Neurocomputing 154 (2015) 305–316

Contents lists available at ScienceDirect

Neurocomputing

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

Multi-label learning with discriminative features for each label

Ju-Jie Zhang, Min Fang*, Xiao Li

School of Computer Science and Technology, Xidian University, Xi'an 710071, China

ARTICLE INFO

Article history:
Received 12 May 2014
Received in revised form
10 October 2014
Accepted 26 November 2014
Communicated by Tao Mei
Available online 7 December 2014

Keywords:
Multi-label learning
Multi-label classification
Spectral clustering
Problem transformation

ABSTRACT

During the last decade, multi-label learning has attracted the attention of more and more researchers in machine learning field due to wide real-world applications. Existing approaches often predict an unseen example for all labels based on the same feature vector. However, this strategy might be suboptimal since different labels usually depend on different aspects of the feature vector. Furthermore, for each label there is close relationship between positive and negative instances, which is quite informative for classification. In this paper, we propose a new algorithm called ML-DFL, which trains a model for each label with newly constructed discriminative features. In order to form these features, we also propose a spectral clustering algorithm SIA to find the closely located local structures between positive and negative instances, which are assumed to be of more discriminative information, and then transform the original data set by consulting the clustering results in a simple but effective way. Comprehensive experiments are conducted on a collection of benchmark data sets. The results clearly validate the superiority of ML-DFL to various competitors.

1. Introduction

Multi-label learning or multi-label classification deals with learning problems where each instance may be associated with multiple labels simultaneously. For example, a document may be associated with several given topics, including *entertainment* and *sports* [1]; a gene has a few functional classes, such as *energy* and *cellular biogenesis* [2]; an image can be annotated as *sun* or *sea* [3]. This is in contrast with traditional single-label (i.e. binary or multiclass) learning problems where each instance is only relevant to a single label.

Over the last decade, more and more researchers are engaged in studying multi-label learning problems due to its wide real-world applications [1–3]. Substantial multi-label learning approaches have been proposed [4]. Most approaches learn one classifier for each label directly based on the original instance $\mathbf{x}$ under some assumption, such as a shared subspace underlying multiple labels [5] or label sparsity [6]. These algorithms have achieved great success, but they may be suboptimal in the multi-label settings.

Different labels have distinct characteristics of their own. Each label highly depends on some specific aspect of the original instance. Zhang [7] has proven the effectiveness of constructing label-specific features for each label. For each label, LIFT performs clustering analysis in the positive and negative instances respectively, and then constructs label-specific features by checking the distance between the original instance and the clustering centres. However, it does not utilize the discriminative information that lies between positive and negative instances. A portion of positive instances may be close to a portion of negative ones, which has significant impact on the performance of multi-label learning algorithm. Therefore, we conjecture that if we can locate the close latent structures between positive and negative instances, a more effective algorithm can induced.

To justify our conjecture, we propose a new algorithm called Multi-label Learning with Discriminative Features for each Label (ML-DFL), which is based on a new spectral clustering algorithm, Spectral Instance Alignment (SIA). For each label, SIA constructs the similarity matrix between positive and negative instances and then applies spectral clustering analysis to extract the closely located latent structures between them. To highlight the difference, ML-DFL calculates the distances between each original instance and the clustering results, and treats them as new features which are expected to be more discriminative than the original instance. Then we train the lth classifier based on new features. To validate the effectiveness of ML-DFL, we conduct comprehensive experiments on a collection of benchmark data sets. Experimental study shows clear advantage of ML-DFL over various competitors.

The rest of this paper is organized as follows: Section 2 gives a brief literature review on multi-label learning. Section 3 describes our proposed algorithm ML-DFL in detail. We present experimental settings, data sets and evaluation criteria in Section 4. Section 5 contains detailed experiments and discussion. Finally, Section 6 concludes this paper.

* Corresponding author. Tel.: +86 15029954025.
E-mail addresses: jujiezhang@stu.xidian.edu.cn (J.-J. Zhang), mfang@mail.xidian.edu.cn (M. Fang), xiao_li@stu.xidian.edu.cn (X. Li).

http://dx.doi.org/10.1016/j.neucom.2014.11.062
0925-2312/© 2014 Elsevier B.V. All rights reserved.

2. Related work

Before embarking on a formal description of our proposed algorithm ML-DFL, we first review related work on multi-label learning. Generally speaking, existing multi-label learning approaches can be categorized into two classes: *problem transformation methods* and *algorithm adaptation methods.*

2.1. Problem transformation methods

The common strategy adopted by these methods is to transform a multi-label problem into one or more single-label classification problems. Then many existing single-label algorithms can be employed, such as support vector machines (SVM) [3], k nearest neighbour (kNN) [8], and Naive Bayes [9]. The final prediction of an unseen instance is formed by combining the predictions of all single-label classifiers. This strategy is attractive since it is flexible to combine existing single-label algorithms and facilitates the algorithm design.

Binary relevance (BR) [3], the most straightforward and intuitive method, is to decompose a multi-label problem into L independent single-label sub-problems. AdaBoost.MH [1] decomposes a multi-label problem into L independent binary classification problems each of which is handled via AdaBoost [11]. The main limitation of BR and AdaBoost.MH is the ignorance of label correlations since it is well believed that there is substantial and complex relationship between labels. It is the label relationship that distinguishes multi-label problems from classical binary or multiclass ones. LIFT [7] trains the classifier for the lth label with new training examples constructed via performing clustering on positive and negative instances separately. It needs considering the relationship between positive and negative instances for performance improvement. BRkNN [10] combines BR and k nearest neighbour method together. When testing a new instance, it first searches for the k nearest neighbours and decides the final prediction for each label. It suffers from high dimensionality of feature space and huge memory burden. [12] proposes a modified one-against-one SVM classifier for multi-label text categorization using the SVM's predictions and probability, which is computationally expensive.

Ranking by pairwise comparison (RPC) [13] takes pairwise correlations between labels into account. However, it is highly dependent on a “zero” point to separate relevant labels from irrelevant labels when predicting. To overcome this limitation, calibrated label ranking (CLR) [14] is proposed. CLR uses BR to learn the “zero” point. However, both RPC and CLR consider at most the relationship between each pair of labels while neglecting higher order correlations, i.e. correlations among three or more labels. Two stage architecture (TSA) [15] adopts a two-stage architecture to utilize pairwise relationship between labels. In the first stage, it learns one classifier for each label based on BR, and in the second stage, it combines each pair of labels for further training according to their prediction probability in the first stage. Although it reduces the number of models in CLR, it is still time consuming.

Label power-set (LP) [16] (or label combination [17]) is a simple yet effective method that treats each distinct label combination as a new unique label, thus transforming the multi-label problem into a multiclass problem. However, when the relationship is rather complex or there are plenty of possible labels, LP fails due to the high time complexity of training and predicting. To alleviate the drawbacks of LP, two ensemble approaches are proposed: each base classifier of random k-labelset (RAkEL) [18] focuses on random k labels out of L labels, while that of ensemble of pruned sets (EPS) [17] focuses on frequent label combinations which are found by pruning technique. Both of them suffer from high model complexity. Ensemble of classifier chains (ECC) [19] is also an ensemble method each base classifier of which is built on the predictions of all previous learned models along the classifier chains. One main limitation is that it treats the predictions the same as attributes in $\mathbf{x}_i$. Another limitation is that CC searches along a random label order in which the former few predictions may have negative impact on the later predictions. Probabilistic classifier chains (PCC) [20] exhaustively searches all possible paths to find the most confident label combinations. It cannot handle multi-label problems with a large number of labels as a result of its tremendous complexity.

2.2. Algorithm adaptation methods

Algorithms of this kind are formed by adapting single-label algorithms to multi-label cases. They are trained like traditional single-label algorithms and directly give the predictions of a new instance. The main attractive point of these methods is that they can employ the characteristics of a multi-label learning problem in a more simple, concise and elegant way. Usually these methods utilize the second order (pairwise) or higher-order label correlations.

Ranking support vector machine (RankSVM) [21] modifies SVM to multi-label cases by only considering pairwise labels. It assumes that the decision value for a relevant label should be ranked higher than that for an irrelevant one. RankSVM is computationally expensive because the number of constraints is $O(NL^2)$. To combat this problem, an efficient learning algorithm, named maximum margin multi-label ranking (M3LR) [22], is developed. M3LR relaxes the constraints and thus reduces the computation time drastically. Both RankSVM and M3LR are ranking-based methods which require a “zero” point to separate relevant labels from irrelevant ones like RPC when making predictions. Thus their predicting accuracy is highly dependent on user-specified threshold. SVM-ML [23] introduces a virtual “zero” label to automatically learn such a threshold. It reduces the complex objective in RankSVM to L sub-problems, each only involving the virtual label and one true label. AdaBoost.MR [1] deals with the ranking problem between relevant and irrelevant labels using boosting technique. Xu [24] proposes a novel method for multi-label learning by combining binary SVM and a ranking loss to combat against the limitation of the traditional one-versus-rest method. The ranking-based methods mentioned above are all hindered by forbidding computational cost. Another drawback is their ignorance of higher-order label relationship.

Multi-label k nearest neighbour (ML-kNN) [8] extends traditional k nearest neighbour method to multi-label cases by first identifying the k nearest neighbour examples and then employing maximum a posteriori (MAP) to make predictions. ML-kNN is a lazy learning method which is not applicable in problems with a large training set. Multi-label learning based on shared subspace (ML-SS) [5] assumes that all labels share a common subspace, which is extended from the multi-task setting [25]. However, the assumption may not hold in real cases since usually only several labels share a common subspace. Multi-label hypothesis reuse (MAHR) [26] uses boosting technique generally and presumes that only linear relationship exists among labels which may decrease the performance when there are substantial complicated relationships. Instance-based learning by logistic regression (IBLR) [27], which combines instance-based learning algorithms and logistic regression together to exploit label correlations and interdependencies, is not application for large-scale problems since it needs to hold the whole training set in memory. Sun et al. [28] uses hyper-graph to model the higher order relatedness between labels and employs hyper-graph spectral learning formulation for multi-label learning which can be resolved by the least squares method under mild conditions. Though simple, its performance is unsatisfactory since it is based on least squares.