Journal of Biomedical Informatics xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

A multi-label approach using binary relevance and decision trees applied to functional genomics

Erica Akemi Tanaka[a], Ségio Ricardo Nozawa[b,1], Alessandra Alaniz Macedo[a], José Augusto Baranauskas[a,*]

[a] Department of Computer Science and Mathematics, University of Sao Paulo (USP), Av. Bandeirantes, 3900, Ribeirão Preto, SP 14040-901, Brazil
[b] Dow AgroSciences (Seeds, Traits & Oils), Av. Antonio Diederichsen, 400, Ribeirão Preto, SP 14020-250, Brazil

ARTICLE INFO

Article history:
Received 14 August 2014
Accepted 18 December 2014
Available online xxxx

Keywords:
Multi-label classification
Decision tree
Functional genomics

ABSTRACT

Many classification problems, especially in the field of bioinformatics, are associated with more than one class, known as multi-label classification problems. In this study, we propose a new adaptation for the Binary Relevance algorithm taking into account possible relations among labels, focusing on the interpretability of the model, not only on its performance. Experiments were conducted to compare the performance of our approach against others commonly found in the literature and applied to functional genomic datasets. The experimental results show that our proposal has a performance comparable to that of other methods and that, at the same time, it provides an interpretable model from the multi-label problem.

1. Introduction

Since the advance of hardware and software, the automated sequencing of DNA fragments has become possible. The amount of biological data available has been increasing, which also increases the need for computational tools for knowledge extraction. Machine learning techniques are widely used to predict gene functions so that the best predictions can then be tested in the lab to validate the results [1]. However, predicting gene functions is a complex process because a single gene may have multiple functions. Consequently, multi-label classification seems to be appropriated.

There are several reasons to investigate and propose new multi-label classification techniques, especially in the bioinformatics or bio-related research fields. *Gene Ontology*[2] is an example of a multi-label problem, where genes and proteins may have more than one function or feature. Another example is the MIPS Functional Catalogue [2], in which genes and proteins may belong to more than one functional class. Therefore, it is very important to carry out research on computational techniques to classify multi-label problems using proteins, genes and other biological and medical data: with such knowledge it is possible to develop new drugs, treat diseases, and help in diagnostics.

Traditional algorithms are unable to handle a set of multi-label instances, since such algorithms were designed to predict a single label. A simple solution to this is to transform the original dataset into several sets of instances where each set contains all the attributes, but only one label to be predicted. This algorithm is known as *Binary Relevance* (BR). However, studies have shown that this approach is not a good solution [3,4], since each label is treated individually, generating one classifier for each label, and ignoring possible correlations among them. An algorithm that finds a classifier for more than one label can intuitively capture some correlations between them, and a simpler classifier may be found (one which uses a smaller number of rules, for example). Under these circumstances, it is important to research and develop techniques that use the *Binary Relevance* algorithm, extending it to capture possible relations among labels.

This study presents a new adaptation of the *Binary Relevance* algorithm using decision trees to treat multi-label problems. Decision trees are symbolic learning models that can be analyzed as set of rules in order to improve the understanding, by human experts, about the knowledge extracted. For this reason, the algorithm proposed here was designed to capture relations between labels, a feature the original *Binary Relevance* algorithm does not take into account, and consequently upgrade its generalization ability. Furthermore, since the present study takes model interpretability into account (and not only performance), our approach reduces the number of induced trees for expert interpretation: in the best scenario, it builds only one model (tree) that classifies all labels.

This paper is organized as follows: Section 2 describes related studies in the literature; Section 3 presents the basic concepts of multi-label classification; Section 4 presents our multi-label learning algorithm. Section 5 describes the experimental methodology to

[*] Corresponding author. Fax: +55 16 3315 0407.
E-mail addresses: kemi@gmail.com (E.A. Tanaka), srnozawa@gmail.com (S. Ricardo Nozawa), ale.alaniz@usp.br (A.A. Macedo), augusto@usp.br (J.A. Baranauskas).
[1] Fax: +55 16 3602 5696.
[2] http://www.geneontology.org/.

http://dx.doi.org/10.1016/j.jbi.2014.12.011
1532-0464/© 2014 Published by Elsevier Inc.

evaluate our approach; Results and discussion are presented in Section 6. Finally, Section 7 presents the final remarks and future work.

2. Related work

Different techniques have been proposed in the literature for treating multi-label classification problems. In some of them, single-label classifiers are combined to treat multi-label classification problems. Other techniques modify single-label classifiers, changing their algorithms to allow their use in multi-label problems.

BR + algorithm [5], an extension of the BR algorithm, considers the relationship between labels, and constructs binary classification problems, similarly to BR. Its main differences are its descriptor attributes, which merge all original attributes as well as all labels, except for the label to be predicted itself.

Another study using decision trees for hierarchical multi-label classification was used to analyze information about *Saccharomyces cerevisiae*, and tries to predict new gene functions [3]. Resampling strategies were developed, and a modified version of the algorithm C4.5 [6] was used.

The Mulam [7] tool was developed based on the Weka machine learning library [8], and contains several algorithms, such as BR (Binary Relevance) [9], LP (Label Powerset) [9], RaKel (RAndom k-labELsets) [10], and ML-kNN (Multi-Label k-Nearest Neighbours) [11]. In the Binary Relevance algorithm, the original dataset is divided into sets of instances, where each instance contains all the attributes but only the label to be predicted. Then, c classifiers are induced (where c represents the total number of labels), and each induced classifier is trained to distinguish one label against all the others involved. The Label Powerset algorithm is based on a combination of more than one label to create a new one, but this may result in a considerable increase in the number of labels, and some may end up with few instances. The RAkEL algorithm constructs an ensemble of LP classifiers, and each classifier is trained with a small subset of k random labels. Algorithm ML-KNN is based on algorithm kNN: for each test instance, its k nearest neighbors in the training set are identified. Then, according to statistical information from the label set of neighboring instances, the maximum a posteriori principle is applied to determine the label set for a particular test instance.

A tool called Clus [12] uses concepts from *Predictive Clustering Trees* (PCT). Decision trees are constructed where each node corresponds to a group of instances from the dataset. PCT is a clustering approach that adapts the basic top-down induction of decision trees for clustering. The procedure used for constructing the PCT is similar to other induction algorithms of decision trees such as C4.5 [6] and CART [13]. Clus-HMC [14] refers to the use of Clus as a multi-label hierarchical classification system that learns a tree to classify all labels, and Clus-SC generates a decision tree for each label.

MHCAIS (Multi-label Hierarchical Classification with an Artificial Immune System) [15] is an adapted algorithm for multi-label and hierarchical classification. The first version of this algorithm builds a global classifier to predict all labels, while the second version builds a classifier for each label. In both versions, the classifier is expressed as a set of IF–THEN rules, which has the advantage of being knowledge understandable to specialists.

Other researchers developed a Network Hierarchical Multi-label Classification algorithm that exploits individual properties of proteins as well as protein–protein interactions (PPI) to predict gene/protein functions [16]. These researchers advocate that (i) the PPI network is exploited in the training phase and can thus make predictions for genes/proteins whose interactions are yet to be investigated; (ii) their method yields better performance than the others by using network and properties separately; and (iii) the use of network information improves the accuracy of gene function prediction not only for highly connected genes, but also for genes with only a few connections. Like Clus-HMC, NHMC also exploits the hierarchical organization of class labels (gene functions), which may have the form of a tree or of a direct acyclic graph (DAG).

The R3P-Loc is a multi-label ridge regression classifier that uses two databases for feature extraction, applying random projection to reduce its feature dimensions [17]. In terms of locating proteins within cellular contexts, R3P-Loc indicates a reduction in the number of dimensions of feature vectors as much as seven-folds, while it also improves the classification performance. Considering the multi-level classification of phylogenetic profiles, authors have proposed an algorithm to capture, at each level, the different aspects of affinity of a protein with another, in the same or in different species [18]. As a result, inter and intra-genome gene clusters are predicted. Aiming at facilitating biological interpretation, the same authors extract close gene associations from metabolic pathways through unsupervised clustering at a sequence level [19]. This level of association can be enhanced if the phylogenetic relationship of the corresponding genomes is taken under consideration.

3. Background: multi-label classification

Basically, the classification task aims to discover knowledge that can be used to predict the unknown class of an instance, based on the values of the attributes that describe such an instance. As a result, we can divide the classification tasks according to the number of labels to be predicted for each instance into two groups: (a) Single-label Classification and (b) multi-label classification. Single-label classification refers to the classification task where there is only one label (the target concept) to be predicted [20]. The basic principles of multi-label classification are similar to single-label classification, however the multi-label classification has two or more concept labels to be predicted. Considering symbolic models expressed as rules, a multi-label classification rule contains two or more conclusions, each one involving a different label.

Next, we formalize the notation used in the remaining text. Let X be the domain of instances to be classified, Y be the set of labels, and H be the set of classifiers for $f : X \rightarrow Y$, where f is unknown. The goal is to find the classifier $h \in H$, maximizing the probability of $h(x) = y$, where $y \in Y$ is the ground truth label of x [21].

Table 1 shows the modified representation of *attribute–value* to deal with multi-label problems. A dataset is characterized by N instances $z_1, z_2, \ldots, z_N$, each containing m attributes $X_1, X_2, \ldots, X_m$ and c labels $Y_1, Y_2, \ldots, Y_c$. On this table, row i refers to the i-th instance ($i = 1, 2, \ldots, N$); entry x_{ij} refers the value of j-th attribute ($j = 1, 2, \ldots, m$) of instance i, and output y_{ik} refers to the value of k-th label ($k = 1, 2, \ldots, c$) of instance i. The instances are tuples $\vec{z}_i = (x_{i1}, x_{i2}, \ldots, x_{im}, y_{i1}, y_{i2}, \ldots, y_{ic}) = (\vec{x}_i, \vec{y}_i)$ also denoted by $z_i = (x_i, y_i)$, where the fact that z_i, x_i and y_i are vectors is implicit. Note each y_i is a member of the set $Y_1 \times Y_2 \times \ldots \times Y_c$; without loosing generality we will assume $Y_i \in \{0, 1\}$, i.e., each label will only assume binary values.

4. Proposal: The BR-DT algorithm

Next, before introducing our algorithm, we introduce some additional notations:

- D: the full dataset with all attributes and labels $\{X_1, \ldots, X_m, Y_1, \ldots, Y_c\}$;

Table 1
Set of instances in the attribute–value format for multi-label problems.

	X_1	X_2	$\cdots$	X_m	Y_1	Y_2	$\cdots$	Y_c
z_1	x_{11}	x_{12}	$\cdots$	x_{1m}	y_{11}	y_{12}	$\cdots$	y_{1c}
z_2	x_{21}	x_{22}	$\cdots$	x_{2m}	y_{21}	y_{22}	$\cdots$	y_{2c}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
z_N	x_{N1}	x_{N2}	$\cdots$	x_{Nm}	y_{N1}	y_{N2}	$\cdots$	y_{Nc}

Q1 Please cite this article in press as: Tanaka EA et al. A multi-label approach using binary relevance and decision trees applied to functional genomics. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.12.011