Information Sciences xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Information Sciences

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

An approach for multi-label classification by directed acyclic graph with label correlation maximization

Q1 Jaedong Lee[a], Heera Kim[b], Noo-ri Kim[a], Jee-hyong Lee[a,*]

[a] Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon 16419, Republic of Korea

Q2 [b] Department of Platform Software, Sungkyunkwan University, Republic of Korea

ARTICLE INFO

Article history:
Received 7 August 2015
Revised 12 January 2016
Accepted 18 February 2016
Available online xxx

Keywords:
Multi-label learning
Directed acyclic graph
Bayesian network
Q3 Conditional entropy

ABSTRACT

Traditional supervised learning approaches primarily work in the single-label environment. However, in many real-world problems, data instances are usually associated with multiple labels simultaneously, and multi-label learning is increasingly required in many modern applications. In multi-label learning, the key to successful classification is effectively exploiting the complex correlations among the output labels. This paper proposes a novel multi-label learning method inspired by the classifier chain approach. The main contribution of this work is to model the correlations of the labels using a directed acyclic graph. Starting from the simple intuitive notion of measuring the correlations among the labels, the proposed method is designed as a multi-label learning method that maximizes the correlations among labels. To evaluate its effectiveness, the proposed method is compared with the state-of-the-art approaches. Extensive experiments demonstrated the proposed method to be highly competitive with the other multi-label approaches.

1. Introduction

Traditional supervised learning approaches primarily work in the single-label environment, where an instance is associated with only a single label. However, in many real-world conditions, data instances are usually associated with multiple labels simultaneously. Thus, multi-label learning is being increasingly required in many modern applications. For example, an image in an image-tagging system could have conceptual labels of both “tree” and “forest” [3,17,26]. Similarly, in text categorization, a document can belong to multiple categories or have multiple tags [1,24].

The goal of single-label learning is to find a model $\boldsymbol{h}$ that maps input $\mathbf{x}$ to a scalar output y. In contrast, in multi-label learning, the problem is to find a model $\boldsymbol{h}$ that maps input $\mathbf{x}$ to a vector output $\mathbf{y} = (y_1, y_2, \ldots, y_L)$. If the values of y_i are uncorrelated, that is, if knowing the values of y_j for $j \neq i$ is not helpful for prediction of y_i, then multi-label learning is equivalent to L single-label learnings in a straightforward way. In this case, the classifier for each label can be independently built. However, output labels are correlated with each other in most cases, which means that y_i can be better predicted if some values of y_j for $j \neq i$ are known. This introduces complexity that makes the task of multi-label learning rather challenging. This viewpoint has motivated much multi-label learning research, which has led to the development of various methods to utilize correlation information to improve the performance of classifiers [4,6,7,10,15,20,22,23,25,28,29,30,31,32].

The existing multi-label learning approaches can be grouped into two categories: label power set (LP) approaches and binary relevance (BR) approaches. In LP approaches [16,20,21,27,28,30], a multi-label problem is transformed into a

Q4 * Corresponding author. Tel.: +82 31 290 7154.

Q5 *E-mail addresses:* ultrajaepo@skku.edu (J. Lee), hrahoy@skku.edu (H. Kim), pd99j@skku.edu (N.-r. Kim), john@skku.edu (J.-h. Lee).

http://dx.doi.org/10.1016/j.ins.2016.02.037
0020-0255/

Table 1
Example of multi-label data.

Inputs	Outputs
$\mathbf{x}_1 = [1, 0, 1, 1, 0, 0, 1]$	$\mathbf{y}_1 = [1, 0, 0, 1, 0]$
$\mathbf{x}_2 = [0, 0, 1, 1, 0, 1, 0]$	$\mathbf{y}_2 = [0, 1, 0, 1, 1]$
$\mathbf{x}_3 = [1, 0, 0, 1, 1, 1, 1]$	$\mathbf{y}_3 = [1, 0, 1, 0, 1]$

multi-class single-label problem by treating an output vector as a scalar value. Then, the problem can be directly solved using any single-label learning approach. LP approaches have the advantage that the correlations of labels can be embedded in scalar values and learned by single-label learners, but they also have some drawbacks: high computational complexity and overfitting to the training data [28]. To overcome these disadvantages, several methods have been proposed based on randomized approaches [16,20,21,27,28,30].

Additionally, BR is another common approach, which transforms a multi-label problem into multiple independent single-label classification problems. In BR, all labels are assumed to be independent of themselves, and each classifier is independently trained to predict each label based on only given inputs. However, this assumption of label independence is not suitable for real-world multi-label dataset. So, in order to consider the correlations of labels, there were some variations of BR that were proposed.

The classifier chain (CC) approach based on BR was introduced by Read et al. [22]. They proposed the idea of "chaining" classifiers to take the label correlations into account. This method involved L single-label classifiers linked along a chain in a specific order; for example, $\{y_2 \rightarrow y_5 \rightarrow \ldots \rightarrow y_1\}$. The inputs of each classifier in a chain were extended with the outputs of all the preceding classifiers. CC approaches can properly model correlations among labels and have been shown to be competitive in terms of their computational efficiency [8]. However, they have a critical drawback in that it is very hard to determine the proper order of classifier chains, which led to the decision to employ a random-order scheme [7,11,14,23,25]. Many extensions of CC approaches have been proposed to deal with the ordering problem. Read et al. proposed ensemble classifier chains (ECC) in an effort to reduce the influence of bad random chain order by adopting a simple ensemble framework [23]; Dembczynski et al. introduced a probabilistic method to estimate the entire joint distribution of the label set in the classification phase [7]. Silva et al. built several randomly ordered chains and then tried to find the best one for a test instance using a brute-force method [24]. Gonçalves et al. adopted a genetic algorithm on CC for label ordering optimization Q6 [11]. However, CC approaches still have limitations. Previous work has focused on avoiding bad ordered chains rather than finding the optimal order to improve prediction performance.

Successful multi-label learning requires effectively exploiting the complex correlations among the output labels. We thus propose a novel multi-label learning method inspired by classifier chain approaches. Our proposed method overcomes some limitations of the existing multi-label classification approaches by finding the order of the classifiers that maximizes the correlations between the labels. The main contribution of this work is to build a directed acyclic graph (DAG) of labels in which the correlations between parent and child nodes are maximized. We quantify the correlations with the conditional entropy and find a DAG that maximizes the sum of conditional entropies between all parent and child nodes. Thus, highly correlated labels can be sequentially ordered in chains obtained from the DAG, and the prediction results can be optimized when utilizing a CC approach with the chains. What is interesting is that finding such a DAG is very similar to the Bayesian network learning problem. Thus, we can solve this problem by using one of several Bayesian network learning methods, such as the K2 algorithm.

The rest of this paper is organized as follows. Section 2 introduces the multi-label classification approach and discusses related work. Section 3 presents a detailed description of the proposed multi-label learning algorithm. Section 4 compares the proposed multi-label learning method to well-known multi-label classification algorithms. The multi-label dataset and evaluation measures used for experiments are also described in detail. Finally, Section 5 highlights the main points of this work and concludes this paper.

2. Related work

In this section, label power set (LP) approaches and binary relevance (BR) approaches for the multi-label learning problem are summarized.

2.1. Label power set approach

An instance in multi-label data is composed of a vector of inputs $\mathbf{x}$ and a vector of outputs $\mathbf{y}$. Table 1 lists 3 instances of multi-label data. There are 7 inputs and 5 outputs: $\mathbf{x} = (x_1, x_2, \ldots, x_7)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_5)$. The goal of multi-label learning is to find a model that maps an input vector $\mathbf{x}$ to a vector output $\mathbf{y}$.

In the LP approach, a multi-label problem is transformed into a multi-class single-label problem by treating an output vector as a scalar value. Then, the problem is to find a model $\boldsymbol{h}$ that maps an input vector $\mathbf{x}$ to a scalar output z representing an output vector $\mathbf{y}$. For example, $\mathbf{y}_1$ in Table 1 can be transformed into a binary number $z_1 = 10010$, and similarly, $\mathbf{y}_2$ into

Please cite this article as: J. Lee et al., An approach for multi-label classification by directed acyclic graph with label correlation maximization, Information Sciences (2016), http://dx.doi.org/10.1016/j.ins.2016.02.037