or $L_{\boldsymbol{x}}(\lambda_i) > m_j$. Any difference between the two approaches is then due to different ways of aggregating the predictions of the binary classifiers. In principle, however, such differences can only occur in the case of inconsistencies, i.e., if the monotonicity condition (6) is violated.

3.4 Generalizing IBLR-ML

Our discussion so far has been restricted to meta-techniques for reducing GMLC to MLC problems, without looking at concrete methods. Nevertheless, there are several methods that can be generalized immediately from the binary to the gradual case. As an example, we mention the IBLR-ML method that will also be used in our experiments later on. This method, which was recently proposed in [Cheng and Hüllermeier, 2009], combines instance-based learning with logistic regression and again trains one classifier H_i for each label. For the i-th label λ_i, this classifier is derived from the logistic regression equation

$$\log\left(\frac{\pi_0^{(i)}}{1-\pi_0^{(i)}}\right) \;=\; \omega_0^{(i)} + \sum_{j=1}^{m} \gamma_j^{(i)} \cdot \omega_{+j}^{(i)}(\boldsymbol{x}_0) \;\;, \quad (10)$$

where $\pi_0^{(i)}$ denotes the (posterior) probability that λ_i is relevant for $\boldsymbol{x}_0$, and

$$\omega_{+j}^{(i)}(\boldsymbol{x}_0) = \sum_{\boldsymbol{x} \in \mathcal{N}(\boldsymbol{x}_0)} \kappa(\boldsymbol{x}_0, \boldsymbol{x}) \cdot y_j(\boldsymbol{x}) \quad (11)$$

is a summary of the presence of the j-th label λ_j in the neighborhood of $\boldsymbol{x}_0$; here, κ is a kernel function, such as the (data-dependent) "KNN kernel" $\kappa(\boldsymbol{x}_0, \boldsymbol{x}_i) = 1$ if $\boldsymbol{x}_i \in \mathcal{N}_k(\boldsymbol{x}_0)$ and $= 0$ otherwise, where $\mathcal{N}_k(\boldsymbol{x}_0)$ is the set of k nearest neighbors of $\boldsymbol{x}_0$. Moreover, $y_j(\boldsymbol{x}) = +1$ if λ_j is present (relevant) for the neighbor $\boldsymbol{x}$, and $y_j(\boldsymbol{x}) = -1$ in case it is absent (non-relevant). Obviously, this approach is able to capture interdependencies between class labels: The estimated coefficient $\gamma_j^{(i)}$ indicates to what extent the relevance of label λ_i is influenced by the relevance of λ_j. A value $\gamma_j^{(i)} > 0$ means that the presence of λ_j makes the relevance of λ_i more likely, i.e., there is a positive correlation. Correspondingly, a negative coefficient would indicate a negative correlation. Given a query instance $\boldsymbol{x}_0$, a multilabel prediction is made on the basis of the predicted posterior probabilities of relevance: $H(\boldsymbol{x}_0) = \{\lambda_i \in \mathcal{L} \,|\, \pi_0^{(i)} > 1/2\}$.

This approach can be generalized to the GMLC setting using both the horizontal and the vertical reduction. The vertical reduction leads to solving an ordinal instead of a binary logistic regression problem for each label, while the horizontal reduction comes down to solving the following k multilabel problems ($r = 1, \ldots, k$):

$$\log\left(\frac{\pi_0^{(i,r)}}{1-\pi_0^{(i,r)}}\right) \;=\; \omega_0^{(i,r)} + \sum_{j=1}^{m} \gamma_j^{(i,r)} \omega_{+j}^{(i,r)}(\boldsymbol{x}_0) \quad (12)$$

Recall, however, that these problems are not independent of each other. Solving them simultaneously so as to guarantee the monotonicity constraint (6) is an interesting but non-trivial task. In the experiments in Section 5, we therefore derived independent predictions and simply combined them by (7).

4 Loss Functions

As mentioned before, a number of different loss functions have already been proposed within the setting of MLC. In principle, all these functions can be generalized so as to make them applicable to the setting of GMLC. In this section, we propose extensions of some important and frequently used measures. Moreover, we address the question of how to handle these extensions in the context of the horizontal and vertical reduction technique, respectively.

4.1 Representation of Generalized Losses

To generalize the Hamming loss (2), it is necessary to replace the symmetric difference operator defined on sets, Δ, by the symmetric difference between two fuzzy sets. This can be done, for example, by averaging over the symmetric differences of the corresponding level-cuts, which in our case leads to

$$\mathrm{E}_H^*(H(\boldsymbol{x}), L_{\boldsymbol{x}}) = \frac{\sum_{i=1}^{k} \left|[H(\boldsymbol{x})]_{m_i} \,\Delta\, [L_{\boldsymbol{x}}]_{m_i}\right|}{k|\mathcal{L}|}. \quad (13)$$

Note that this "horizontal" computation can be replaced by an equivalent "vertical" one, namely

$$\mathrm{E}_H^*(H(\boldsymbol{x}), L_{\boldsymbol{x}}) = \frac{\sum_{i=1}^{|\mathcal{L}|} \mathrm{AE}(H(\boldsymbol{x})(\lambda_i), L_{\boldsymbol{x}}(\lambda_i))}{k|\mathcal{L}|}, \quad (14)$$

where $\mathrm{AE}(\cdot)$ is the absolute error of a predicted membership degree which, as mentioned above, is defined by $\mathrm{AE}(m_i, m_j) = |i - j|$. In other words, minimizing the symmetric difference level-wise is equivalent to minimizing the absolute error label-wise.

It is worth to mention that the existence of an equivalent horizontal and vertical representation of a loss function, like in the case of (13) and (14), is not self-evident. For example, replacing in (14) the absolute error on the ordinal scale M by the simple 0/1 loss leads to

$$\mathrm{E}_{0/1}^*(H(\boldsymbol{x}), L_{\boldsymbol{x}}) = \frac{1}{|\mathcal{L}|} \sum_{i=1}^{|\mathcal{L}|} \begin{cases} 0 & H(\boldsymbol{x})(\lambda_i) = L_{\boldsymbol{x}}(\lambda_i) \\ 1 & H(\boldsymbol{x})(\lambda_i) \neq L_{\boldsymbol{x}}(\lambda_i) \end{cases} .$$

Just like (14), this is a typical vertical expression of a loss function, that is, an expression of the form

$$\mathrm{A}\left(\{\ell\left(H(\boldsymbol{x})(\lambda_i), L_{\boldsymbol{x}}(\lambda_i)\right)\}_{i=1}^{|\mathcal{L}|}\right),$$

where $\ell(\cdot)$ is a loss defined on $\mathcal{L}$ and A is an aggregation operator. Interestingly, $\mathrm{E}_{0/1}^*$ does not have an equivalent horizontal representation. Thus, there is provably no loss function $L(\cdot)$ on $2^{\mathcal{L}}$ (and aggregation A) such that

$$\mathrm{E}_{0/1}^*(H(\boldsymbol{x}), L_{\boldsymbol{x}}) = \mathrm{A}\left(\{L\left([H(\boldsymbol{x})]_{m_i}, [L_{\boldsymbol{x}}]_{m_i}\right)\}_{i=1}^{k}\right).$$

This observation has an important implication. Namely, if the loss function to be minimized has a vertical but not a

horizontal representation, then a vertical decomposition of the learning problem is arguably more self-evident than a horizontal one, and vice versa. Strictly speaking, the non-existence of an equivalent representation does of course not exclude the existence of another loss function and aggregation operator producing the same predictions. Such alternatives, however, will normally be less obvious.

As an example of a loss function that lends itself to a horizontal representation, consider a variant of the Hamming loss based on the well-known Jaccard-index:

$$\mathrm{E}_J(H(\boldsymbol{x}), L_{\boldsymbol{x}}) = \frac{|H(\boldsymbol{x}) \cap L_{\boldsymbol{x}}|}{|H(\boldsymbol{x}) \cup L_{\boldsymbol{x}}|} \tag{15}$$

This variant avoids a certain disadvantage of the Hamming loss, which treats relevant and non-relevant labels in a symmetric way even though the former are typically less numerous than the latter, thereby producing a bias toward the prediction of non-relevance. A natural generalization of this measure is obtained by averaging (15) over the levels:

$$\mathrm{E}^*_J(H(\boldsymbol{x}), L_{\boldsymbol{x}}) = \frac{1}{k} \sum_{i=1}^{k} \frac{\left|[H(\boldsymbol{x})]_{m_i} \cap [L_{\boldsymbol{x}}]_{m_i}\right|}{\left|[H(\boldsymbol{x})]_{m_i} \cup [L_{\boldsymbol{x}}]_{m_i}\right|} \tag{16}$$

This extension, however, does not admit an equivalent vertical representation, which is plausible since the Jaccard-index is indeed a genuine set measure.

4.2 Rank Loss

The rank loss E_R can be generalized in a canonical way by the so-called C-index, which is commonly used as a measure of concordance in statistics [Gnen and Heller, 2005], and which is essentially equivalent to the pairwise ranking error introduced in [Herbrich *et al.*, 2000]:

$$\mathrm{E}^*_R(f, L_{\boldsymbol{x}}) = \frac{\sum_{i<j} \sum_{(\lambda,\lambda') \in M_i \times M_j} S(f(\boldsymbol{x}, \lambda), f(\boldsymbol{x}, \lambda'))}{\sum_{i<j} |M_i| \times |M_j|},$$

where $M_i = \{\lambda \in \mathcal{L} \,|\, L_{\boldsymbol{x}}(\lambda) = m_i\}$. As can be seen, the C-index is the fraction of labels that are correctly ordered by $f(\cdot)$: If label λ' has a higher degree of membership in $L_{\boldsymbol{x}}$ than λ, then the former should be ranked above the latter. It is also worth mentioning that the C-index has recently been proposed as a performance measure in the problem of *multipartite ranking* [Fürnkranz *et al.*, 2009], and indeed, the problem here can be considered as a problem of that kind when interpreting $\{M_0, M_1, \ldots, M_k\}$ as an ordered partition of the label set $\mathcal{L}$.

Other ranking losses proposed in the literature can be generalized, too. For example, the *one error* checks whether the top-ranked label is relevant or not:

$$\mathrm{E}_{1E}(f, L_{\boldsymbol{x}}) = \begin{cases} 0 & \arg\max_{\lambda \in \mathcal{L}} f(\boldsymbol{x}, \lambda) \in L_{\boldsymbol{x}} \\ 1 & \text{otherwise} \end{cases}$$

A natural generalization of this measure is obtained on the basis of the degree of membership of the top-ranked label in $L_{\boldsymbol{x}}$:

$$\mathrm{E}^*_{1E}(f, L_{\boldsymbol{x}}) = 1 - L_{\boldsymbol{x}}\left(\arg\max_{\lambda \in \mathcal{L}} f(\boldsymbol{x}, \lambda)\right) \quad .$$

5 Experimental Study

An experimental validation of the methods proposed in this paper is not at all straightforward. First, since we introduced a new machine learning problem, no benchmark data sets can be found so far. Essentially for the same reason, there are no existing methods to be used for comparison. The two reduction schemes proposed in Section 3, vertical and horizontal, are not easily comparable either, since these are meta-techniques using different types of base learners.

For these reasons, we decided to focus on another aspect, namely the general usefulness of the extended setting that we proposed in this paper. More specifically, our idea is to provide empirical evidence for the claim that allowing a user to label instances on a graded scale does provide useful extra information. In a sense, this claim is trivial if a prediction on a graded scale is eventually needed. For example, a reviewer recommendation (which can be seen as an estimation of the quality of a paper) on an ordered scale with labels such as "weak accept" and "strong accept" is normally more useful than just a "yes" or "no" answer to the question of acceptance.

However, we claim that *training* a learner on graded data can be useful even if only a *binary prediction* is eventually requested. Intuitively, this claim derives from the simple observation that graded data provides more information than binary data, which can be helpful, e.g., to determine proper decision boundaries.

5.1 Data

In light of the aforementioned lack of benchmark data, we used a data set from another research field, namely social psychology [Abele and Stief, 2004].[1] This data set, called BeLa-E, consists of 1930 instances and 50 attributes. Each instance corresponds to a graduate student. The first attribute is the sex of the student and the second one the age. Each of the other 48 attributes is a graded degree of importance of different properties of the future job, evaluated by the student on an ordinal scale with 5 levels ranging from 1 (completely unimportant) to 5 (very important). Examples of such properties include "reputation", "safety", "high income" and "friendly colleagues". Thus, every student was asked how important he or she considers these properties to be, and the student answered by assigning one of the aforementioned 5 levels.

On the basis of this data set, we generated (graded) MLC problems as follows: m of the above 48 attributes were randomly selected as the set of class labels, while all remaining $m - 48$ attributes plus the student's sex and age were taken as predictive features. The goal, then, is to train an MLC model that takes the features as input and produces a prediction of the relevance of the class labels as output.

Moreover, for every GMLC problem thus obtained, a binary version is produced by mimicking a student who is forced to answer either yes or no: The graded levels 1 and 2 are mapped to "No", the levels 4 and 5 are mapped to "Yes", and a coin is flipped for level 3.

5.2 Methods

As multilabel classifiers we used the IBLR-ML method outlined in Section 3.4 and, moreover, binary relevance learning with 10-nearest neighbor classification (BR-10NN) as base learner. Two types of learning are distinguished, binary and graded: In binary learning, the original data is first binarized as explained above (turning graded into 0/1 answers). Then, the multilabel classifier is trained on this data and used to make binary multilabel predictions. In graded learning, a GMLC classifier is trained

[1] The data set is available online at `http://www.uni-marburg.de/fb12/kebi/research`.

on the original (graded) data, using the horizontal reduction technique (for the BR learners automatically combined with the vertical reduction). The graded relevance predictions of these learners are then mapped to binary relevance degrees at the very end, using the same $M \longrightarrow \{0, 1\}$ mapping (randomized for label 3) as used in binary learning at the beginning. Eventually, both types of learning thus produce binary relevance predictions and, therefore, can be compared with each other.

5.3 Results

Each method was evaluated on a single problem in terms of a 10-fold cross validation. These evaluations were then averaged over a total number of 50 randomly generated problems. While averaging the performance over different data sets is questionable in general, we consider it legitimate in our case. In fact, all data sets are actually variants of the same problem, and indeed, the standard deviation of the performance was rather small throughout.

Table 1 summarizes the performance of the different methods for $m = 5$ and $m = 10$ in terms of the Hamming loss, subset zero-one loss, rank loss and C-index as performance metrics. As can be seen, the use of graded training data improves performance throughout, regardless of the learning method and the loss function. Comparing the respective mean values in terms of a paired t-test, the differences are significant at a significance level of 5%.

Note that, as an extension of the rank loss, the C-index is actually not intended for binary learning. We still included it, as it only requires a predicted ranking and a ground-truth labeling as input; thus, it can also be derived for the binary learner. Of course, this learner is at a disadvantage here, and indeed, the gains of the gradual learner for the C-index are slightly higher than those for the rank loss.

6 Summary and Conclusions

In this paper, we have proposed an extension of conventional multilabel classification, called *graded multilabel classification* (GMLC). The basic idea of GMLC is that the membership of an instance in a class or, say, the relevance of a label for an instance, is not a matter of "yes" or "no". Instead, the membership is measured on a *graded scale*, thus allowing for intermediate degrees of relevance. Here, we have focused on an *ordinal scale* as a special case, though numeric scales could in principle be used as well. In any case, a generalization of this kind appears to be useful and reasonable from a practical point of view.

Moreover, we have introduced two meta-techniques for reducing GMLC problems to existing machine learning problems, namely a vertical and a horizontal decomposition scheme. Whereas the former turns a GMLC problem into a set of ordinal classification problems, one for each label, the latter leads to solving a set of conventional multilabel problems, one for each level of the ordinal scale. In the context of these two techniques, we have also discussed the extension of MLC loss functions to the graded case.

Experimentally, we have shown that graded relevance does provide useful extra information from a learning point of view, even if only a binary prediction is requested. Collecting real-world GMLC data and complementing this study by further experiments is planned as future work. Besides, the GMLC framework gives rise to a number of interesting theoretical challenges, including but not limited to the simultaneous, monotonicity-preserving solution of the sub-problems produced by our reduction schemes.

Acknowledgments

We are grateful to Professor Abele-Brehm, University of Erlangen, for providing us the BELA-E data. This work has been supported by the Germany Research Foundation (DFG).

References

[Abele and Stief, 2004] A.E. Abele and M. Stief. Die Prognose des Berufserfolgs von Hochschulabsolventinnen und -absolventen. Befunde zur ersten und zweiten Erhebung der Erlanger Längsschnittstudie BELA-E. *Zeitschrift für Arbeits- und Organisationspsychologie*, 48:4–16, 2004.

[Balcan *et al.*, 2008] M.F. Balcan, N. Bansal, A. Beygelzimer, D. Coppersmith, J. Langford, and G.B. Sorkin. Robust reductions from ranking to classification. *Machine Learning*, 72(1–2):139–153, 2008.

[Berger, 1985] J.O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, 2. edition, 1985.

[Cheng and Hüllermeier, 2009] W. Cheng and E. Hüllermeier. Combining instance-based learning and logistic regression for multilabel classification. *Machine Learning*, 76(2–3):211–225, 2009.

[Frank and Hall, 2001] E. Frank and M. Hall. A simple approach to ordinal classification. In *Proc. ECML–2001*, pages 145–156, Freiburg, Germany, 2001.

[Fürnkranz *et al.*, 2008] J. Fürnkranz, E. Hüllermeier, E. Mencia, and K. Brinker. Multilabel classification via calibrated label ranking. *Machine Learning*, 73(2):133–153, 2008.

[Fürnkranz *et al.*, 2009] J. Fürnkranz, E. Hüllermeier, and S. Vanderlooy. Binary decomposition methods for multipartite ranking. In *Proc. ECML/PKDD–2009*, Bled, Slovenia, 2009.

[Gnen and Heller, 2005] Mithat Gnen and Glenn Heller. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*, 92(4):965–970, 2005.

[Herbrich *et al.*, 2000] Ralf Herbrich, Thore Graepel, and Klaus Obermayer. Large margin rank boundaries for ordinal regression. In *Advances in Large Margin Classifiers*, pages 115–132. MIT Press, 2000.

[Lin and Li, 2007] H.T. Lin and L. Li. Ordinal regression by extended binary classiffications. In *Proc. NIPS–07*, pages 865–872, 2007.

[Trohidis *et al.*, 2008] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas. Multilabel classification of music into emotions. In *Proc. Int. Conf. Music Information Retrieval*, 2008.

[Tsoumakas and Vlahavas, 2007] G. Tsoumakas and I. Vlahavas. Random k-labelsets: An ensemble method for multilabel classification. In *Proc. ECML–2007*, pages 406–417, Warsaw, 2007.

[Zadeh, 1965] L.A. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338–353, 1965.

Table 1: Performance (mean and standard deviation) in the case of $m = 5$ labels (above) and $m = 10$ labels (below).

	IBLR-ML		BR-10NN	
	binary	graded	binary	graded
Hamming loss	0.245±0.048	0.219±0.042	0.220±0.051	0.213±0.052
rank loss	0.190±0.062	0.180±0.057	0.328±0.115	0.310±0.104
C-index	0.204±0.047	0.183±0.045	0.381±0.089	0.361±0.080
subset zero-one loss	0.736±0.093	0.695±0.078	0.857±0.051	0.808±0.070
Hamming loss	0.225±0.017	0.207±0.018	0.230±0.018	0.217±0.018
rank loss	0.169±0.029	0.157±0.021	0.225±0.040	0.154±0.020
C-index	0.190±0.012	0.178±0.019	0.237±0.011	0.171±0.016
subset zero-one loss	0.908±0.028	0.875±0.042	0.913±0.022	0.893±0.034