Pattern Recognition 47 (2014) 3405–3413

Contents lists available at ScienceDirect

Pattern Recognition

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

Enhancing multi-label classification by modeling dependencies among labels

Shangfei Wang [a,*], Jun Wang [a], Zhaoyu Wang [a], Qiang Ji [b]

[a] Key Lab of Computing and Communication Software of Anhui Province, School of Computer Science and Technology, University of Science and Technology of China, Hefei, Anhui 230027, PR China
[b] Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180, USA

A R T I C L E I N F O

Article history:
Received 2 July 2013
Received in revised form
4 April 2014
Accepted 8 April 2014
Available online 16 April 2014

Keywords:
Multi-label classification
Bayesian network
Structure learning
Incomplete label assignments
Maximum likelihood estimation

A B S T R A C T

In this paper, we propose a novel framework for multi-label classification, which directly models the dependencies among labels using a Bayesian network. Each node of the Bayesian network represents a label, and the links and conditional probabilities capture the probabilistic dependencies among multiple labels. We employ our Bayesian network structure learning method, which guarantees to find the global optimum structure, independent of the initial structure. After structure learning, maximum likelihood estimation is used to learn the conditional probabilities among nodes. Any current multi-label classifier can be employed to obtain the measurements of labels. Then, using the learned Bayesian network, the true labels are inferred by combining the relationship among labels with the labels' estimates obtained from a current multi-labeling method. We further extend the proposed multi-label classification method to deal with incomplete label assignments. Structural Expectation-Maximization algorithm is adopted for both structure and parameter learning. Experimental results on two benchmark multi-label databases show that our approach can effectively capture the co-occurrent and the mutual exclusive relation among labels. The relation modeled by our approach is more flexible than the pairwise or fixed subset labels captured by current multi-label learning methods. Thus, our approach improves the performance over current multi-label classifiers. Furthermore, our approach demonstrates its robustness to incomplete multi-label classification.

1. Introduction

Multi-label classification is a classification problem where one sample can be assigned with more than one target labels simultaneously. There are many multi-label learning applications. For example, a piece of music may be characterized by both dreamy and cheerful [1]. An image may include grass, cow and sky [2]. Hence, a data sample (image or music) may simultaneously contain multiple different labels that characterize different properties of the data.

Usually the labels are dependent on each other. Take music emotional tagging for example, some emotions may appear together frequently, while others may not. A piece of music may induce the feelings of relaxing, comfortable and happy, but it rarely induces disgust at the same time. Such dependencies among labels are one of the key issues in multi-label learning. Current research can be divided into three groups: ignoring dependencies, exploring dependencies directly only from labels, and exploring label dependencies indirectly with the help of features or hypotheses. The first group takes no account of the relation among labels, it therefore suffers from unstable performance. The second group considers pairwise relation, or the fixed label combinations present in training data directly from labels without considering features or hypotheses. However, the dependencies among multiple labels are more complex and flexible, beyond pairwise and fixed label combinations. In fact, there are two kinds of relationship: co-existence and mutual exclusion. For example, in music emotional tagging, a piece of sad music may elicit both sadness and anger but rarely happiness, which reflects the co-existent relation between sadness and anger, and the mutual exclusive relation between sadness and happiness. Thus, the second group cannot fully explore the feasible dependencies among labels. The third group can model more feasible label relation with the aid of features and hypotheses. However, its computation cost is much higher than the first two groups.

Furthermore, since annotating labels is time confusing and require expertise, labels may be missing for some applications. For example, because of difficulty with annotating certain labels,

* Corresponding author. Tel.: +86 551 63602824.
E-mail addresses: sfwang@ustc.edu.cn (S. Wang), junwong@mail.ustc.edu.cn (J. Wang), wazhy@mail.ustc.edu.cn (Z. Wang), qji@ecse.rpi.edu (Q. Ji).

http://dx.doi.org/10.1016/j.patcog.2014.04.009
0031-3203/

annotators may only provide the main emotion of a piece of music, and users may tag an image just with several main objects. Therefore, learning from incomplete labels is another key issue for multi-label classification. However, current multi-label classification research rarely addresses learning from incomplete labels.

In this paper, we propose a Bayesian Network (BN) to systematically capture the dependencies among different labels directly. The nodes of the BN represent the labels. The links and their parameters capture the probabilistic relation among labels. Our structure learning algorithm [3] is employed to learn the BN structure. By exploiting the decomposable property of the Bayesian Information Criterion (BIC) score function, the algorithm significantly reduces the search space of possible structures and guarantees the global optimality. After structure learning, the conditional probabilities are directly learned on the training data. Then, we can infer the true labels by instantiating the measurement nodes with the labels' estimates obtained from a traditional multi-labeling method. The experimental results on two multi-label datasets show that both the co-occurrent and the mutual exclusive relation among labels can be effectively captured by our structure learning algorithm. The relation modeled by our approach is more flexible than pairwise or fixed subset labels captured by current multi-label learning methods, it improves the performance of current multi-label classifiers which model label dependence directly. Furthermore, we extend our approach to deal with incomplete labels by using structural Expectation-Maximization (EM) algorithm. The experimental results on the same two multi-label datasets show the advantage of our method.

2. Related work

Multi-label classification methods can be categorized into two groups: problem transformation methods and algorithm adaptation methods. The former includes Binary Relevance (BR) [4], Label Powerset (LP) [4], and Random k labelsets (RAkEL) [5]. They transform the multi-label classification task into one or more single-label classification tasks, and then any traditional classification algorithms can be used. The latter consists of Binary Relevance k Nearest Neighbors (BRkNN) [6], Multi-Label k Nearest Neighbors (MLkNN) [7], AdaBoost.MH [8], etc. They extend specific learning algorithms to handle multi-label data directly. A comprehensive overview of current research in multi-label classification can be found in [9,10].

Due to the large number of possible label sets, multi-label classification is rather challenging. Successfully exploiting the dependencies inherent in multiple labels is the key to facilitate the learning process. Considering dependencies among labels, most present multi-label learning strategies can be categorized into three groups: methods ignoring label correlation, methods considering label correlation directly, and methods considering label correlation indirectly. The first group (i.e., BR [4]) decomposes multi-label problem into multiple independent binary classification problems (one per category). By ignoring the correlation among labels, the generalization ability of such method may be weak. The second group addresses the pairwise relation between labels (such as Calibrated Label Ranking (CLR)), or the fixed label combinations existing in training data (such as LP), or a random subset of the combinations (such as RAkEL)). However, the relation among labels may be beyond pairwise, and cannot be expressed by a fixed subset of labels existing in training data. Besides, the number of the pairwise subsets increases exponentially when the number of the labels is quite large. Meanwhile, there may not be sufficient training data when there are few instances for the combined labels. Thus, the second group may not capture the label relation effectively. The third group considers label dependencies with the help of features or hypothesis. Godbole and Sarawagi [11] stacked the outputs of BR along with the full original feature space into a separate meta classifier, creating a two-stage classification process. Read et al. [12] proposed the classifier chain model to link n classifier into a chain. The feature space of each classifier in the chain is extended with the label associations of all previous classifiers. Ghamrawi and McCallum [13] adopted conditional random field to capture the impact of an individual feature on the co-occurrence probability of a pair of labels. Sun et al. [14] proposed to construct a hyperedge for each label, and include all instances annotated with a common label into one hyperedge, thus capturing their joint similarity. Zhangs [15] proposed Bayesian Network to model the dependencies among label errors, and then a binary classifier was constructed for each label combining the features and the parental labels, which were regarded as additional features. Huang et al. [16] modeled the label relation by a hypothesis reuse process. When the classifier of a certain label is learned, all trained hypotheses generated for other labels are taken into account via weighted combinations. These methods can model the flexible dependencies among labels to some extent, but their computation costs are usually much higher compared with the second group.

Among the above, Zhangs' work is the most similar one to ours. They proposed to use a Bayesian Network (BN) structure to encode the conditional dependencies of labels as well as the feature set: $P(\lambda_1, \lambda_2, \dots, \lambda_n | x)$, where x is the features and $(\lambda_1, \lambda_2, \dots, \lambda_n)$ is the multiple target labels, n is the number of labels. Since they thought directly modeling $P(\lambda_1, \lambda_2, \dots, \lambda_n | x)$ by Bayesian approach was intractable, they adopted an approximate method to model the dependencies among label errors, which was independent of features x. Based on the learned BN structure of errors, a binary classifier was constructed for each label λ_i combining the features x and the parental labels $pa(\lambda_i)$, which were regarded as additional features.

Unlike Zhangs' method, we propose a Bayesian Network to systematically capture the dependencies among different labels, $P(\lambda_1, \dots, \lambda_n)$, directly. The nodes of the BN represent the labels. The links and their parameters capture the probabilistic relation among labels. Our BN structure learning algorithm [3] is adopted. After structure learning, maximum likelihood estimation is used to learn the conditional probabilities. Then, we can infer the true labels by instantiating the measurement nodes with the labels' estimates obtained from a traditional multi-labeling method.

Compared to Zhangs' method, we first directly capture the dependencies among labels. Then we obtain label measurements, $M\lambda_i$, using any multi-label classifier. After that, we infer an instance's multiple labels simultaneously using Most Probable Explanation (MPE) inference: $P(\lambda_1, \dots, \lambda_n | M\lambda_1, \dots, M\lambda_n) = P(M\lambda_1, \dots, M\lambda_n | \lambda_1, \dots, \lambda_n) P(\lambda_1, \dots, \lambda_n) / P(M\lambda_1, \dots, M\lambda_n)$. Thus, our approach can explicitly model the co-existent and the mutual exclusive relation among labels, instead of the errors of the labels. Besides, our approach can infer the multiple labels of an instance simultaneously, not recognize each label separately. Our approach can be easily combined with any multi-label classifier to enhance its performance.

Current multi-label classification methods require complete label assignments. However, multi-label classification with incomplete label assignments is frequently encountered in realistic scenario, especially when the number of labels is very large. Till now, little research [17] has addressed the challenge of multi-label classification with incomplete labels [18].

In this paper, we propose a BN to systematically capture the dependencies among labels directly. Furthermore, we extend our approach to address multi-label classification with incomplete labels using a structural EM algorithm.