Discrete Minimax Binary Relevance Classifier for Imbalanced Multi-label Classification

Salvador Madrigal, Vu-Linh Nguyen, Cyprien Gilet, Sébastien Destercke

Discrete Minimax Binary Relevance Classifier for Imbalanced Multi-Label Classification

Salvador Madrigal, Vu-Linh Nguyen, Cyprien Gilet, and Sébastien Destercke

Heudiasyc Laboratory, University of Technology of Compiègne, France
salvador.madrigal-castillo@etu.utc.fr,
{sebastien.destercke,vu-linh.nguyen,cyprien.gilet}@hds.utc.fr

Abstract. Multi-label classification (MLC) is a supervised learning problem where each instance can be associated with none, one, or multiple labels. MLC has received increasing attention due to its wide range of applications, such as text categorization and medical diagnosis. Despite a rich literature on MLC, handling imbalanced data, often encountered in real-world MLC datasets, has not been tackled satisfactorily. Based on a thorough literature review, it appears that the existing methods for imbalanced MLC are either hard to be coupled with sound theoretical guarantees or of limited scalability. This paper discusses the potential (dis)advantages of existing methods for imbalanced MLC, when being coupled with Binary relevance classifier (BRC), and introduces Discrete Minimax BRC (DMBRC), which would be a promising attempt to robustify the BRC by leveraging theoretically sound properties of the Discrete Minimax Classifier. We also provide empirical evidence to illustrate how DMBRC may be advantageous in balancing the label-wise error rates. Finally, we envision future works on further strengthening DMBRC in both label-wise error rates and conventional MLC evaluation metrics.

Keywords: Multi-label classification · Binary relevance · Imbalance data · Discrete Minimax Classifier.

1 Introduction

This paper seeds in the context of supervised multi-label classification (MLC) for safety-critical detection, such as diagnosing pathologies in precision medicine, detecting anomalies, frauds, or failures of components in condition monitoring.

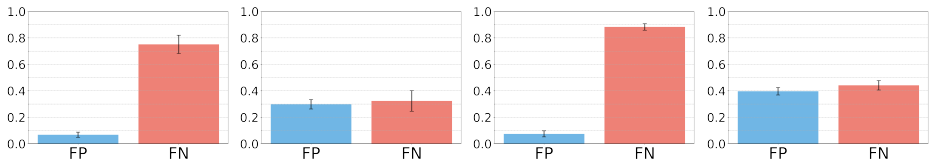
1.1 Problem Statements

Given observations of features which can be numeric, categorical, images, or any other kind of signals describing an instance, the purpose of MLC is to determine the actual class labels of the sample in order to support the experts of the application domain in their diagnosis.

Differ from usual *single-label classification tasks*, which classify instances into one of several mutually exclusive classes, *MLC* allows one to assign multiple class

labels to each instance [2,8,26,38]. An example of MLC problems is multi-cancer early detection [24], in which each patient may develop either non or multiple cancer types. Another MLC problem is fault detection, such as in aeroengines [42] and battery systems [43], where multiple faults can coexist.

While MLC is becoming increasingly promising in safety-critical detection, an important issue often occurs in such contexts: the presence of imbalanced labels. In MLC datasets that contain rare class labels, conventional classifiers, such as Multi-label k Nearest Neighbours (MLKNN) [46], tend to underestimate these class labels, predicting them less frequently, which results in a high number of false negatives class-wise (see Fig. 1(a) and (c)). Therefore, the goal is to reduce the number of false positives by balancing the error rates (see figure 1(b) and (d)). The issue of imbalanced labels generally harms classification performances associated with the most imbalanced labels, favouring the most represented situation. This may typically happen when some labels are important to detect but are scarcely observed and difficult to predict. For example, diagnosing rare pathologies in precision medicine, and predicting failures of a component in condition monitoring, are crucial but difficult to carry out.



(a) MLKNN Water qua. (b) DMBRC Water qua. (c) MLKNN Human Pse (d) DMBRC Human Pse

Fig. 1: Mean False Positive (FP) and False Negative (FN) rates for the fourth label with a prior probability of 0.21 in the Water Quality dataset, and for the second label (0.26) in the Human PseAAC dataset, obtained after a 5-fold cross-validation procedure using MLKNN and our new method named DMBRC.

1.2 Related works and state of the art

The issue of imbalanced classes has become more and more studied in MLC from the past decades [4,7,20,38], while remaining a challenging problem. In [38], the authors proposed a complete and interesting survey of methods designed to handle imbalanced data in MLC. They notably highlight that the methods aiming to address the imbalance problem in MLC can be divided into four categories: resampling methods [4,6,7,8,20], classifier adaptation [9,35,45], ensemble methods [23,29,37] and cost-sensitive approaches [3,11,40]. Let us note that commonly used single-label classifiers robust to imbalance data (such as Weighted Logistic Regression, Weighted Decision Trees [31]) can be straigthforwardly adapted

to MLC when considering the binary relevance strategy, and with more efforts when considering more complex techniques such as classifier chains [34].

Classifier adaptation techniques handling imbalanced labels often either (1) re-define each binary classification problem as a multi-class classification problem, in which the instances in the majority class are partitioned into multiple (sub)-classes, to reduce the potential impact of the majority classes during the training time [45] or (2) partition the input space into multiple regions and solve the binary classification problems on the regions independently [9] or (3) only allow the neighbor training instances to have impact on the prediction of the query instance [35]. Clearly, such techniques require one to choose at least one sensitive hyper-parameter, such as the number of (sub)-classes, the number of smaller regions, and the threshold that determines the nearest neighbors. Moreover, it is unclear why/how splitting the majority class into multiple (sub)-classes and partitioning the input space into multiple regions without further processing may help to mitigate class imbalance.

Cost-sensitive approaches [3,11,40] explicitly increase the impact of the instances from the minority class by adjusting class weights in the training loss. This would be analogous to learning a classifier on a modified training data set. Therefore, it might be inconvenient if one wishes to use the classifier for other purposes, such as doing descriptive statistics, i.e., summarizing the characteristics of the data set, or accommodating emerging evaluation metrics at the prediction time.

Resampling methods often oversample to increase the impact of instances with minority classes or undersample to decrease the impact of instances within the majority classes. Yet, it may provide promising predictive results in practice [4,6,7,8,20]. Similar to cost-sensitive approaches [3,11,40], they also learn a classifier on a modified training data set, yet in a much less controlled way, as they typically rely on random processes.

Ensemble methods consist of ensemble generation and aggregation. During ensemble generation, one can optionally couple the data-generating/sampling process (to train the ensemble members) with other methods for class imbalance. The chosen data-generating/sampling process may also amplify the imbalance. For example, bagging seems to lower the chance of selecting instances from the minority class when creating training data sets, from which the ensemble members are trained, and thus may lower the impact of instances from the minority class in general. It is also known that different aggregators may be in favor of different evaluation metrics [27].

1.3 Towards the minimax classifier

The issue of imbalance classes is well known in single-label classification since the past century [1,16,32], and the *Minimax Classifier* have been analytically demonstrated to be an optimal approach to deal with imbalanced data in the context of single-label classification [1,32]. The minimax classifier is indeed the Bayesian classifier for which the risks per class are all minimized and balanced.

In statistical decision theory [1,32], minimax classifiers are usually fitted by maximizing the Bayes risk with respect to the prior probabilities over the simplex, which requires to know the conditional distributions of the features in each class in order to analytically calculate the Bayes risk. However, this task remains a challenge in Machine Learning since we do not know the class-conditional distributions of the features, especially when dealing with several classes, mixed features, and arbitrary loss function. Indeed, computing an accurate estimate of the feature joint distribution in each class to achieve a good estimate of the empirical Bayes risk over the simplex remains highly complicated. Furthermore, in most real-world applications, the estimation of the empirical Bayes risk over the simplex is generally intractable because of the curse of dimensionality. Nowadays, only a few minimax algorithms have been proposed to deal with these general cases in single-label classification [16,19,21].

In the past few years, the authors in [16,19] developed a new approach for computing a minimax classifier for single-label classification tasks, suitable to process both numeric and categorical attributes, that can process a large number of classes, that can be coupled with any pretrained deep neural network [18], and which has been applied to precision medicine [15] or condition monitoring [17]. The procedure partitions the feature space beforehand and learns the minimax classifier by using a closed-form expression of the empirical Bayes risk over the simplex. While the authors show that discrete empirical Bayes risk is a concave non-differentiable multivariate piecewise affine function concerning the priors, they provide an efficient algorithmic procedure to obtain the Discrete Minimax Classifier (DMC). This approach can outperform several other state-of-the-art methods to obtain guaranteed robustness against imbalanced class risks, even when dealing with a large number of classes (for example $K > 100$). However, the use of minimax classifiers has not been studied yet in the context of MLC. The objective of this paper is to propose an opening approach to introduce the minimax classifier in MLC based on the DMC proposed in [16,19].

The paper is organized as follows. Section 2 recalls the main concepts of imbalanced MLC. In Section 3, we propose a binary relevance strategy to consider the DMC in MLC. Our proposed classifier aims to minimize and balance all the risks per class and with respect to any kind of loss/cost function that penalizes the classification errors. Section 4 empirically assesses our proposed classifier, in comparison with 11 other MLC methods/algorithms, on six real data sets.

2 Multi-label classification

This section recalls the main concepts of MLC. Let $\mathcal{X} = \mathbb{R}^d$ be a d-dimensional feature space, and let $\Lambda = \{\lambda_1, \ldots, \lambda_K\}$ be a set containing K labels. A multi-label sample is a pair $(\boldsymbol{x}, \boldsymbol{y})$, where $\boldsymbol{x} \in \mathcal{X}$ is a d-dimensional feature vector and $\boldsymbol{y} \in \mathcal{Y} := \{0,1\}^K$, where, for any $1 \leq k \leq K$, $y^k = 1$ ($y^k = 0$) indicates that the label λ_k is relevant/present (irrelevant). A multi-label dataset (MLD), $\mathcal{D} = \{(\boldsymbol{x}_n, \boldsymbol{y}_n) | 1 \leq n \leq N\}$, is formed by N pairs of multi-label samples. A

multi-label classifier, $\delta : \mathcal{X} \to \mathcal{Y}$, is a classifier for which given an unseen sample $\boldsymbol{x}$, the classifier returns a prediction $\hat{\boldsymbol{y}} \in \mathcal{Y}$.

There are two main approaches for dealing with MLC: problem transformation and algorithm adaptation [2]. The first one, problem transformation, aims to transform the MLC problem into one or more binary or multiclass classifications, some popular methods include: Binary relevance [44], classifier chains [33,34] and power label set [2]. On the other hand, algorithm adaptation, aims to adapt existing algorithms so that they can work with the MLC setting, like the MLKNN [46]. In this paper, we primarily explore the binary relevance (BR). The BR method consists in transforming the multi-label problem (MLP) into a collection of independent binary classification problems [44], one per label. Formally, given a multi-label data set $\mathcal{D} = \{(\boldsymbol{x}_n, \boldsymbol{y}_n) | 1 \leq n \leq N\}$, the BR methods consist in creating K different data sets $\mathcal{D}_k = \{(\boldsymbol{x}_n, \boldsymbol{y}_n^k) | 1 \leq n \leq N\}$, for $k = 1, \ldots, K$, and learning for each dataset $\mathcal{D}_k$ a binary classifier $\delta_k : \mathcal{X} \to \{1, 0\}$.

While BR methods, focusing in this paper, are among the most algorithmically simplest MLC methods, they are arguably sound methods for optimizing the commonly used decomposable losses [12]. Moreover, it seems to be a convenient choice regarding the trade-off between the theoretical soundness and computational expenses in various applications of MLC, where the training data can come with incomplete or missing data. A notable example of such scenarios would be predicting antimicrobial resistance phenotypes (susceptible, or resistant) of multiple drugs given genomic sequences of strains [13,25], where a significant proportion of training instances are partially annotated. Finally, BR can sometimes provide competitive empirical results, compared to power label sets and classifier chains, with respect to both decomposable and non-decomposable losses [27,41].

2.1 Imbalance in multi-label data

Many real-life applications suffer from label imbalance [38], highlighting the importance of determining whether an MLD is imbalanced. One straightforward method to determine label imbalance in an MLP is to observe the label distribution, as shown in the sequel by Fig. 3. This visualization provides an idea of how the labels are distributed, but is only useful when there are a few labels. For datasets with a large number of labels, it is beneficial to employ metrics to measure the imbalance. Different measures have been proposed [4] to quantify the imbalance present in a MLD.

Imbalance Ratio per Label (IRLbl) [4]. For any $\lambda_k \in \Lambda$, the IRLbl, defined as

$$\mathrm{IRLbl}(\lambda_k | \mathcal{D}) = \frac{\max\limits_{k^* \in \{1,\ldots,K\}} \left(\sum_{n=1}^{N} [\![y_n^{k^*} = 1]\!] \right)}{\sum_{n=1}^{N} [\![y_n^k = 1]\!]}, \tag{1}$$

where $[\![\cdot]\!]$ is the indicator function, i.e., $[\![A]\!] = 1$ if the predicate A is true and $[\![A]\!] = 0$ otherwise. The $\mathrm{IRLbl}(\lambda_k | \mathcal{D})$ is the ratio between the most frequent label

and λ_k. It is 1 for the most frequent label and a higher value means a higher level of imbalance for λ_k. Note however that the measure is not upper-bounded and is relative.

Mean Imbalance Ratio (MeanIR) [4]. This score measures the average level of imbalance in an MLD and is defined as

$$\text{MeanIR}(\mathcal{D}) = \frac{1}{K} \sum_{k=1}^{K} \text{IRLbl}(\lambda_k|\mathcal{D}) \,. \tag{2}$$

Of course, the same critic applies to MeanIR than to IRLbl.

Coefficient of Variation of IRLbl (CVIR) [4]. This score measures the variation of IRLbl and indicates if all labels have the same level of imbalance:

$$\text{CVIR}(\mathcal{D}) = \frac{1}{\text{MeanIR}(\mathcal{D})\sqrt{K-1}} \sqrt{\sum_{k=1}^{k} \left(\text{IRLbl}(\lambda_k|\mathcal{D}) - \text{MeanIR}(\mathcal{D})\right)^2}. \tag{3}$$

The higher the value the greater the level of imbalance between labels.

2.2 Performance metrics for multi-label problems

For binary and multi-class classification, the performance of a classifier depends on whether the unseen samples are correctly classified. In the case of MLC, predictions can considered as correct, partially correct, or partially incorrect [22]. We shall detail 3 commonly used metrics used in this paper.

For any query instance $\boldsymbol{x}$, let $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ be the true labels and the predicted labels, respectively. The subset 0/1 loss and the Hamming loss (respectively denoted by $\mathcal{L}_{0/1}$, $\mathcal{L}_{\text{Ham}}$ and reminded in equation (4)), which are the lower the better scores, assess the classifier δ according to its ability to exactly predict all the labels and accurately predict the labels on average, respectively.

$$\mathcal{L}_{0/1}(\delta|\mathcal{D}) = \frac{1}{N} \sum_{n=1}^{N} [\![\hat{\boldsymbol{y}}_n \neq \boldsymbol{y}_n]\!] \,, \quad \mathcal{L}_{\text{Ham}}(\delta|\mathcal{D}) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{K} \sum_{k}^{K} [\![\hat{y}_n^k \neq y_n^k]\!] \,. \tag{4}$$

The F1 score given by Equation (5) is a metric that combines precision and recall. It is calculated as the harmonic mean of precision and recall. Higher values indicate better performance of the model.

$$\text{F1}(\delta|\mathcal{D}) = 2 \sum_{n=1}^{N} \frac{\sum_{k=1}^{K} \hat{y}_n^k \cdot y_n^k}{\sum_{k=1}^{K} \hat{y}_n^k + \sum_{k=1}^{K} y_n^k} \,. \tag{5}$$

3 Discrete Minimax Binary Relevance Classifier

The DMC [16] is a single-label classifier that aims to address the imbalance issue with statistical guarantees and when considering any kind of loss function L (which allows one to penalize the classification errors). More precisely, the DMC aims to minimize and balance the risks per class during the training procedure. In particular, in the context of binary classification and when considering the 0/1 loss function, the DMC allows one to minimize and balance the false positive and false negative rates, especially when dealing with imbalance data.

The primary goal of this paper is to extend the statistical guarantees provided by the DMC [16] to the MLC using the Binary Relevance strategy. In other words, given a MLD, $\mathcal{D} = \{(\boldsymbol{x}_n, \boldsymbol{y}_n) | 1 \leq n \leq N\}$, we create K different datasets $\mathcal{D}_k = \{(\boldsymbol{x}_n, y_n^k) | 1 \leq n \leq N\}$, for $k = 1, \ldots, K$, and we aim to learn for each set $\mathcal{D}_k$ a binary DMC denoted by $\delta_k^M : \mathcal{X} \to \{1, 0\}$.

To this aim, given a training set $\mathcal{D}_k$, we first partition the feature space $\mathcal{X}$ into T different regions $\Omega = \{\phi_1, \ldots, \phi_T\}$, as proposed in [16] , using for example the Kmeans partitioning or the decision tree partitioning. This defines a mapping $\boldsymbol{\Phi} : \mathcal{X} \to \Omega$, which maps any instance $\boldsymbol{x} \in \mathcal{X}$ to a discrete profile $\boldsymbol{\Phi}(\boldsymbol{x}) \in \Omega$. For all $t \in \{1, \ldots, T\}$ and for each binary class $\ell \in \{0, 1\}$, the estimated probability that an instance $\boldsymbol{x} \in \mathcal{X}$ has the discrete profile ϕ_t given its real class y^k is $\ell \in \{0, 1\}$ is given by

$$\hat{p}_{\ell t} := \frac{1}{|\mathcal{I}_\ell|} \sum_{i \in \mathcal{I}_\ell} [\![\boldsymbol{\Phi}(\boldsymbol{x}_i) = \phi_t]\!]. \tag{6}$$

Here, $\mathcal{I}_\ell$ is the set containing all the training instances such that $y_i^k = \ell$.

For the following, let us define $L = \{L_{\ell j} : \ell, j \in \{0, 1\}\}$ the costs of classification errors such that $L_{\ell,j}$ is the cost of predicting the class j given that the real class is ℓ. For example, when considering the 0/1 loss function, we have $L_{0,0} = L_{1,1} = 0$ and $L_{0,1} = L_{1,0} = 1$. Furthermore, let us define $\mathbb{S} = \left\{\pi = [\pi_0, \pi_1] \in [0, 1]^2 : \pi_0 + \pi_1 = 1\right\}$ the 2-dimensional simplex.

Based on the partitioned feature space $\Omega = \boldsymbol{\Phi}(\mathcal{X})$ and similarly to [16], we can demonstrate that the empirical Bayes risk $V_k : \mathbb{S} \to \mathbb{R}_+$ (as a function of the priors), associated with the binary-classification set $\mathcal{D}_k$, is analytically given by

$$V_k(\pi) = \sum_{\ell=0}^{1} \sum_{t=1}^{T} \sum_{z=0}^{1} L_{\ell z}\, \pi_\ell\, \hat{p}_{\ell t}\, [\![\sum_{j \in \{0,1\}} L_{jz} \pi_j \hat{p}_{jt} = \min_{q \in \{0,1\}} \sum_{j \in \{0,1\}} L_{jq} \pi_j \hat{p}_{jt}]\!]. \tag{7}$$

We can furthermore demonstrate that V_k is a concave multivariate piecewise affine function over $\mathbb{S}$ with a finite number of pieces.

Finally, similarly to [16], the DMC $\delta_k^M : \mathcal{X} \to \{1, 0\}$ associated with the binary classification dataset $\mathcal{D}_k$ is given by

$$\delta_k^M : \boldsymbol{x} \mapsto \arg\min_{j \in \{0,1\}} \sum_{t=1}^{T} \sum_{\ell=0}^{1} L_{\ell j}\, \bar{\pi}_\ell\, \hat{p}_{\ell t}\, [\![\boldsymbol{\Phi}(\boldsymbol{x}_i) = \phi_t]\!]. \tag{8}$$

In equation (8), $\bar{\pi} = [\bar{\pi}_0, \bar{\pi}_1]$ corresponds to the least favorable priors that maximize the Bayes risk V_k over the simplex $\mathbb{S}$ and can be easily computed with a

projected subgradient algorithm [16]. Note that the classifier obtained by equation (8) will balance class-wise risks, as illustrated by Fig. 1 (b) and (d). Without entering into too much details (due to page limits), the DMC will seek the class prior distribution $\bar{\pi}$ that will balance the risks.

For the MLC, the idea of the **Discrete Minimax Binary Relevance Classifier** (DMBRC) is to extend the statistical guarantees from the multi-class problem to the multi-label problem. To achieve, first, the feature space of the MLP, $\mathcal{X}$, is discretized, and then following the idea of the BR method, a DMC, $\delta^B_{\bar{\pi}} : \mathcal{X} \to \{0, 1\}$ for the binary is trained per each label individually.

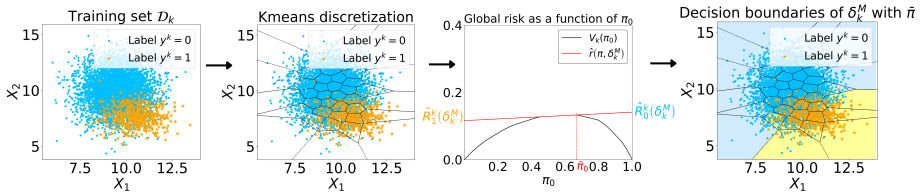


Fig. 2: Illustrative example of how to compute δ^M_k.

4 Experiments

In this section, we conduct an empirical study to assess the performance of our new Discrete Minimax Binary Relevance Classifier (DMBRC), compared to other methods suitable to imbalanced MLC.

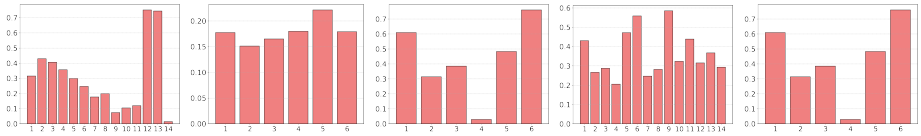
4.1 Experimental setting

The datasets used are the following: yeast, scene, CHD_49 and Tmc2007 from the Mulan repository [39]; Water-quality (water qua) from the repository from the University of Cordoba (see, `https://www.uco.es/kdis/mllresources/`) and HumanPseAAC (Human Pse) from the cometa repository [5]. Information about the different datasets can be viewed in Table 1 and their distributions per label are given in Figure 3, with the exception of TMC2007 due to its large number of labels. Note that they display various levels of imbalance and disparity across labels: some data sets have labels that are highly imbalanced, such as yeast (label 9), CHD_49 (label 4) or Human Pse. (label 4), while others have labels whose percentage of appearance remain relatively high across labels (scene and water quality), with different degrees of disparity (scene percentage are all similar, while water quality has a higher disparity).

For each data set, a 5−fold-cross-validation is employed, and for each method, the mean and standard deviation (std) of the performance scores on the test sets are reported. The subset zero-one metric (4), F1 score (5), and Hamming

Table 1: Overview of datasets

Dataset	Instances	Features	Labels	MeanIR	CVIR
Yeast	2417	103	14	7.20	1.88
Scene	2407	294	6	1.25	0.12
CHD_49	555	49	6	5.77	1.75
Water qua	1060	16	14	15.29	1.08
Human Pse.	3106	440	14	1.77	0.30
Tmc2007	28596	49060	22	17.13	0.81



(a) Yeast (b) Scene (c) CHD_49 (d) Water qua. (e) Human Pse.

Fig. 3: Datasets distribution per label.

metric (4) are employed as conventional MLC performance scores, which are not specially designed for assessing the ability to balance the class conditional risk. To assess this aspect, for each given classifier δ, we quantify the ability to balance the false positive rate and false negative rate associated with each label $\lambda \in \Lambda$ using the following criterion

$$\psi(\delta) = \max_{\lambda \in \Lambda} \left| R_0^\lambda(\delta) - R_1^\lambda(\delta) \right|, \tag{9}$$

where $R_0^\lambda(\delta)$ denotes the false negative rate associated with the label λ for the classifier δ and $R_1^\lambda(\delta)$ the false positive rate. A smaller $\psi(\delta)$ indicates that the multi-label classifier δ provides a better balance of the false positive and false negative rates. A perfectly balanced classifier would have $\psi(\delta) = 0$, a value $\psi(\delta) = 1$ indicating that at least for one label, a classifier always predict either false or positive.

To seek an extensive empirical study, we implemented the DMBRC and 11 other competitors in Python. The competitors can be divided into three groups of different natures. The first group consists of three classifiers that do not take into account class imbalance: the MLKNN classifier and two binary relevance methods with logistic regression (LR) and decision tree (DT) as the base learners. The second group consists of two cost-sensitive binary relevance classifiers, weighted logistic regression (WLR) and weighted decision tree (WDT), which explicitly take into account class imbalance as part of the training loss. The third group consists of six resampling methods, which employ multi-label random oversampling [7] (ROS), multi-label random undersampling [7] (RUS), and multi-label SMOTE [8] (SM) to resample the datasets and then train binary relevance LR and DT. We refer to these methods as ROS+LR, ROS+DT, RUS+LR, RUS+DT, SM+LR, and SM+DT.

With the exception of MLKNN [46], whose implementation is provided in the Scikit-multilearn package [36], all the other methods are of binary relevance nature and only differ in their base learners. We use the implementation for binary relevance from the Scikit-multilearn package [36], which allows customization of the base learner using base learners from other packages such as the Scikit-learn library [31], from which we call functions for training LR and DT.

For the multi-label resampling methods, we use an unofficial implementation on GitHub (see `https://github.com/Seal-Li/Multi-label-imbalance`). For DMBRC, we employ the DMC using the source code provided by [16], with some modifications. The source code used in our experiments has been made public at `https://github.com/SalvadorMadrigal/DMBRC-for-Imbalanced-MLC`.

4.2 Results

The predictive performance, in terms of the subset zero-one metric (4), F1 score (5), and Hamming metric (4), and the ψ scores are given in Table **??** and Table**??**. Overall, the predictive performance provided by DMBRC is on par with the ones provided by other cost-sensitive classifiers, i.e., WLR and WDT. This is a satisfying result since DMBRC, WLR, and WDT all take into account the class imbalance, but in different manners and are useful for different purposes (besides opting for the subset zero-one metric (4), F1 score (5), and Hamming metric (4)). Therefore, DMBRC complements the existing collection of classifiers/methods to balance accuracies meaningfully.

Considering the full set of methods/classifiers covered, DMBRC is never the worst, but also never the best. If one ranks the methods/classifiers covered by their performance, DMBRC consistently occupies middle positions in the ranks. This is perfectly expected: the additional constraint of being balanced will typically lower the average accuracy, compared to an unconstrained classifier optimizing average accuracy. It is therefore normal that the DMBRC cannot outperform, in average, methods such as MLKNN, LR or DT that are not also taking care of balancing the true positive and negative rates for each label. This should therefore not be considered as bad news as DMBRC and other cost-sensitive classifiers essentially scarify average scores to gain a better balance of the class-conditional risk, i.e., their primary focus.

However, when looking at the ψ scores (9), we clearly see that DMBRC and other cost-sensitive classifiers do their balancing jobs well, as they consistently outperform other competitors on this criterion. Interestingly, DMBRC provides the best scores on all the tested data sets, showing that its theoretical properties are paired with very good empirical results. This would be a strong motivation to further boost DMBRC in both its primary focus of balancing the false positive and false negative rates associated with each label.

5 Conclusion

This paper complements the literature on handling imbalanced MLC data with a new approach based on the theoretical minimax strategy [1,14,16,32]. Our new

Table 2: Evaluation of the subset zero-one metric ($\mathcal{L}_{0/1}$), F-1 score (F1), Hamming metric ($\mathcal{L}_{\text{Ham}}$), and ψ metric. The arrow next to the metric represents ↑ the higher the better and ↓ the lower the better. The results are presented as [mean ± std]. The best and the worst performances are given in **bold** and red, respectively.

Metric	Classifier	Yeast	Scene	CHD_49
$\mathcal{L}_{0/1}$ ↓	MLKNN	**0.81 ± 0.02**	**0.37 ± 0.02**	0.88 ± 0.05
	LR	**0.85 ± 0.01**	0.46 ± 0.02	**0.83 ± 0.05**
	DT	0.93 ± 0.01	0.60 ± 0.03	0.88 ± 0.04
	DMBRC	0.91 ± 0.02	0.68 ± 0.03	0.95 ± 0.02
	WLR	0.93 ± 0.02	0.57 ± 0.02	0.91 ± 0.04
	WDT	0.97 ± 0.01	0.76 ± 0.02	0.96 ± 0.02
	ROS+LR	0.86 ± 0.01	0.47 ± 0.02	0.86 ± 0.04
	ROS+DT	0.94 ± 0.02	0.60 ± 0.01	0.87 ± 0.04
	RUS+LR	0.85 ± 0.01	0.47 ± 0.02	0.84 ± 0.04
	RUS+DT	0.93 ± 0.01	0.64 ± 0.03	0.87 ± 0.03
	SM+LR	0.91 ± 0.01	0.54 ± 0.02	0.93 ± 0.02
	SM+DT	0.96 ± 0.01	0.63 ± 0.01	0.88 ± 0.02
F1 ↑	MLKNN	0.61 ± 0.02	**0.70 ± 0.02**	0.59 ± 0.04
	LR	0.61 ± 0.01	0.62 ± 0.01	0.64 ± 0.02
	DT	0.56 ± 0.02	0.47 ± 0.04	0.59 ± 0.04
	DMBRC	0.52 ± 0.01	0.64 ± 0.01	0.48 ± 0.04
	WLR	0.52 ± 0.02	**0.70 ± 0.02**	0.54 ± 0.01
	WDT	0.46 ± 0.01	0.59 ± 0.01	0.47 ± 0.04
	ROS+LR	0.60 ± 0.01	0.62 ± 0.01	0.61 ± 0.03
	ROS+DT	0.54 ± 0.03	0.48 ± 0.02	0.59 ± 0.04
	RUS+LR	0.60 ± 0.01	0.61 ± 0.02	0.64 ± 0.01
	RUS+DT	0.56 ± 0.01	0.44 ± 0.03	0.60 ± 0.03
	SM+LR	**0.64 ± 0.01**	0.67 ± 0.02	**0.67 ± 0.02**
	SM+DT	0.57 ± 0.01	0.56 ± 0.02	0.65 ± 0.03
$\mathcal{L}_{\text{Ham}}$ ↓	MLKNN	**0.20 ± 0.01**	**0.09 ± 0.01**	0.32 ± 0.02
	LR	**0.20 ± 0.01**	0.10 ± 0.00	**0.29 ± 0.02**
	DT	0.23 ± 0.01	0.13 ± 0.00	0.32 ± 0.02
	DMBRC	0.33 ± 0.01	0.18 ± 0.01	0.44 ± 0.02
	WLR	0.33 ± 0.01	0.13 ± 0.00	0.37 ± 0.01
	WDT	0.38 ± 0.01	0.20 ± 0.01	0.46 ± 0.02
	ROS+LR	**0.20 ± 0.00**	0.10 ± 0.00	0.31 ± 0.02
	ROS+DT	0.24 ± 0.01	0.13 ± 0.00	0.30 ± 0.02
	RUS+LR	**0.20 ± 0.01**	0.10 ± 0.00	**0.29 ± 0.01**
	RUS+DT	0.23 ± 0.00	0.14 ± 0.01	0.31 ± 0.02
	SM+LR	0.25 ± 0.01	0.12 ± 0.01	0.35 ± 0.01
	SM+DT	0.30 ± 0.01	0.16 ± 0.00	0.31 ± 0.02
ψ ↓	MLKNN	1.00 ± 0.00	0.49 ± 0.03	1.00 ± 0.00
	LR	1.00 ± 0.00	0.50 ± 0.05	1.00 ± 0.00
	DT	1.00 ± 0.00	0.62 ± 0.07	1.00 ± 0.00
	DMBRC	**0.66 ± 0.20**	**0.09 ± 0.02**	**0.45 ± 0.23**
	WLR	0.68 ± 0.19	0.11 ± 0.03	0.65 ± 0.17
	WDT	0.67 ± 0.11	0.13 ± 0.04	0.48 ± 0.15
	ROS+LR	1.00 ± 0.00	0.49 ± 0.04	0.92 ± 0.02
	ROS+DT	1.00 ± 0.00	0.62 ± 0.05	1.00 ± 0.00
	RUS+LR	1.00 ± 0.00	0.59 ± 0.06	0.97 ± 0.04
	RUS+DT	1.00 ± 0.00	0.68 ± 0.05	1.00 ± 0.00
	SM+LR	1.00 ± 0.00	0.25 ± 0.04	1.00 ± 0.00
	SM+DT	1.00 ± 0.00	0.35 ± 0.02	1.00 ± 0.00

Table 3: Evaluation of the subset zero-one metric ($\mathcal{L}_{0/1}$), F-1 score (F1), Hamming metric ($\mathcal{L}_{\text{Ham}}$), and ψ metric. The arrow next to the metric represents ↑ the higher the better and ↓ the lower the better. The results are presented as [mean ± std]. The best and the worst performances are given in **bold** and red, respectively.

Metric	Classifier	Water qua	Human Pse	Tmc2007
$\mathcal{L}_{0/1}$ ↓	MLKNN	**0.84 ± 0.01**	**0.98 ± 0.00**	0.74 ± 0.01
	LR	**0.84 ± 0.01**	**0.98 ± 0.01**	**0.69 ± 0.00**
	DT	0.89 ± 0.02	0.99 ± 0.01	0.83 ± 0.01
	DMBRC	0.97 ± 0.01	0.99 ± 0.01	0.96 ± 0.01
	WLR	0.92 ± 0.00	0.99 ± 0.01	0.83 ± 0.01
	WDT	0.99 ± 0.00	0.99 ± 0.00	0.98 ± 0.00
	ROS+LR	0.85 ± 0.01	0.99 ± 0.01	0.70 ± 0.01
	ROS+DT	0.91 ± 0.01	0.99 ± 0.01	0.83 ± 0.01
	RUS+LR	0.84 ± 0.02	0.99 ± 0.01	0.69 ± 0.00
	RUS+DT	0.92 ± 0.01	0.99 ± 0.01	0.83 ± 0.00
	SM+LR	0.90 ± 0.01	1.00 ± 0.00	0.76 ± 0.01
	SM+DT	0.90 ± 0.01	1.00 ± 0.00	0.84 ± 0.00
F1 ↑	MLKNN	0.20 ± 0.01	0.54 ± 0.01	0.62 ± 0.00
	LR	0.22 ± 0.01	0.47 ± 0.01	**0.68 ± 0.00**
	DT	0.19 ± 0.01	0.49 ± 0.01	0.52 ± 0.01
	DMBRC	0.30 ± 0.01	0.53 ± 0.02	0.45 ± 0.00
	WLR	**0.35 ± 0.01**	0.53 ± 0.01	0.63 ± 0.00
	WDT	0.24 ± 0.01	0.52 ± 0.01	0.41 ± 0.01
	ROS+LR	0.21 ± 0.01	0.48 ± 0.01	0.68 ± 0.00
	ROS+DT	0.11 ± 0.01	0.49 ± 0.01	0.53 ± 0.01
	RUS+LR	0.22 ± 0.02	0.49 ± 0.01	0.68 ± 0.00
	RUS+DT	0.11 ± 0.02	0.49 ± 0.01	0.53 ± 0.01
	SM+LR	0.13 ± 0.01	**0.57 ± 0.01**	0.69 ± 0.00
	SM+DT	0.13 ± 0.01	0.56 ± 0.01	0.60 ± 0.01
$\mathcal{L}_{\text{Ham}}$ ↓	MLKNN	0.09 ± 0.00	**0.29 ± 0.01**	0.07 ± 0.00
	LR	0.09 ± 0.00	**0.29 ± 0.01**	0.06 ± 0.00
	DT	0.09 ± 0.00	0.30 ± 0.00	0.08 ± 0.00
	DMBRC	0.23 ± 0.01	0.36 ± 0.01	0.20 ± 0.00
	WLR	0.17 ± 0.00	0.35 ± 0.01	0.10 ± 0.00
	WDT	0.27 ± 0.01	0.36 ± 0.01	0.22 ± 0.00
	ROS+LR	0.09 ± 0.00	**0.29 ± 0.01**	0.06 ± 0.00
	ROS+DT	0.09 ± 0.00	0.31 ± 0.01	0.08 ± 0.00
	RUS+LR	0.09 ± 0.00	**0.29 ± 0.01**	0.06 ± 0.00
	RUS+DT	0.09 ± 0.00	0.31 ± 0.00	0.08 ± 0.00
	SM+LR	**0.08 ± 0.00**	0.41 ± 0.01	0.07 ± 0.00
	SM+DT	0.09 ± 0.00	0.39 ± 0.01	0.08 ± 0.00
ψ ↓	MLKNN	1.00 ± 0.00	0.84 ± 0.03	0.90 ± 0.03
	LR	1.00 ± 0.00	0.97 ± 0.03	0.88 ± 0.04
	DT	1.00 ± 0.00	0.94 ± 0.08	1.00 ± 0.00
	DMBRC	**0.87 ± 0.04**	**0.19 ± 0.04**	**0.09 ± 0.03**
	WLR	0.97 ± 0.01	0.25 ± 0.05	0.22 ± 0.06
	WDT	0.90 ± 0.03	0.29 ± 0.03	0.16 ± 0.02
	ROS+LR	1.00 ± 0.00	0.96 ± 0.03	0.81 ± 0.06
	ROS+DT	1.00 ± 0.00	0.85 ± 0.02	1.00 ± 0.00
	RUS+LR	1.00 ± 0.00	0.96 ± 0.03	0.87 ± 0.04
	RUS+DT	1.00 ± 0.00	0.88 ± 0.07	1.00 ± 0.00
	SM+LR	1.00 ± 0.00	0.92 ± 0.03	0.92 ± 0.02
	SM+DT	1.00 ± 0.00	0.70 ± 0.09	1.00 ± 0.00

classifier DMBRC attempts to robustify the binary relevance classification by leveraging theoretically sound properties provided by the DMC [16].

We moreover provide empirical evidence to illustrate how classifiers, which take into account class imbalance, may be advantageous when being assessed by the ability to balance the class conditional risk. The empirical evidence also suggests that DMBRC is the most promising classifier in this aspect.

Motivated by these promising pieces of evidence, we envision future works on further strengthening DMBRC in both label-wise error rates and conventional MLC evaluation metrics. We also plan to leverage the fact that DMC can be coupled with deep neural networks to handle images to broaden the application domain of DMBRC to high-stakes applications such as predicting multiple diseases given medical images [30]. In addition, since DMC theoretical guarantees are not limited to binary classification, we also think of extending the current approaches to the graded multi-label [10] setting or to multi-dimensional classification [28], for which approaches such as WLR or WDT.

References

1. Berger, J.O.: Statistical decision theory and Bayesian analysis; 2nd ed. Springer Series in Statistics, Springer, New York (1985)
2. Bogatinovski, J., Todorovski, L., Džeroski, S., Kocev, D.: Comprehensive comparative study of multi-label classification methods. Expert Systems with Applications **203**, 117215 (2022)
3. Cao, P., Liu, X., Zhao, D., Zaiane, O.: Cost sensitive ranking support vector machine for multi-label data learning. In: Proceedings of the 16th International Conference on Hybrid Intelligent Systems (HIS 2016). pp. 244–255. Springer (2017)
4. Charte, F., Rivera, A., del Jesus, M.J., Herrera, F.: A first approach to deal with imbalance in multi-label datasets. In: Hybrid Artificial Intelligent Systems: 8th International Conference, HAIS 2013, Salamanca, Spain, September 11-13, 2013. Proceedings 8. pp. 150–160. Springer (2013)
5. Charte, F., Rivera, A.J., Charte, D., del Jesus, M.J., Herrera, F.: Tips, guidelines and tools for managing multi-label datasets: The mldr. datasets r package and the cometa data repository. Neurocomputing **289**, 68–85 (2018)
6. Charte, F., Rivera, A.J., del Jesus, M.J., Herrera, F.: Mlenn: a first approach to heuristic multilabel undersampling. In: Intelligent Data Engineering and Automated Learning–IDEAL 2014: 15th International Conference, Salamanca, Spain, September 10-12, 2014. Proceedings 15. pp. 1–9. Springer (2014)
7. Charte, F., Rivera, A.J., del Jesus, M.J., Herrera, F.: Addressing imbalance in multilabel classification: Measures and random resampling algorithms. Neurocomputing **163**, 3–16 (2015)
8. Charte, F., Rivera, A.J., del Jesus, M.J., Herrera, F.: Mlsmote: Approaching imbalanced multilabel learning through synthetic instance generation. Knowledge-Based Systems **89**, 385–397 (2015)
9. Chen, K., Lu, B.L., Kwok, J.T.: Efficient classification of multi-label and imbalanced data using min-max modular classifiers. In: The 2006 IEEE International Joint Conference on Neural Network Proceedings. pp. 1770–1775. IEEE (2006)
10. Cheng, W., Hüllermeier, E., Dembczynski, K.J.: Graded multilabel classification: The ordinal case. In: Proceedings of the 27th international conference on machine learning (ICML-10). pp. 223–230 (2010)

11. Daniels, Z., Metaxas, D.: Addressing imbalance in multi-label classification using structured hellinger forests. In: Proceedings of the AAAI conference on artificial intelligence. vol. 31 (2017)
12. Dembczynski, K., Cheng, W., Hüllermeier, E.: Bayes optimal multilabel classification via probabilistic classifier chains. In: Proceedings of the 27th International Conference on Machine Learning (ICML). pp. 279–286 (2010)
13. Do, V.H., Nguyen, S.H., Le, D.Q., Nguyen, T.T., Nguyen, C.H., Ho, T.H., Vo, N.S., Nguyen, T., Nguyen, H.A., Cao, M.D., et al.: Panka: Leveraging population pangenome to predict antibiotic resistance. iScience (2024)
14. Gilet, C.: Discrete minimax classifier for personalized diagnosis in medicine. PhD Thesis, Université Côte d'Azur (2021), https://tel.archives-ouvertes.fr/tel-03553934
15. Gilet, C., Barbosa, S., Fillatre, L.: Minimax classifier with box constraint on the priors. In: Machine Learning for Health (ML4H) at NeurIPS 2019. Proceedings of Machine Learning Research (2019)
16. Gilet, C., Barbosa, S., Fillatre, L.: Discrete box-constrained minimax classifier for uncertain and imbalanced class proportions. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(6), 2923–2937 (2020)
17. Gilet, C., Fillatre, L.: Anomaly detection with discrete minimax classifier for imbalanced datasets or uncertain class proportions. In: World Congress on Condition Monitoring 2019. Springer (2019)
18. Gilet, C., Guyomard, M., Barbosa, S., Fillatre, L.: Multiclass minimax learning for deep neural networks. Proceedings of the 31st European Signal Processing Conference (EUSIPCO) (2023)
19. Gilet, C., Guyomard, M., Destercke, S., Fillatre, L.: Softmin discrete minimax classifier for imbalanced classes and prior probability shifts. Machine Learning (2023). https://doi.org/10.1007/s10994-023-06397-8
20. Giraldo-Forero, A.F., Jaramillo-Garzón, J.A., Ruiz-Muñoz, J.F., Castellanos-Domínguez, C.G.: Managing imbalanced data sets in multi-label problems: a case study with the smote algorithm. In: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 18th Iberoamerican Congress, CIARP 2013, Havana, Cuba, November 20-23, 2013, Proceedings, Part I 18. pp. 334–342. Springer (2013)
21. Guerrero-Curieses, A., Alaíz-Rodríguez, R., Cid-Sueiro, J.: A fixed-point algorithm to minimax learning with neural networks. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) **34**, 383–392 (2004)
22. Han, M., Wu, H., Chen, Z., Li, M., Zhang, X.: A survey of multi-label classification based on supervised and semi-supervised learning. International Journal of Machine Learning and Cybernetics **14**(3), 697–724 (2023)
23. Liu, B., Tsoumakas, G.: Making classifier chains resilient to class imbalance. In: Asian Conference on Machine Learning. pp. 280–295. PMLR (2018)
24. Liu, M.C., Oxnard, G., Klein, E., Swanton, C., Seiden, M., Liu, M.C., Oxnard, G.R., Klein, E.A., Smith, D., Richards, D., et al.: Sensitive and specific multicancer detection and localization using methylation signatures in cell-free dna. Annals of Oncology **31**(6), 745–759 (2020)
25. Moradigaravand, D., Palm, M., Farewell, A., Mustonen, V., Warringer, J., Parts, L.: Prediction of antibiotic resistance in escherichia coli from large-scale pangenome data. PLoS computational biology **14**(12), e1006258 (2018)
26. Nguyen, V.L., Hüllermeier, E.: Multilabel classification with partial abstention: Bayes-optimal prediction under label independence. Journal of Artificial Intelligence Research **72**, 613–665 (2021)

27. Nguyen, V.L., Hüllermeier, E., Rapp, M., Loza Mencía, E., Fürnkranz, J.: On aggregation in ensembles of multilabel classifiers. In: Proceedings of the 23rd International Conference on Discovery Science (DS). pp. 533–547 (2020)
28. Nguyen, V.L., Yang, Y., De Campos, C.: Probabilistic multi-dimensional classification. In: Uncertainty in Artificial Intelligence. pp. 1522–1533. PMLR (2023)
29. Pakrashi, A., Mac Namee, B.: Stacked-mlknn: a stacking based improvement to multi-label k-nearest neighbours. In: First International Workshop on Learning with Imbalanced Domains: Theory and Applications. pp. 51–63. PMLR (2017)
30. Paul, H.Y., Kim, T.K., Siegel, E., Yahyavi-Firouz-Abadi, N.: Demographic reporting in publicly available chest radiograph data sets: opportunities for mitigating sex and racial disparities in deep learning models. Journal of the American College of Radiology **19**(1), 192–200 (2022)
31. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. the Journal of machine Learning research **12**, 2825–2830 (2011)
32. Poor, H.V.: An Introduction to Signal Detection and Estimation. Springer-Verlag New York, 2nd edn. (1994)
33. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. Machine Learning **85**, 333–359 (2011)
34. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains: A review and perspectives. Journal of Artificial Intelligence Research **70**, 683–718 (2021)
35. Sun, K.W., Lee, C.H.: Addressing class-imbalance in multi-label learning via two-stage multi-label hypernetwork. Neurocomputing **266**, 375–389 (2017)
36. Szymański, P., Kajdanowicz, T.: A scikit-based python environment for performing multi-label classification. arXiv preprint arXiv:1702.01460 (2017)
37. Tahir, M.A., Kittler, J., Bouridane, A.: Multilabel classification using heterogeneous ensemble of multi-label classifiers. Pattern Recognition Letters **33**(5), 513–523 (2012)
38. Tarekegn, A.N., Giacobini, M., Michalak, K.: A review of methods for imbalanced multi-label classification. Pattern Recognition **118**, 107965 (2021)
39. Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J., Vlahavas, I.: Mulan: A java library for multi-label learning. The Journal of Machine Learning Research **12**, 2411–2414 (2011)
40. Wu, G., Tian, Y., Liu, D.: Cost-sensitive multi-label learning with positive and negative label pairwise correlations. Neural Networks **108**, 411–423 (2018)
41. Wu, G., Zhu, J.: Multi-label classification: do hamming loss and subset accuracy really conflict with each other? In: Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS). pp. 3130–3140 (2020)
42. Yuan, Y., Liu, X., Ding, S., Pan, B.: Fault detection and location system for diagnosis of multiple faults in aeroengines. IEEE Access **5**, 17671–17677 (2017)
43. Zhang, K., Hu, X., Liu, Y., Lin, X., Liu, W.: Multi-fault detection and isolation for lithium-ion battery systems. IEEE Transactions on Power Electronics **37**(1), 971–989 (2021)
44. Zhang, M.L., Li, Y.K., Liu, X.Y., Geng, X.: Binary relevance for multi-label learning: an overview. Frontiers of Computer Science **12**, 191–202 (2018)
45. Zhang, M.L., Li, Y.K., Yang, H., Liu, X.Y.: Towards class-imbalance aware multi-label learning. IEEE Transactions on Cybernetics **52**(6), 4459–4471 (2020)
46. Zhang, M.L., Zhou, Z.H.: Ml-knn: A lazy learning approach to multi-label learning. Pattern recognition **40**(7), 2038–2048 (2007)