Neurocomputing 149 (2015) 1573–1586

Contents lists available at ScienceDirect

Neurocomputing

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

Online semi-supervised annotation via proxy-based local consistency propagation

Lei Huang [a,*], Xianglong Liu [a], Binqiang Ma [b], Bo Lang [a]

[a] State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, China
[b] College of Sciences, Henan Agricultural University, Zhengzhou 450002, China

ARTICLE INFO

Article history:
Received 12 January 2014
Received in revised form
12 August 2014
Accepted 16 August 2014
Communicated by X. Gao
Available online 24 August 2014

Keywords:
Image annotation
Label propagation
Semi-supervised learning
Incremental learning
Online learning

ABSTRACT

In this paper, we propose a novel label propagation algorithm named Proxy-based Local Consistency Propagation (PLCP), in which the label information is first propagated from labeled examples to the unlabeled ones, and then spreads only among unlabeled ones mutually until a steady state is reached. To meet the requirements of efficiency in many real-world image annotation applications, we propose an online semi-supervised annotation framework where the new examples can be predicted and used to update the model. Specifically, we extend PLCP to work under an inductive setting and propose an incremental model updating method that can incorporate the new examples including labeled and unlabeled examples. The comprehensive experiments on MNIST, CIFAR-10 and PIE datasets show that our proposed PLCP achieves superior performance compared with the baselines, and our proposed incremental model updating method can achieve significant promotion in efficiency, with the nearly identical accuracy compared to re-training.

1. Introduction

The amount of digital images have grown rapidly in recent years. The demand for effective solutions to manage images is increasing tremendously. Automatic image annotation is crucial to understanding image semantic concepts for storage, indexing and retrieval purposes. For most approaches of automatic image annotation, statistical models are usually built from manually labeled examples, and then the labels are assigned to unlabeled examples utilizing these models. However, this process faces a major problem that labeled data is often insufficient so that its distribution may not be able to well approximate that of the entire data set, which usually leads to inaccurate annotation results.

Semi-Supervised Learning (SSL) methods, by leveraging unlabeled data with certain assumptions, are promising to build more accurate models than those purely supervised methods. Typical SSL methods include self-training [1], co-training [2], transductive SVM [3], and graph-based methods [4–6]. A detailed survey of the literature on SSL methods is presented in [7]. As an important family of SSL, graph-based methods have gained much attention in the past few years. The common denominator of graph-based SSL methods is to model the whole data set as a graph $G=\langle V,E\rangle$, where V is the set of vertices and each vertex corresponding to a data sample; E is the set of edges associated with a weight matrix which reflects the similarity between samples. Essentially, the graph-based SSL methods, especially semi-supervised classification, can be viewed as estimating a function on G subject to the following properties: (1) it should be close to the given labels on the labeled vertices and (2) it should be smooth on the whole graph. The former property can be formulated as a loss function penalizing the deviation of the predicted labeling from the given labeling. The latter property can be expressed as a regularizer enforcing label consistency.

Generally, graph-based SSL can be decomposed into two main components: graph construction and label propagation. In recent years, most of the researches focus on graph construction which plays an important role in the perception of the neighbor structure in many applications [8–19]. Wang and Zhang [8,15] constructed the graph by performing an operation very similar to locally linear embedding (LLE) [20] on the data points except constraining the LLE weights to be non-negative. He et al. proposed the Sparse Probability Graph (SPG) which is modeled to a lasso problem with the nonnegativity constraint, without considering corruption errors [16]. Zhuang et al. proposed the NNLRS-graph which assumes that the weights of edges in the graph are obtained by seeking a nonnegative low-rank and sparse matrix [17]. There also are some works relating the label propagation [21,15,22]. For example, Frey et al. proposed a method called Affinity Propagation (AP) to solve the cluster problem [21] and Ni et al. proposed a

* Corresponding author.
E-mail addresses: huanglei@nlsde.buaa.edu.cn, huanglei36060520@gmail.com (L. Huang).

http://dx.doi.org/10.1016/j.neucom.2014.08.035
0925-2312/

simultaneous approach which integrates both label propagation and optimal feature representation [22].

Our work mainly focuses on the strategy of label propagation and incremental learning. Different from the conventional methods in which label propagation is performed over the whole data set, we truncate the propagation mutually among labeled examples based on the assumption that the labeled examples should keep their labels unchanged [4,5]. We propose a novel label propagation algorithm named Proxy-based Local Consistency Propagation (PLCP), in which we only consider two types of relationships among examples: labeled-to-unlabeled and unlabeled-to-unlabeled. Fig. 1 shows our propagation framework, where each labeled example propagates the label information to its unlabeled neighbors which are called proxies, and then the proxies spread the information among the unlabeled examples mutually until a steady state is reached. Since each labeled example initially propagates its labels to more than one proxy, our approach can be intuitively understood as increasing the labeled examples when propagation is performed over the unlabeled examples.

Many graph-based SSL methods have been applied to image or video annotation [23–32]. However, most of them face the limitation that learning must be performed in a batch mode, which means that they require the training dataset to be available all at once [33] and need a full re-training procedure when new labeled examples are added into the training dataset. There also exist some researches combining online learning and semi-supervised learning [24,34–40]. Andrew et al. proposed an online learning algorithm for manifold regularization of SVMs solved by a convex programming with stochastic gradient descent in kernel space [35]. However, it only achieves asymptotic zero-regret guarantee. Valko et al. employed data quantization and maintained a compact representation of the complete data adjacency graph [37]. Wang et al. proposed an interactive annotation framework for RNAi microscopic cellular images, in which the data adjacency graph is constructed aforehand and only the labels are utilized to update the model [24]. However, it cannot incrementally handle the new examples. As far as we know, there is no work to solve the propagation matrix incrementally based on the closed-form solutions of graph-based SSL methods, which is the main difficulty for graph-based SSL methods in online annotation.

In this paper, we attempt to address the problem by proposing an online semi-supervised annotation framework where the new example can be predicted and used to update the model. Since our proposed PLCP suppresses the mutual effect among labeled examples, it is feasible to incrementally update the annotation model via handling the labeled examples and unlabeled examples

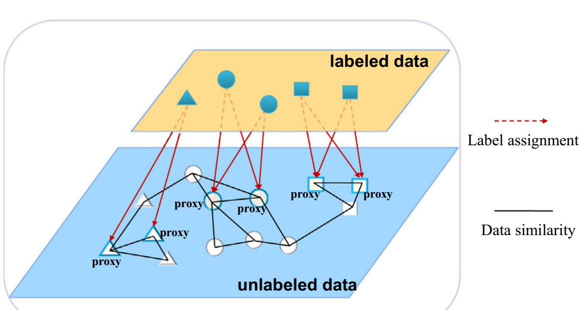

Fig. 1. The mixed graph constructed from a dataset with both labeled and unlabeled examples. The triangles, circles and squares represent three classes. The label information is first propagated from labeled examples to the proxies (regarded as a directed graph), and then spreads only among unlabeled ones (regarded as an undirected graph).

differently. Particularly, for incorporation of the new labeled examples, we can achieve fast incremental learning without accuracy loss by means of the decomposed formulation; for incorporation of the new unlabeled examples, we propose a method to calculate the propagation matrix incrementally which avoids the expensive calculation of matrix inversion. Besides, we propose an online semi-supervised annotation framework, in which we can predict the new examples and use them to incrementally update the model. We also extend PLCP to work under an inductive setting [41] to sustain real-time prediction for new examples.

To summarize, we highlight here the main contributions of our work:

- We propose a novel label propagation algorithm named PLCP, in which the label information is first propagated from labeled examples to its unlabeled neighbors, and then spreads only among unlabeled ones like a spreading activation network [42].
- We introduce an online semi-supervised annotation framework, in which the new example can be predicted and used to update the model. In particular, we extend PLCP to support an inductive mode for efficient prediction of the new examples.
- We design an algorithm to update the model incrementally when incorporating the new examples (including labeled and unlabeled) into the training set, which dramatically relieves the computational burden for model re-training. The comprehensive experiments show that our proposed incremental model updating method can achieve significant promotion in efficiency.

This paper extends upon the previous work [43] with additional exploration on the algorithm generalization (extending PCLP to work under an inductive setting), more robust online annotation framework (supporting incremental learning when incorporating unlabeled examples) and amplified experimental results. The remaining sections are organized as follows. Section 2 elaborates on the PLCP algorithm design and its interpretation. In Section 3, we propose a semi-supervised online annotation framework, extend PLCP to inductive learning and design the incremental learning algorithm for PLCP. Comprehensive experimental results are presented to demonstrate the effectiveness and efficiency of our approach in Section 4. Finally Section 5 concludes this paper.

2. Proxy based local consistency propagation

In this section, we will formally present our Proxy based Local Consistency Propagation (PLCP) method. First we describe the notations used in this paper. Given data points set $\chi=(\chi_L,\chi_U)=\{\mathbf{x}_1,\dots,\mathbf{x}_l,\mathbf{x}_{l+1},\dots,\mathbf{x}_n\}$ where $\mathbf{x}_i\in R^d$. The first l points $\chi_L=\{\mathbf{x}_1,\dots,\mathbf{x}_l\}$ are labeled with $y_i\in Ł=\{1,\dots,c\}$ and the remaining points $\chi_U=\{\mathbf{x}_{l+1},\dots,\mathbf{x}_n\}$ are unlabeled. The goal is to predict the label $y_j(l+1\le j\le n)$ of the unlabeled points.

Let $\mathcal{F}$ denote the set of $n\times c$ matrices. A matrix $\mathbf{F}=[\mathbf{F}_1^T,\dots,\mathbf{F}_n^T]^T\in\mathcal{F}$ is a vectorial function $\mathbf{F}:\chi\rightarrow R^c$ which assigns a vector $\mathbf{F}_i$ to each point $\mathbf{x}_i$. The label matrix $\mathbf{Y}=[\mathbf{Y}_1^T,\dots,\mathbf{Y}_n^T]^T$ is described as $\mathbf{Y}\in R^{n\times c}$ with $\mathbf{Y}_{ij}=1$ if $\mathbf{x}_i$ is with label $y_i=j$ and $\mathbf{Y}_{ij}=0$ otherwise.

2.1. Algorithm

A typical assumption used in graph-based SSL is that nearby points are likely to have the same label. As the basis of label propagation, pairwise similarity measure is necessary for graph-based SSL methods. We use the pairwise similarity between