Neurocomputing 000 (2017) 1–10

Contents lists available at ScienceDirect

Neurocomputing

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

Multi-label learning based on label-specific features and local pairwise label correlation

Wei Weng [a,b], Yaojin Lin [c], Shunxiang Wu [a,*], Yuwen Li [a], Yun Kang [a]

[a] Department of Automation, Xiamen University, Xiamen 361000, PR China
[b] College of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361024, PR China
[c] School of Computer Science, Minnan Normal University, Zhangzhou 363000, PR China

ARTICLE INFO

Article history:
Received 2 September 2016
Revised 21 April 2017
Accepted 18 July 2017
Available online xxx

Communicated by Jiayu Zhou

Keywords:
Multi-label learning
Label-specific features
Local pairwise label correlation

ABSTRACT

Multi-label learning has drawn great attention in recent years. One of its tasks aims to build classification models for the problem where each instance associates with a set of labels. In order to exploit discriminative features for classification, some methods are proposed to construct label-specific features. However, these methods neglect the correlation among labels. In this paper, we propose a new method called LF-LPLC for multi-label learning, which integrates Label-specific features and local pairwise label correlation simultaneously. Firstly, we convert the original feature space to a low dimensional label-specific feature space, and therefore each label has a specific representation of its own. Then, we exploit the local correlation between each pair of labels by means of nearest neighbor techniques. According to the local correlation, the label-specific features of each label are expanded by uniting the related data from other label-specific features. With such a framework, it enriches the labels' semantic information and solves the imbalanced class-distribution problem. Finally, for each label, based on its label-specific features we construct a binary classification algorithm to test unlabeled instances. Comprehensive experiments are conducted on a collection of benchmark data sets. Comparison results with the state-of-the-art approaches validate the competitive performance of our proposed method.

1. Introduction

In traditional supervised learning, each instance belongs to a single class label, which is called single-label classification. However, multi-label data exists in many applications, such as image annotation [1,27,33], text classification [5,23,32], and bioinformatics [4,29]. For example, an image can be annotated with several keywords [1]; a document belongs to multiple related topics, such as Shanghai World Expo, economics and even volunteers [23]; a gene may be attached with a set of functional classes, such as energy, metabolism, and cellular biogenesis [4].

During the past decade, numerous multi-label learning algorithms have been developed, which can be classified into two categories [38]: problem transformation and algorithm adaption. Problem transformation is a common and intuitive approach, which views the problem of multi-label problems as one or more traditional single-label problems. Classical problem transformation algorithms include binary relevance (BR) [1], pruned problem transformation (PPT) [25], label power set (LP) [28], and so on. Algorithm adaption performs learning algorithms on multi-label data directly by extending traditional single-label learning algorithms, such as multi-label informed latent semantic indexing (MLSI) [34], and multi-label dimensionality reduction via dependence maximization (MDDM) [36].

In multi-label learning, the performance of many classification algorithms benefit from the label correlation, i.e., the information of one label may be helpful for learning other related labels. For example, in a library of scenery images, the image tagged with “desert” usually is more likely to be tagged with “camel” than “tree”. To date, a number of multi-label algorithms have been proposed to exploit the relationship among labels [11,12,21,22,37], which utilize the relationship among labels to construct probability-based or optimization-based classification models.

On the other hand, multi-label data usually have high dimensional features [16,19], which contain different concepts related to different labels. Specifically, an instance with a feature vector describes the mixed concepts of multi-labels (a label vector) at the same time. Hence, for each label, the label-specific features will be

* Corresponding author.
E-mail addresses: xmutwei@163.com (W. Weng), yjlin@mnnu.edu.cn (Y. Lin), wsx1009@163.com (S. Wu), Yuwenli2012@163.com (Y. Li), kangyun1225@163.com (Y. Kang).

http://dx.doi.org/10.1016/j.neucom.2017.07.044
0925-2312/© 2017 Published by Elsevier B.V.

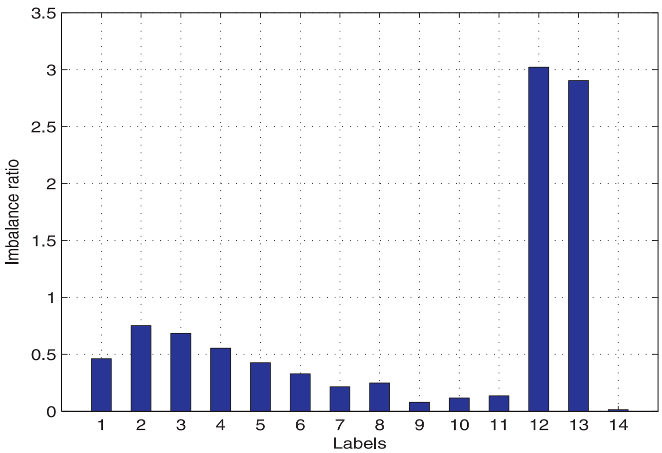

Fig. 1. The imbalance ratio variation of 14 class labels in yeast data set.

much more informative and should be converted from the original features [40]. It is well worth noting that label-specific features conversion is different from traditional feature reduction [3,17,18], where all labels share the same features.

Except from label correlation and label-specific features, multi-label learning also suffers from a high level of class-imbalance [6,8,20,31,38]. That is, the distribution of positive instances and negative instances for each label is usually highly imbalanced and varies widely. It is even worse when the number of labels is large and the label density is low. For example, Fig. 1 shows the imbalance ratio (the number of positive instances vs. the number of negative instances) of each label on the yeast data. From Fig. 1, we can observe that there are 12 out of the 14 class labels whose imbalance ratios are less than 1. Especially, in the 14th label, the number of positive instances is much less than the number of negative instances.

As discussed above, label correlation, label-specific features, and class-imbalance are three characteristics in multi-label learning. However, many learning algorithms study these problems independently. In this paper, we propose a new multi-label learning algorithm named LF-LPLC. The main idea of LF-LPLC is to utilize the label-specific features and the relationship among labels to obtain discriminative information for each label, and eliminate the problem of class-imbalance. Firstly, LF-LPLC constructs a label-specific feature space for each label. Then, for each label, we enlarge the size of positive instances via copying others related positive instances belonging to other correlated labels. Note that the "copying" operation requires the dimensions of these label-specific feature spaces to be identical. Therefore, a process called class-alignment is introduced to preprocess training data before constructing the label-specific feature space. Finally, SVM is trained on each label-specific feature space and used to predict the corresponding label. A comprehensive set of experiments are conducted to verify the performance of LF-LPLC against other state-of-art approaches on eight benchmark data sets.

The rest of this paper is organized as follows. Section 2 introduces the related work. Section 3 presents our proposed algorithm in detail. Comprehensive experimental results are discussed in Section 4. Conclusions and future work are given in Section 5.

2. Related work

At present, a large number of multi-label learning methods have been developed, and the detailed survey can be found in [7,38]. In this section, we will briefly review the related work, i.e., some works about label-specific features and label correlation.

Label-specific features was first proposed in [40], called LIFT, which constructs specific features for each label. At first, for each label, LIFT gets centers of its positive and negative instances via the algorithm of $k-$means [14]. Then, those instances are transformed into new ones, whose features are composed of distances between original instances and clustering centers. These features are called label-specific features. Note that the number of centers are related to the number of positive and negative instances. If the training set is class-imbalanced, the number of centers will be different among different labels. So, the dimensions of label-specific feature space are usually different because of class-imbalance. At last, SVM is applied on the process of training and test. LIFT shows an intuitive implementation of label-specific features construction, and related experiments show its excellent classification effect. Zhang et al. [39] figured out that LIFT does not utilize the discriminative information lying between positive and negative instances, and proposed a algorithm named ML-DFL. ML-DFL constructs a matrix whose elements represent the similarities based on distances between positive and negative instances. Then, it adapts a spectral clustering algorithm [24] to exploit the closely located local structures beneath positive and negative instances. Different from LIFT and ML-DFL, Huang et al. [13] utilized feature selection to obtain label-specific features and proposed the algorithm called LLSF. In LLSF, linear regression is applied to get the weights of features for each label. If the weight of some features is zero for a label, it has no effect on the discrimination of that label. Furthermore, in order to incorporate label correlation, LLSF requires that strongly correlated labels should have large similarity between their weight vectors.

To exploit the correlation among labels, a number of multi-label algorithms are presented. For example, Zhang and Zhang [37] used a Bayesian network structure to encode the conditional dependencies of labels, and constructed classifiers for each label by incorporating its parental labels as additional features. This method considers that the label correlations are shared by all instances. Huang and Zhou [11] expanded the original features by adding a code vector for each instance. It firstly divides training data into several groups based on clustering. Then, it generates the prototype of label vectors for each group. The code vector is composed of the similarities between the label vector and those prototypes. Huang et al. [10] considered that if two labels are related, the classifier model generated for one label can be helpful for the other label, and the greater the help, the stronger the relationship. Huang et al. [12] proposed a method to exploit local pairwise label correlations. For a given instance, it utilizes the co-occurrence frequency of two ground true labels in the k nearest neighbors to indicate their relationships. The higher the frequency, the stronger the relationship. It maximizes the posterior probability to predict unseen instances, which accords to the distribution of each label in the k nearest neighbors and their strongest local pairwise label correlations.

3. The proposed algorithm

In multi-label learning, let $\mathcal{X} = \mathbb{R}^d$ be the domain of instances and let $\mathcal{Y} = \{l_1, l_2, \dots, l_Q\}$ denote the finite set of labels. $D = \{(x_i, y_i) | 1 \le i \le N, x_i \in \mathcal{X}, y_i \subseteq \mathcal{Y}\}$ denotes the training data that consists of N instances and its related labels. For convenience, y_i is often written as a vector consisting of +1 and −1. That is, $y_i \in \{-1, +1\}^{|\mathcal{Y}|}$ is used to identify whether a label is assigned to the instance x_i. $y_{ij} = +1 (1 \le j \le Q)$ means that l_j belongs to the instance x_i; otherwise $y_{ij} = -1$. Instances associating with a given label are considered as positive instances, otherwise, are considered as negative ones. For each label l_k, the set of positive instances are denoted as $\mathcal{P}_k$ and the set of negative instances are denoted as $\mathcal{N}_k$. The goal of multi-label learning is to define a set of real value functions $f_i : \mathcal{X} \to \mathbb{R} (i = 1, 2, \dots, Q)$. The bigger the value of f_i means