Information Sciences 422 (2018) 336–351

Contents lists available at ScienceDirect

Information Sciences

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

Robust non-negative sparse graph for semi-supervised multi-label learning with missing labels

Jianghong Ma, Tommy W.S. Chow*

Department of Electronic Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong

ARTICLE INFO

Article history:
Received 22 February 2017
Revised 10 July 2017
Accepted 18 August 2017
Available online 30 August 2017

Keywords:
Label recovery
Semi-supervised setting
Missing labels
Semantic structure
Semantic correlation

ABSTRACT

In multi-label learning, each instance is assumed to belong to multiple nonexclusive classes among a finite number of candidate categories. Labels are related to certain conceptual space according to their semantic similarities. Most existing approaches that deal with missing labels have the limitations in mining interdependencies among labels in the original incomplete label matrix with missing labels. In addition, semantic gaps are often neglected when features are used to facilitate label recovery. In this paper, we propose a novel label recovery method under a semi-supervised setting. The proposed method can perform label matrix imputation in the labeled space and label matrix prediction in the unlabeled space simultaneously. The semantic structure (label relationships among different instances) and the semantic correlation (label relationships among different labels) are also exploited to increase the robustness to semantic gaps and unreliable label correlations respectively. In formulating the objective function, l_1-norm and nonnegative constraints are utilized to capture hidden relational graphs in semantic level and to reveal the annotation structure. An iterative mechanism is introduced to assure all variables are reliable. Intensive simulations were conducted and compared with five widely used multi-label datasets. Obtained results show that the proposed method can achieve highly competitive performance compared to other state-of-art methods.

1. Introduction

We have witnessed the era of Big data and how it has been reshaping the entire concept of business, finance, marketing, medical and many others in our everyday life. The concept of Big data stemmed from the advanced computing technologies that enable us to obtain data in different forms, and from diverse sources. Despite the fact that technologies available for us to download/collect information have become highly sophisticated and uncostly, we are still unable to assure the completeness and the correctness of collected information. For example, though images are abundant in different electronic media, their associated annotations can never be guaranteed to be complete and accurate. Thus, missing labels are inevitable in multi-label learning [26] problems. Missing labels cause negative effects on labeling of test samples, because missing labels change the original label structure (between instances) and label relationship (between labels). Label imputation based on noisy label structure or label relationship extracted from the incomplete label matrix often produces unreliable image classification results. Recent multi-label research have found numerous real-world applications such as text categorization in which documents can be assigned to different pre-defined topics [28,31,37]; in protein function classification, genes are

* Corresponding author.
E-mail address: eetchow@cityu.edu.hk (T.W.S. Chow).

http://dx.doi.org/10.1016/j.ins.2017.08.061

Fig. 1. An example of multi-labeled image in the ESP Game dataset, which has labels of *boat, sky, island, beach, water, cloud, blue, mountain, sand and ocean* manually provided by the trained experts. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

often associated with different biological functional classes [1,8,47]; in music information retrieval, a piece of melody can convey different human spiritual messages [29,34,36]; and in image annotation, a visual photo can describe different objects and scenarios simultaneously [3,24,25,42,46,48,50].

In classification problems, we often need to handle huge number of data and large number of labels. However, obtaining sufficient numbers of correct labels is difficult and expensive. For instance, the image from ESP Game dataset as shown in Fig. 1 illustrates that the 10 tags (boat, sky, island, beach, water, cloud, blue, mountain, sand, ocean) are required prior to performing supervised learning. Although most images should have multiple classes, they are often tagged only according to the obvious signatures connected to the labels, while other details are often neglected. For example, the tag 'mountain' in Fig. 1 is easily overlooked because of the obscured view. In addition, the existence of synonyms in the class dictionary and intrinsic ambiguities among different classes [41] also make it difficult to provide a complete set of labels manually. As a result, the output performance using incomplete label matrices is often unsatisfactory. It is necessary to devise a strategy to handle missing labels. Also it must be noted that training data with multiple label information are not always available in abundance because manual labeling is laborious and costly. In order to enlarge the size of training datasets for improving classification accuracy, there is a need to exploit the abundant unlabeled samples. We in this paper will use semi-supervised learning for finding missing labels.

Semi-Supervised Learning (SSL), which leverages the labeled instances in addition to unlabeled instances, has recently received increasing attentions as it provides reliable and improved performance. The most active area in SSL is graph-based learning that models the weighted graph according to the similarities among different data points. Unlike the traditional graph-based methods, the relational graphs in multi-label learning can be constructed in two levels which are input feature level and output category level [7,45] for exploring the inherent correlations among multiple instances and labels respectively. Typical graph-based multi-label learning methods estimate labels by regularizing labels to be smooth on the relational graphs in both feature and category level. These methods can enhance the classification performance compared to other semi-supervised multi-label learning approaches if the semantic structure and the semantic correlation can be accurately determined. In this paper, we use graph-based semi-supervised method to complete label recovery, and graphs are constructed in more accurate forms compared to the above mentioned methods [7,45].

This paper focuses on the cases when given multi-labeled instances consist of missing labels and unlabeled instances. Our objective is to make use of noisy and incomplete data to perform label imputation for enhancing classification performance. Labeled and unlabeled samples can be used for clustering according to the data compactness. The discriminant edge weight matrix S can be used to measure the similarity among nodes in the instance space. However, the undirected-edge relational graph in the input feature space cannot be mapped directly to the output label space, because the label structure Z reflecting the semantic relationship among different instances cannot be assumed to be the same as S due to the semantic gaps (distributions of features and semantic concepts is non-consistency) [33]. The weight matrix S is still required for constructing the label structure Z. This is partly attributed to the large portion of latent labels in the labeled matrix and the unlabeled matrix. Intuitively, these labels are not independent of each other. In Fig. 1, because of the tag 'beach', it gives a high probability that a tag 'ocean' is also a candidate to the image. Many existing multi-label learning methods have proved that significant improvements on annotation can be achieved by involving dependence among classes [24,25,40,50]. The class co-occurrence matrix is often calculated by cosine similarity in a one-to-one way [40,50], or local geometry reconstruction in a one-to-all way [24,25]. But mining the correlation using the above methods from incomplete tags will result in performance degradations, especially under the cases when the number of labels is insufficient. Therefore, instead of taking the correlation information as a priori knowledge, we model the global correlations among labels in an iterative fashion to obtain a reliable relational semantic graph. Noted that the graph can be shared with other methods relying on label correlation. In addition to the above techniques that can explicitly improve the performance of multi-labeling, we also introduce sparse [10,12,13] and nonnegative constraints on describing the mentioned intrinsic geometrical structures in the input and output spaces. It provides visualized and reliable weight matrices that aim at reflecting real situations as a result. It is worth noting that the same constraints of satisfying the assumptions of non-negativity of label values and limited literal expression of one object are applied to missing values and predicted label assignment. In summary, our method increases the robustness of semantic gaps and unreliable class interdependences.