Pattern Recognition 47 (2014) 1494–1508

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Dependent binary relevance models for multi-label classification

Elena Montañes [a], Robin Senge [b], Jose Barranquero [a], José Ramón Quevedo [a], Juan José del Coz [a,*], Eyke Hüllermeier [b]

[a] Artificial Intelligence Center, University of Oviedo at Gijón, 33204 Gijón, Spain
[b] Mathematics and Computer Science, Marburg University, 35032 Marburg, Germany

ARTICLE INFO

Article history:
Received 31 May 2013
Received in revised form
24 September 2013
Accepted 25 September 2013
Available online 4 October 2013

Keywords:
Multi-label classification
Label dependence
Stacking
Chaining

ABSTRACT

Several meta-learning techniques for multi-label classification (MLC), such as chaining and stacking, have already been proposed in the literature, mostly aimed at improving predictive accuracy through the exploitation of label dependencies. In this paper, we propose another technique of that kind, called *dependent binary relevance* (DBR) learning. DBR combines properties of both, chaining and stacking. We provide a careful analysis of the relationship between these and other techniques, specifically focusing on the underlying dependency structure and the type of training data used for model construction. Moreover, we offer an extensive empirical evaluation, in which we compare different techniques on MLC benchmark data. Our experiments provide evidence for the good performance of DBR in terms of several evaluation measures that are commonly used in MLC.

1. Introduction

Multi-label classification (MLC) is a machine learning problem in which models are sought that assign a subset of (class) labels to each object, unlike conventional (single-class) classification that involves predicting only a single class. Multi-label classification problems are ubiquitous and naturally occur, for instance, in assigning keywords to a paper, tags to resources in a social network, objects to images or emotional expressions to human faces.

There is a considerable amount of literature, in which state-of-the-art binary or multi-class classification algorithms are adapted and extended to the setting of MLC, including methods using decision trees [1], instance-based algorithms [2], neural networks [3], support vector machines [4], naive Bayes [5], conditional random fields [6] and boosting [7]. Besides, there is also another line of research, in which approaches of that kind are completely put aside; instead, the development of specialized methods that consider the particularities of multi-label data is advocated.

In general, the problem of multi-label learning is coming with two fundamental challenges. The first one bears on the computational complexity of the algorithms. If the number of labels is large, then a complex approach might not be applicable in practice. Therefore, the scalability of algorithms is a key issue in this field. The second problem is related to the very nature of multi-label data. Not only is the number of classes typically larger than in multi-class classification tasks, but also each example belongs to a variable-sized subset of labels simultaneously. Moreover, and perhaps even more importantly, the labels will normally not occur independent of each other; instead, there are statistical dependencies between them. From a learning and prediction point of view, these relationships constitute a promising source of information, in addition to that coming from the mere description of the objects. Thus, it is hardly surprising that research on MLC has very much focused on the design of new methods that are able to detect—and benefit from—interdependencies among labels.

In recent years, many papers have analyzed the presence of label correlations, including theoretical analyses of label dependence in the context of MLC [8]. In this regard, different types of dependence have been formally distinguished, such as conditional dependence [6,9–12] and marginal (unconditional) dependence [3,13,14]. Other papers are aiming at the exploitation of relations in different sets of labels, such as pairwise relations [3,4,7,15,16], relations in sets of different sizes [11,17,18], or relations in the whole set of labels [10,13,14]. Exploiting label dependence implicates the induction of complex models. In fact, the more the label combinations are considered, the more complex the models are. This does not mean that exploiting pairwise correlations is preferable to exploiting full-order correlations, since the former may fail to capture the true dependencies while the latter may not work well if the labels display complex relations that are difficult to deal with.

This paper proposes *dependent binary relevance* (DBR) models as an efficient and effective approach to induce multi-label classifiers that exploit conditional label dependence. Instead of studying them in combination with independent classifiers, like in [10], our goal is

* Corresponding author. Tel.: +34 985182501; fax: +34 985182125.
E-mail address: juanjo@aic.uniovi.es (J. José del Coz).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.09.029

to explore their behavior when used in isolation, extending the work presented in [19] in which this approach was favorably compared with several state-of-the-art methods [3,11,13,18]. The DBR approach is conceived as a natural extension of the simple binary relevance strategy, which does not allow for exploiting conditional label dependence. We shall elaborate on the positioning of our approach more closely in Section 3, where we argue that this approach combines properties of two other meta-techniques for MLC, namely chaining [11] and stacking [14], and that it fills a "gap" within the spectrum of methods that have been devised so far.

A key contribution of this paper is a deep analysis of the properties of dependent binary models, in which we characterize those conditions under which they should work well in practice. These models require label estimations (produced by any multi-label classifier) at prediction time. This issue is analyzed throughout the paper, concluding that the more reliable these estimations are, the better the overall performance becomes.

Another contribution of this work is to present a comprehensive study of methods based on chaining [11] and stacking [14] strategies. Our goal is to analyze these two approaches, which are closely connected, and to study those factors that have an influence on their performance. A key distinction between both approaches is the type of training data they rely on, which in turn has a decisive impact on the kind of label dependence captured.

The rest of the paper is organized as follows. The next section introduces multi-label classification in a more formal way. Stacking and chaining methods are reviewed in Section 3. Section 4 is devoted to the new DBR technique; we describe this approach formally and provide a detailed analysis of its properties. Finally, experimental results are reported in Section 5, before concluding the paper in Section 6.

2. Multi-label classification

Before describing some previous approaches to tackle multi-label classification, we present this learning task in a more formal way. The point of departure is a finite and non-empty set of labels $\mathcal{L}=\{\ell_1,\ell_2,\ldots,\ell_m\}$ and a training set $S=\{(\boldsymbol{x}_1,\boldsymbol{y}_1),\ldots,(\boldsymbol{x}_n,\boldsymbol{y}_n)\}$. The elements of this set are supposed to be independently and randomly drawn according to an unknown probability distribution $\mathbf{P}(\mathbf{X},\mathbf{Y})$ on $\mathcal{X}\times\mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are the input and the output space respectively. The former is the space of the object descriptions (instances), whereas the latter is given by the power set $\mathcal{P}(\mathcal{L})$ of $\mathcal{L}$. To ease notation, we define $\boldsymbol{y}_i$ as a binary vector $\boldsymbol{y}_i=(y_{i,1},y_{i,2},\ldots,y_{i,m})$ in which $y_{i,j}=1$ indicates the presence (relevance) and $y_{i,j}=0$ the absence (irrelevance) of ℓ_j in the labeling of $\boldsymbol{x}_i$. Using this convention, the output space can also be defined as $\mathcal{Y}=\{0,1\}^m$. The goal in MLC is to induce from S a hypothesis $\boldsymbol{h}:\mathcal{X}\longrightarrow\mathcal{Y}$ that correctly predicts the subset of relevant labels for unlabeled query instances $\boldsymbol{x}$.

The most straightforward and arguably simplest approach to tackle multi-label classification is *binary relevance* (BR). The BR strategy reduces a given multi-label problem with m labels to m binary classification problems. More precisely, m hypotheses $h_1,h_2,\ldots,h_m$ are induced, each of them being responsible for predicting the relevance of one label, using just $\mathcal{X}$ as the input space:

$$h_j:\mathcal{X}\longrightarrow\{0,1\}. \tag{1}$$

In this way, the labels are predicted independent of each other and no label dependencies are taken into account. Yet, despite its inability to exploit any label dependencies, the BR algorithm also exhibits several advantages: (i) each binary learning method can be used as base learner, (ii) it has linear complexity with respect to the number of labels and (iii) it can be easily parallelized.

In spite of its simplicity, the BR method obtains competitive results in benchmark datasets whenever being applied on top of a state-of-the-art base learner with a proper procedure for tuning parameters. Interestingly, it has been shown theoretically and empirically that BR performs quite strong in terms of decomposable loss functions [9]. This behavior can be explained by studying BR from a probabilistic point of view. Given that each binary model h_j is able to estimate $\mathbf{P}(y_j|\boldsymbol{x})$, BR is well-suited for every loss function whose risk minimizer can be expressed in terms of marginal distributions of labels. Since the classifier used for learning h_j commonly optimizes its accuracy, the whole BR model minimizes the Hamming loss[1]:

$$\text{HammingLoss}(\boldsymbol{y},\boldsymbol{h}(\boldsymbol{x}))=\frac{1}{m}\sum_{i=1}^{m}[\![\,y_i\neq h_i(\boldsymbol{x})\,]\!]. \tag{2}$$

This measure averages the standard 0/1 classification error over the m labels and hence corresponds to the proportion of labels whose relevance is incorrectly predicted. Besides, if an appropriate base learner is employed, then BR is also able to optimize all other macro-average label-based metrics, such as the macro-F_1 measure [20].

On the other hand, the decomposition approach followed by BR affects its performance for those loss functions whose minimization requires an estimation of the joint distribution. Examples of these measures are micro-average metrics and Subset 0/1 loss, which looks if the predicted and relevant label subsets are equal or not:

$$\text{Subset}_{0/1}(\boldsymbol{y},\boldsymbol{h}(\boldsymbol{x}))=[\![\,\boldsymbol{y}\neq\boldsymbol{h}(\boldsymbol{x})\,]\!]. \tag{3}$$

In these cases, it is necessary to develop algorithms which are able to estimate the joint label probability distributions to obtain predictions that minimize this sort of metrics. Dembczyński et al. [8,21] present a formal probabilistic analysis of multi-label classification, studying the connection between risk minimization and loss functions.

3. Modeling label dependence

The arguably most natural way to capture label dependencies is to learn classifier models that *condition* the prediction of a label y_i not only on the object features $\boldsymbol{x}$ but also on some of the other labels y_j. This idea of conditioning can be realized in different ways. In particular, the following distinctions can be made:

(i) *Full vs. partial conditioning*: The prediction of y_i can be conditioned on all other labels $\{y_1,\ldots,y_{i-1},y_{i+1},\ldots,y_m\}$ or only on a subset of these labels. The most "sparse" conditioning scheme among those that capture full dependence between labels is a sequential structure: y_i is conditioned on $\{y_1,\ldots,y_{i-1}\}$. This structure, which constitutes the core of the idea of *classifier chains* (to be discussed further below), can be motivated by the product rule of probability [9]:

$$\mathbf{P}(\boldsymbol{y}|\boldsymbol{x})=\prod_{i=1}^{m}\mathbf{P}(y_i|\boldsymbol{x},y_1,\ldots,y_{i-1}) \tag{4}$$

(ii) *True vs. predicted label information*: For training the predictor of y_i, the other labels y_j are available in the training data and, therefore, can in principle be used for learning this predictor. Alternatively, the model for y_i can be trained on the estimations $\hat{y}_j$ produced by the other predictors.

[1] The expression $[\![\,p\,]\!]$ evaluates to 1 if the predicate p is true, and to 0 otherwise.