Neurocomputing 171 (2016) 242–251

Contents lists available at ScienceDirect

Neurocomputing

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

Efficient multi-modal hypergraph learning for social image classification with complex label correlations

Leiquan Wang [a,b,*], Zhicheng Zhao [a], Fei Su [a]

[a] School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China
[b] College of Computer and Communication Engineering, China University of Petroleum, Qingdao, China

ARTICLE INFO

Article history:
Received 7 April 2015
Received in revised form
2 June 2015
Accepted 21 June 2015
Communicated by X. Gao
Available online 10 July 2015

Keywords:
Efficient hypergraph learning
Multi-modal feature fusion
Multi-label correlation
Social image classification

ABSTRACT

Multi-label and multi-modality are two dramatic characteristics of social images. Multi-labels illustrate the co-occurrence of objects in an image; while multimodal features represent the image from different viewpoints. They describe social images from two different aspects. However, it is of considerable challenge to integrate multimodal features and multi-labels simultaneously for social images classification. In this paper, we propose a hypergraph learning algorithm to integrate multi-modal features and multi-label correlation seamlessly. More specifically, we first propose a new feature fusion strategy by integrating multi-modal features into a unified hypergraph. An efficient multimodal hypergraph (EMHG) is constructed to solve the high computational complexity problem of the proposed fusion scheme. Secondly, we construct a multi-label correlation hypergraph (LCHG) to model the complex associations among labels. Moreover, an adaptive learning algorithm is adopted to learn the label scores and hyperedge weights simultaneously with the combination of the two hypergraphs. Experiments conducted on real-world social image datasets demonstrate the superiority of our proposed method compared with representative transductive baselines.

1. Introduction

With the development of social images, it is of great challenge to analyse and organize multimodal data effectively. As a fundamental problem in image processing, classification has been widely used in social image annotation, retrieval, reranking etc. Multi-label and multi-modality (see Fig. 1) are two dramatic characteristics of social images, leading to a more complex situation for social image classification. However, both multi-label correlation and multi-modal feature fusion are helpful for social image classification.

Multi-modality is a typical feature of social image. Users share images with corresponding tags (see Fig. 1). Images provide visual features, while text information offers semantic features. Furthermore, these social images contain a wide range of categories from small objects to large scene information. It is true that one visual descriptor can be regarded as a better representation under certain circumstances than the others [1]. For example, local feature, such as SIFT [2], shows superiority in describing rich textural objects, whereas global feature, such as GIST [3], is more effective for scene recognition. Therefore, employing multi-modal features[1] can be a better solution [4]. However, how to fuse multi-modal features is a challenging problem. Recently, graph-based fusion methods [4–6] have gained great interest. They hold that data are in local manifold spaces for different modal features respectively. Graphs are constructed for each modal feature and then weights are assigned for each graph with a linear manner. This approach treats each modal feature separately. It is difficult to construct a graph to cover local manifolds of different features. Meanwhile, graph-based methods consider only the pairwise relationship of the samples. Therefore, an overall context fusion scheme is needed, and the high-order relationship among social images should be also set up, where multimodal features can be integrated into a unified hypergraph due to its flexible construction scheme (see Fig. 2).

Multi-label is another characteristic of social image. A typical social image is usually associated with several semantic labels (see Fig. 1). Lots of work have been proposed to address the multi-label classification problem, and have achieved great achievements, such as SVM [7], CNN [8] etc. However, these methods depend too much on the abundant and accurate labelled data. The high cost of annotation

* Corresponding author.
E-mail addresses: richiewlq@gmail.com (L. Wang), zhaozc@bupt.edu.cn (Z. Zhao), sufei@bupt.edu.cn (F. Su).

[1] Note that multimodal features here are general terms and contain 2 cases: (1) different modalities, such as video, image and text; (2) different features from the same modality, such as SIFT, GIST and LBP.

http://dx.doi.org/10.1016/j.neucom.2015.06.064
0925-2312/

Fig. 1. Example of social images with associated tags and annotated multi-label from MIR Flickr.

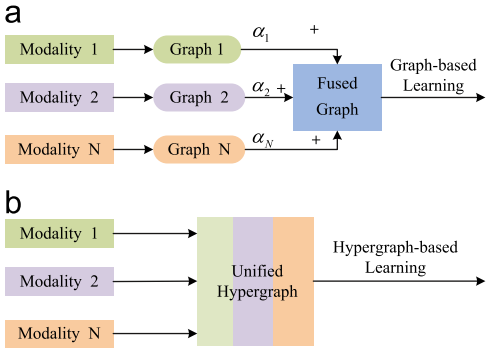

Fig. 2. Comparisons on graph and hypergraph based fusion scheme.

makes us have to face the situation with insufficient and partially labelled data. Therefore, it is desired to effectively utilize the large amount of unlabelled data together with the limited amount of labelled data to improve the multi-label classification performances. Different from single label classification by assuming labels that are mutually exclusive, labels of social images are usually interdependent with each other. For example, "tree" and "bird" are higher probabilistic to appear in the same image than "tree" and "fish". This kind of correlation is helpful for inferring unknown labels from known labels [9]. Pair-wised correlation (e.g. bird and tree) has been used for different algorithms [5,10–12] to improve the performance. However, pair-wised correlation between labels sometimes is insufficient to describe the complex situation in the real-world, and a high-order correlation (e.g. bird, tree and outdoor) among labels would improve the performance of classification.

Hypergraph [13] is a generalization of traditional graph, which models high-order relationships among objects. Hypergraph-based methods are widely used in machine learning and computer vision and have achieved promising performance [14–17]. However, there are still issues (such as computational complexity, the weights of different modalities) that prevent this approach to achieve a better performance for multi-modal case. The computational cost mainly relies on the number of vertices (n): with increasing of n, the computational efficiency will be degraded rapidly. Many works [18,19] are devoted to solve the problem by decreasing the cost $O(n^3)$ to about linear complexity. The number of hyperedges (m) is another factor to influence the cost $O(m^2)$ when assigning the hyperedge weights [20]. For its potential decline on accuracy, no work has been done to reduce the number of hyperedges, while it is an important issue to consider.

In this paper, we propose an efficient multi-modal hypergraph for social image classification by seamlessly integrating multi-modal features and multi-label correlation. The multi-modal features are fused into a unified hypergraph. To reduce the number of hyperedges, we obtain m-representative centroids for each modality by clustering. These centroids have a strong representative power to cover the whole dataset. Images are allocated to its s-nearest centroids and finally m hyperedges are generated. Considering different centroids having different impacts on the hyergraph, we simultaneously learn the hyperedge weights and label scores. The main contributions of our work are summarized as follows:

(1) A new hypergraph based fusion scheme is proposed. Different from the graph-based fusion techniques treating features separately, the new fusion scheme integrates multi-modal features into a unified hypergraph with a comprehensive way.
(2) Efficient multimodal hypergraph (EMHG) is proposed to solve the high computational cost problem of the newly fused hypergraph.
(3) A hypergraph-based learning approach for social image classification is proposed by simultaneously exploiting multi-modal features complementariness and multi-label correlation with a high-order manner.
(4) Comprehensive experiments are conducted to empirically analyze the proposed method. The experimental results on two social image datasets validate the effectiveness of our method.

The remainder of this paper is organized as follows. In Section 2, we discuss the most relevant work, in Section 3 we review the conventional hypergraph learning for classification, and in Section 4 we present our efficient multi-modal hypergraph learning with complex label correlation in details. And Section 5 demonstrates the experimental results. The last section is the conclusion.

2. Related work

2.1. Multi-label classification

Multi-label classification is a fundamental supervised problem in machine learning. However, labeled instances are often difficult to obtain [21]. Early works [22–24] proved that the combined use of unlabeled data and a small amount of labeled data can produce considerable improvement in learning accuracy. In semi-supervised approaches [25,26], the unlabeled data are utilized by assuming the testing data are different from the unlabeled data. In contrast, transductive approaches [27,28], aiming to get the best generalization on the unlabeled data, exploit unlabeled data with the assumption that the testing data are exactly the unlabeled data. Exploiting label dependency [29,30] is an effective way for multi-label learning, and building up pair-wised correlation is a universal way. Ji et al. [12] proposed a general framework for extracting shared structures in multi-label classification with a pair-wised label correlation constraint. Wang et al. [10] proposed an asymmetric label correlation for image annotation. However, the label associations in social images are more complex than other classification problems.

In this paper, we proposed a new transductive learning approach for multi-label classification with the help of a high-order label correlation.