Pattern Recognition 70 (2017) 89–103

Contents lists available at ScienceDirect

Pattern Recognition

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

Hierarchical Multi-label Classification using Fully Associative Ensemble Learning

L. Zhang, S.K. Shah, I.A. Kakadiaris*

Computational Biomedicine Lab, 4849 Calhoun Rd, Rm 373, Houston, TX 77204, United States

ARTICLE INFO

Article history:
Received 26 October 2016
Revised 4 April 2017
Accepted 7 May 2017
Available online 8 May 2017

Keywords:
Hierarchical multi-label classification
Ensemble learning
Ridge regression

ABSTRACT

Traditional flat classification methods (*e.g.*, binary or multi-class classification) neglect the structural information between different classes. In contrast, Hierarchical Multi-label Classification (HMC) considers the structural information embedded in the class hierarchy, and uses it to improve classification performance. In this paper, we propose a local hierarchical ensemble framework for HMC, *Fully Associative Ensemble Learning* (FAEL). We model the relationship between each class node's global prediction and the local predictions of all the class nodes as a multi-variable regression problem with Frobenius norm or l_1 norm regularization. It can be extended using the kernel trick, which explores the complex correlation between global and local prediction. In addition, we introduce a binary constraint model to restrict the optimal weight matrix learning. The proposed models have been applied to image annotation and gene function prediction datasets with tree structured class hierarchy and large scale visual recognition dataset with Direct Acyclic Graph (DAG) structured class hierarchy. The experimental results indicate that our models achieve better performance when compared with other baseline methods.

Published by Elsevier Ltd.

1. Introduction

Hierarchical Multi-label Classification (HMC) is a variant of classification where each sample has more than one label and all these labels are organized hierarchically in a tree or Direct Acyclic Graph (DAG). In reality, HMC can be applied to many domains [1–3]. In web page classification, one website with the label "football" could be labeled with a high level label "sport". In image annotation, an image tagged as "outdoor" might have other low level concept labels, like "beach" or "garden". In gene function prediction, a gene can be simultaneously labeled as "metabolism" and "catalytic or binding activities" by the biological process hierarchy and the molecular function hierarchy, respectively.

A rich source of hierarchical information in tree and DAG structured class hierarchies is helpful to improve classification performance [4]. Based on how this information is used, previous HMC approaches can be divided into global (big-bang) or local [5]. Global approaches learn a single model for the whole class hierarchy. Global approaches enjoy smaller model size because they build one model for the whole hierarchy. However, they ignore the local modularity, which is an essential advantage of HMC. Local approaches first build multiple local classifiers on the class hierarchy. Then, hierarchical information is aggregated across the local prediction results of all the local classifiers to obtain the global prediction results for all the nodes. We refer to "local prediction result" and "global prediction result" as "local prediction" and "global prediction", respectively. Previous local approaches suffer from three drawbacks. First, most of them focus only on the parent-child relationship. Other relationships in the hierarchy (e.g., ancestor-descendant, siblings) are ignored. Second, their models are sensitive to local prediction. The global prediction of each node is only decided by the local predictions of several closely related nodes. The error of local predictions is more likely to propagate to global predictions. Third, most local methods assume that the local structural constraint between two nodes will be reflected in their local predictions. However, this assumption might be shaken by different choices of features, local classification models, and positive-negative sample selection rules [6,7]. In such situations, previous methods would fail to integrate valid structural information into local prediction.

In this paper, we propose a novel local HMC framework, Fully Associative Ensemble Learning (FAEL). We call it "fully associative ensemble" because in our model the global prediction of each node considers the relationships between the current node and all the other nodes. Specifically, a multi-variable regression model is built to minimize the empirical loss between the global predictions of all the training samples and their corresponding true label observations.

* Corresponding author.
E-mail addresses: lzhang34@uh.edu (L. Zhang), sshah@central.uh.edu (S.K. Shah), ioannisk@uh.edu (I.A. Kakadiaris).

http://dx.doi.org/10.1016/j.patcog.2017.05.007
0031-3203/Published by Elsevier Ltd.

Our contributions are: we (i) developed a novel local hierarchical ensemble framework, in which all the structural relationships in the class hierarchy are used to obtain global prediction; (ii) introduced empirical loss minimization into HMC, so that the learned model can capture the most useful information from historical data; and (iii) proposed sparse, kernel, and binary constraint HMC models.

Parts of this work have been published in [8]. In this paper, we extend that work by providing: (i) the sparse basic model with l_1 norm; (ii) a new application of DAG structured class hierarchy in a visual recognition dataset based on deep learning features; (ii) the sensitivity analysis of all the parameters; (iii) the performance of two more kernel functions (Laplace kernel and Polynomial kernel) in the kernel model; and (iv) statistical analysis of all the experimental results.

The rest of this paper is organized as follows: in Section 2 we discuss related work. Section 3 describes the proposed FAEL models. The experimental design, results and analysis are presented in Section 4. Section 5 concludes the paper.

2. Related work

In this section, we review the most recent works in HMC and flat multi-label classification, especially those that are related to our work. Also, we illustrate how our framework is different from previous ones.

In HMC, Both global and local approaches have been developed. Most global approaches are extended from classic single label machine learning algorithms. Wang et al. [9] used association rules for hierarchical document categorization. Hierarchical relationships between different classes are defined based on the similarity of the documents belonging to them. Vens et al. [10] introduced a modified version of decision tree for HMC. One tree is learned to predict all the classes at once. Bi et al. [11] formulated the HMC as a graph problem of finding the best subgraph in a tree or DAG. Kernel dependency estimation is used to reduce the original hierarchy to a manageable number of single label learning problems. A generalized condensing sort and select algorithm is applied to preserve the parent-child relationships in the label hierarchy. Based on a predictive clustering tree, Dimitrovski et al. [2] proposed the cluster-HMC algorithm for medical image annotation. In another work [12], Dimitrovski et al. introduced ensembles of predictive clustering trees for hierarchical classification of diatom images. Bagging and random forests are used to combine the predictions of different trees. Cerri et al. [13] introduced genetic algorithm to HMC. Genetic algorithm is used to evolve the antecedents of classification rules. A set of optimized antecedents is selected to make a prediction for the corresponding classes. Barros et al. [14] introduced the probabilistic clustering HMC framework for protein function problem. The assumption is that training instances can fit to several probability distributions, where instances from the same distribution also share similar class vectors. The major drawback of previous global models is that they ignore the local modularity in the label hierarchy, such as parent-child, ancestor-descendent, and sibling relationships between different labels.

Local approaches also draw heavy attention. Dumais and Chen [15] applied a multiplicative threshold to update local prediction. The posterior probability is computed based on the parent-child relationship. Barutcuoglu and DeCoro [16] proposed a Bayesian aggregation model for image shape classification. The main idea is to obtain the most probable consistent set of global predictions. Cesa-Bianchi et al. [17] developed a top down HMC method using hierarchical Support Vector Machine (SVM), where SVM learning is applied to a node only if its parent has been labeled as positive. Alaydie et al. [18] introduced hierarchical multi-label boosting with label dependency. The pre-defined label hierarchy is used to decide the training set for each classifier. The dependencies of the children are analyzed using Bayesian method and instance based similarity. Ren et al. [19] proposed to address the HMC problem for documents in social text streams with Structural SVM (S-SVM). Multiple structural classifiers are built for each chunk of classes to overcome the unbalanced sample problem. Cerri et al. [20] proposed to build multi-layer perceptron for each level of labels in the label hierarchy. The predictions made by a given level are used as inputs to the next level. Vateekul et al. [21] introduced a hierarchical R-SVM system for gene function prediction. The threshold adjustment from R-SVM is used to mitigate the problem of false negatives in HMC. Valentini [22,23] presented the True Path Rule (TPR) ensembles. In this method, positive local predictions of child nodes affect their parent nodes and negative local predictions of non-leaf nodes affect their descendant nodes.

Our work is inspired by both top-down and bottom-up local models. The top-down models propagate predictions from high level nodes to the bottom [15,24]. In contrast, the bottom-up models propagate predictions from the bottom to the whole hierarchy [25,26]. As a state-of-the-art method, the TPR ensemble integrates both top-down and bottom-up rules [22]. The global prediction of each parent node is updated by the positive local predictions of its child nodes. Then, a top-down rule is applied to synchronize the obtained global predictions. The method is also extended to handle DAG structured class hierarchy [4,23]. In contrast to TPR, our model incorporates all pairs of hierarchical relationships and attempts to learn a fully associative weight matrix from training data. Take the “human” sub-hierarchy from the extended IAPR TC-12 image dataset [27] for example. Fig. 1 depicts the merits of our model and shows the contribution of hierarchical and sibling nodes on each local prediction. The weight matrix computed shows that each local node influences its own decision positively, while nodes not directly connected in the hierarchy provide a negative influence. Since the weight matrix of our model is learned based on all the training samples, we can minimize the influence of outlier examples of each node. The learning model also helps to avoid the error propagation problem, because all the global predictions are obtained simultaneously.

Many works have also been proposed for flat multi-label classification, where no specific hierarchical relationships between labels are given. Because multiple labels share the same input space and semantics conveyed by different labels are usually correlated, it is essential to exploit the correlation information contained in different labels by a multi-task learning framework. Ji et al. [28] developed a general multi-task framework for extracting shared structures in multi-label classification. The optimal solution to the proposed formulation is obtained by solving a generalized eigenvalue problem. Zhu et al. [29] proposed a multi-view multi-label framework with block-row regularization. The regularizer concatenates a Frobenius norm regularizer and l_{21} norm regularizer, which are used to select informative views and features. To handle the missing label problem, semi-supervised learning was introduced to multi-label classification. Luo et al. [30] proposed a manifold regularized multi-task learning algorithm. A discriminative subspace shared by multiple classification tasks is learned while manifold regularization ensures that the learned predictive structure is reliable for both labeled data and unlabeled data. In another work, Luo et al. [31] developed a multi-view matrix completion framework for semi-supervised multi-label image classification. A cross-validation strategy is used to learn combination coefficients of different views. Inspired by the great success of deep Convolutional Neural Networks (CNN) in single label image classification in the past few years [32–34], CNN-based multi-label image classification algorithms were also developed. Wei et al. [35] proposed a hypotheses CNN pooling framework. Different object segment hypotheses are taken as inputs of a shared CNN. The CNN output re-