Neurocomputing 138 (2014) 167–179

Contents lists available at ScienceDirect

Neurocomputing

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

Graph-based multimodal semi-supervised image classification

Wenxuan Xie, Zhiwu Lu, Yuxin Peng*, Jianguo Xiao

Institute of Computer Science and Technology, Peking University, Beijing 100871, China

ARTICLE INFO

Article history:
Received 5 July 2013
Received in revised form
16 October 2013
Accepted 3 December 2013
Communicated by Haowei Liu
Available online 5 April 2014

Keywords:
Tag refinement
Graph-based label propagation
Support vector regression
Multiple graphs

ABSTRACT

We investigate an image classification task where training images come along with tags, but only a subset being labeled, and the goal is to predict the class label of test images without tags. This task is important for image search engine on photo sharing websites. In previous studies, it is handled by first training a multiple kernel learning classifier using both image content and tags to score unlabeled training images and then establishing a least-squares regression (LSR) model on visual features to predict the label of test images. Nevertheless, there remain three important issues in the task: (1) image tags on photo sharing websites tend to be imperfect, and thus it is beneficial to refine them for final image classification; (2) since supervised learning with a subset of labeled samples may be unreliable in practice, we adopt a graph-based label propagation approach by extra consideration of unlabeled data, and also an approach to combining multiple graphs is proposed; (3) kernel method is a powerful tool in the literature, but LSR simply treats the visual kernel matrix as an image feature matrix and does not consider the powerful kernel method. By considering these three issues holistically, we propose a graph-based multimodal semi-supervised image classification (GraMSIC) framework to handle the aforementioned task. Extensive experiments conducted on three publicly available datasets show the superior performance of the proposed framework.

1. Introduction

Image classification has been studied for decades [1–6]. The goal of image classification is to determine whether an image belongs to a predefined category or not. In the literature, different types of categories have been investigated, e.g., scenes [7] or objects [8]. To handle an image classification problem, a supervised framework can be used, where a binary classifier is first learned from manually labeled training images and then used to predict the class label of test images. By increasing the quantity and diversity of manually labeled images, the learned classifier can be enhanced. However, it is a time-consuming task to label images manually. Although it is possible to label large numbers of images for many categories for research purposes [9], it is usually unrealistic, e.g., in photo sharing applications. In practice, we usually have to handle a challenging classification problem by using only a small number of labeled samples. In the literature, semi-supervised learning [10] has been proposed to exploit the large number of unlabeled samples and thus helps to handle the scarcity of labeled samples to some extent.

In this paper, we investigate a multimodal semi-supervised image classification problem originally raised in [11]. In this problem, training images have associated tags (e.g., from Flickr), and only a limited number of the training samples come along with class labels. The goal of this problem is to predict the class label of test images without tags. This is an important problem for image search engine on photo sharing websites. Since a newly uploaded image and also a considerable part of the existing images on websites have no associated tags, it is necessary to build up an image-only classifier for such image search engines with available resources (i.e., tagged images, and only a subset is labeled). To solve this problem, a two-step method has been proposed in [11]. In the first step, a multiple kernel learning (MKL) [12,13] classifier is learned by utilizing labeled training images with tags, which is then used to score unlabeled training images. In the second step, a least-squares regression (LSR) model is learned on the training set by using centered visual kernel columns as independent variables and using centered classification scores as dependent variables, which is then used to predict the scores of test images.

Nevertheless, we still need to consider the following *three* important issues, since they all may lead to performance degeneration in the aforementioned problem:

Tag imperfectness: Image tags on photo sharing websites (e.g., Flickr) are often inaccurate and incomplete, i.e., they may not directly relate to the image content and typically some relevant tags are missing. Some example images are shown in Fig. 1. For example, as we can see from the image on the upper left corner, the tag ‘car’ is inaccurate and the tag ‘bird’ is missing. Since the

* Corresponding author. Tel.: +86 10 82529699; fax: +86 10 82529207.
E-mail address: pengyuxin@pku.edu.cn (Y. Peng).

http://dx.doi.org/10.1016/j.neucom.2013.12.052
0925-2312/

Tags: aviary, **car**
Labels: bird

Tags: 2006, dogs, **sheep**
Labels: dog

Tags: **tree**, reflection, bokeh, home
Labels: sunset, water

Tags: **food**
Labels: indoor, people

Fig. 1. Example images from PASCAL VOC'07 (top row) and MIR Flickr (bottom row) datasets with their associated tags and class labels. Tags in **bold** are inaccurate ones.

original tags are imperfect, it is a suboptimal choice to use them directly. Hence, we propose to refine these tags by using the affinity of image content as the first step.

Label scarcity: Since only a subset of the training images is labeled, supervised models such as an MKL classifier learned by using only labeled samples may be unreliable in practice. To handle the scarcity of labeled samples, we adopt a graph-based label propagation method to leverage the large number of unlabeled samples. By exploiting the graph structure of labeled and unlabeled samples, the label propagation method is shown to perform better in the experiments. More notably, since an average combination of multiple graphs for label propagation is only a suboptimal choice, we propose an approach to learning the combination weights of multiple graphs.

Ignorance of kernel method: The LSR model used in [11] simply treats the visual kernel matrix as an image feature matrix and does not consider the powerful kernel method. Moreover, the singular value decomposition (SVD) step involved in the LSR model is time-consuming. Instead of LSR, we propose to use support vector regression (SVR) to predict the class label of test images, since SVR can readily leverage the original visual kernel and make full use of image features in the reproducing kernel Hilbert space (RKHS) [14].

In summary, taking into account the *three* important issues, we propose a graph-based multimodal semi-supervised image classification (GraMSIC) framework to handle the aforementioned task by combining the following three components: (1) tag refinement; (2) graph-based label propagation by combining multiple graphs; (3) SVR. Fig. 2 shows the schematic overview of the proposed framework.

Upon our short conference version [15], this paper provides two additional contributions: (1) an approach to learning the combination weights of multiple graphs is proposed; (2) more extensive experimental results are added on three publicly available datasets, i.e., PASCAL VOC'07 [8], MIR Flickr [16] and NUS-WIDE-Object [17]. In the next two subsections, we briefly present preliminary notations and paper organization.

1.1. Preliminary notations

We denote training image set and test image set by $I_{tr} = \{x_1, x_2, \ldots, x_{n_1}\}$ and $I_{te} = \{x_{n_1+1}, x_{n_1+2}, \ldots, x_{n_1+n_2}\}$, respectively. Note that $n = n_1 + n_2$ is the total number of samples. Training images come along with tags, where the tag set is represented by $V = \{v_1, v_2, \ldots, v_m\}$ and m stands for the size of the tag set. The initial tag membership for all training images can be denoted by a binary matrix $T_{tr} \in \{0,1\}^{n_1 \times m}$ whose element $T_{tr}(i,j)$ indicates the presence of tag v_j in image x_i, i.e., $T_{tr}(i,j) = 1$ if tag v_j is associated with image x_i, and $T_{tr}(i,j) = 0$ otherwise. Moreover, only a small number of the training images are assigned with class labels from c categories, and the initial label matrix is denoted by $Y_{tr} \in \{1, 0, -1\}^{n_1 \times c}$, whose element $Y_{tr}(i,j)$ indicates the label of image x_i, i.e., $Y_{tr}(i,j) = 1$ if x_i is labeled as a positive sample of category j, $Y_{tr}(i,j) = -1$ if x_i is labeled negative, and $Y_{tr}(i,j) = 0$ if x_i is unlabeled. The goal is to predict the class label of test images without tags, i.e., an $n_2 \times c$ matrix Y_{te}.

Moreover, in order to state conveniently, the values determined by the learning algorithm are called 'parameters', and the values which require hand-tuning in advance are called 'hyperparameters' [18].

1.2. Paper organization

The paper is organized as follows. We begin by introducing related studies in the literature in Section 2. Then, we present the GraMSIC framework in Section 3. In Section 4, we discuss in detail the proposed approach to combining multiple graphs for label propagation. Moreover, we investigate the complexity issues and