Pattern Recognition 56 (2016) 170–183

Contents lists available at ScienceDirect

Pattern Recognition

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

Labelling strategies for hierarchical multi-label classification techniques

Isaac Triguero [a,b,*], Celine Vens [c]

[a] Department of Respiratory Medicine, Ghent University, Ghent 9000, Belgium
[b] Data Mining and Modelling for Biomedicine group, VIB Inflammation Research Center, Zwijnaarde 9052, Belgium
[c] Department of Public Health and Primary Care, KU Leuven Kulak, Kortrijk 8500, Belgium

ARTICLE INFO

Article history:
Received 16 October 2015
Received in revised form
29 January 2016
Accepted 24 February 2016
Available online 4 March 2016

Keywords:
Hierarchical multi-label classification
Threshold optimisation
Hierarchical loss
HMC-loss
F-measure

ABSTRACT

Many hierarchical multi-label classification systems predict a real valued score for every (instance, class) couple, with a higher score reflecting more confidence that the instance belongs to that class. These classifiers leave the conversion of these scores to an actual label set to the user, who applies a cut-off value to the scores. The predictive performance of these classifiers is usually evaluated using threshold independent measures like precision-recall curves. However, several applications require actual label sets, and thus an automatic labelling strategy.

In this paper, we present and evaluate different alternatives to perform the actual labelling in hierarchical multi-label classification. We investigate the selection of both single and multiple thresholds. Despite the existence of multiple threshold selection strategies in non-hierarchical multi-label classification, they cannot be applied directly to the hierarchical context. The proposed strategies are implemented within two main approaches: optimisation of a certain performance measure of interest (such as F-measure or hierarchical loss), and simulating training set properties (such as class distribution or label cardinality) in the predictions. We assess the performance of the proposed labelling schemes on 10 datasets from different application domains. Our results show that selecting multiple thresholds may result in an efficient and effective solution for hierarchical multi-label problems.

1. Introduction

Traditional classification problems deal with assigning a (single) class to an instance. However, many applications require assigning a *set* of classes (labels) to an instance. Examples are found in biology (e.g., gene function prediction [1,2]), text or image classification [3,4], etc. Multi-label classification algorithms have been proposed to tackle this task [5–7]. In many applications, the set of possible labels is structured as a hierarchy, representing a superclass/subclass relation. For instance, gene functions are organised as a tree structure in MIPS's FunCat hierarchy [8], or as a directed acyclic graph (DAG) in the Gene Ontology [9]. The corresponding classification task, which also takes into account this structure, is then called hierarchical multi-label classification (HMC) [10]. It thus involves predicting multiple and partial paths in a hierarchy of labels. Allowing partial paths means that the true and predicted paths need not necessarily end in a leaf node. Several HMC algorithms have been proposed in the literature, e.g., [11–13]. They exploit the label set hierarchy when labelling instances. These systems also ensure (implicitly or using post-processing) that the hierarchy constraint is fulfilled in the predictions they make: whenever a class is predicted, its parent and ancestor classes are also predicted.

Rather than predicting an actual label set, most of the HMC algorithms actually predict a real valued prediction score p_i for every label l_i, that reflects the confidence that an instance should be annotated with label l_i. These values can be easily converted into a label set by applying a threshold on them: if p_i is above some threshold t_i, then the instance is predicted to belong to class l_i, otherwise not. To ensure that the predictions fulfil the hierarchy constraint, it suffices to choose $t_i \leq t_j$ whenever l_i is a super class of l_j.

Often, the decision as to which thresholds to choose is left to the end user, and the predictive performance of the classification algorithms is evaluated in a threshold independent way, for example, by using precision-recall curves. However, in some situations, it is preferable or necessary to fix the thresholds. For

[*] Corresponding author at: Department of Respiratory Medicine, Ghent University, Ghent 9000, Belgium. Tel.: +32 09 331 36 93; fax: +32 09 221 76 73.
E-mail addresses: Isaac.Triguero@irc.vib-UGent.be (I. Triguero), Celine.Vens@kuleuven-kulak.be (C. Vens).

http://dx.doi.org/10.1016/j.patcog.2016.02.017
0031-3203/© 2016 Elsevier Ltd. All rights reserved.

instance, the gene function prediction task may be part of a larger pipeline of experiments, or the predicted image labels may be used as tags in image retrieval systems to locate images of interest. The objective of this article is to investigate and empirically compare different thresholding strategies.

HMC studies that fix the thresholds typically choose one threshold shared by all labels. In the non-hierarchical multi-label setting, however, studies exist that choose a separate threshold per label [14,15]. It is currently an open question how these two options compare in HMC, and this is addressed in this article. Non-hierarchical optimisation techniques cannot be straightforwardly applied in the HMC context, because of the aforementioned hierarchy constraint, and thus, we propose adapted techniques. Depending on the context, the user may want to set the thresholds such that the resulting classifier maximises predictive performance or such that training set properties (such as class distribution) are reflected in the predictions. We consider both approaches. In order to apply the former approach, we first critically review several performance measures used in HMC to compare a predicted label set to a true label set: hierarchical loss, HMC-loss and micro-averaged F-measure.

The contributions of this work are as follows. First, we describe measures that evaluate the predicted label sets, and we identify problems with the widely used (unweighted) hierarchical loss, which leads us to advise against its use (Section 2). Second, we devise a number of multiple-threshold-selection approaches for HMC (Section 3). Third, we empirically investigate the designed schemes and their single-threshold-selection counterparts on ten HMC datasets, showing that the multiple threshold approaches generally outperform their single threshold variants, both in predictive performance and computationally (Section 4). We draw some conclusions and further research directions (Section 5).

2. Evaluating HMC classifiers

In HMC we obtain for every instance and every label a prediction. As mentioned in the introduction, this prediction is often real-valued. Given a hierarchy of k labels, we represent the predicted multi-label of an instance x with a vector $p=(p_1,\ldots,p_k)\in\mathbb{R}^k$. The label hierarchy can be represented by a partial order $\le_h$ that represents the superclass relationship. For all labels l_1 and l_2: $l_1 \le_h l_2$ if and only if l_1 is a superclass of l_2. In the following discussion, we assume that p fulfils the hierarchy constraint: $p_{l_i} \ge p_{l_j}$ whenever $l_i \le_h l_j$.

In order to evaluate the predicted multi-labels in a test set, there are two possible strategies. The first strategy keeps the real-valued predictions, and evaluates them independently of any fixed thresholds. This is often done by constructing an average precision–recall curve (PR curve) and reporting the area under the curve. Precision gives the proportion of positive predictions that are positive, while recall gives the proportion of positive instances that are correctly predicted positive. A precision–recall curve plots the precision of a model as a function of its recall. While a threshold corresponds to a single point in PR space, by varying the threshold a curve is obtained. Vens et al. [11] and Pillai et al. [15] describe how to compute PR curves in the context of multiple labels.

The second strategy is to convert the predicted multi-labels to binary vectors, by thresholding the predicted values, and to evaluate these binary multi-labels. In non-hierarchical multi-label classification several evaluation measures have been proposed for evaluating binary multi-labels. An overview is given by Tsoumakas et al. [6]. However, these measures are less suited for HMC tasks, exactly because they do not take into account the hierarchical structure in the labels. Kiritchenko et al. [16] formulate three requirements that should be fulfilled by a hierarchical evaluation measure (see the simple label hierarchy in Fig. 1, where $\{I,J\}$ is indicated as the true multi-label to be predicted):

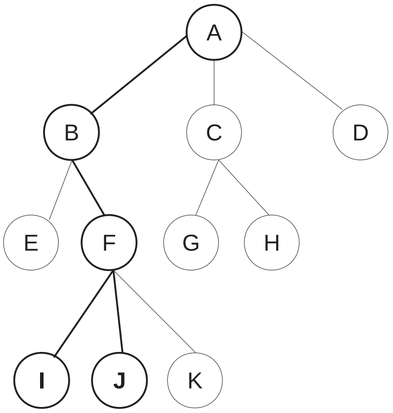

Fig. 1. Toy class hierarchy.

1. *The measure should give credit to a partially correct classification.* Thus, predicting node K should be better than predicting node C, as the prediction of K involves the path ABF that is part of the correct multi-label.
2. *The measure should punish distant errors more heavily.* This requirement is split further into two parts:
   (a) *The measure should give a higher evaluation for correctly classifying one level down, than to stay at the parent.* Thus, predicting F should be better than predicting B.
   (b) *The measure should give a lower evaluation for incorrectly classifying one level down than to stay at the parent.* Thus, predicting H should be worse than predicting C.
3. *The measure should punish errors at higher levels of the hierarchy more heavily.* This means that, e.g., predicting D when the true label is C should be worse than predicting K when the true label is I.

Examples of evaluation measures for binary multi-labels that do take into account a hierarchical label structure are hierarchical loss functions and a hierarchical extension of the F-measure. In the following, we represent the thresholded (binary) predicted multi-label of an instance with a vector $\hat{p}=(\hat{p}_1,\ldots,\hat{p}_k)\in\{0,1\}^k$; similarly, we represent the true multi-label with a vector $l=(l_1,\ldots,l_k)\in\{0,1\}^k$. Without loss of generality, we also assume a single root node in the hierarchy. In the case of a collection of separate hierarchies (such as the Gene Ontology, which consists of three independent sub-graphs), this means that we create an artificial root node, to which all instances belong. This node then has as children the individual root nodes of the sub-hierarchies.

2.1. Hierarchical loss functions

The hierarchical loss (H-loss) function [17] was proposed specifically for HMC tasks. It assumes a tree structured label hierarchy. It is based on the Hamming or symmetric difference loss, which returns the symmetric difference between the predicted and true multi-label vector for an instance. However, the H-loss does not punish mistakes that have already been punished at a higher level in the hierarchy. In other words, whenever a classification mistake is made on a label in the hierarchy, the H-loss does not charge any loss for additional mistakes occurring in the subtree of that label:

$$\text{H-loss}(\hat{p},l)=\sum_{i=1..k} c_i\{\hat{p}_i \ne l_i \text{ and } \hat{p}_j = l_j, j\in anc(i)\}, \tag{1}$$

where $anc(i)$ represents the set of ancestors of node i, and $c_1,\ldots,c_k>0$ are fixed cost coefficients. Cesa-Bianchi et al. [18] propose