Information Sciences 454–455 (2018) 216–228

Contents lists available at ScienceDirect

Information Sciences

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

On using supervised clustering analysis to improve classification performance

Haitao Gan[a,*], Rui Huang[b], Zhizeng Luo[a,*], Xugang Xi[a], Yunyuan Gao[a]

[a] School of Automation, Hangzhou Dianzi University, Hangzhou, 310018, China
[b] School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, Guangdong, 518172, China

ARTICLE INFO

Article history:
Received 8 May 2017
Revised 7 February 2018
Accepted 30 April 2018
Available online 1 May 2018

Keywords:
Supervised classification
Supervised clustering
Regularized Least Squares Classification
k-means

ABSTRACT

During the past decade, graph-based learning methods have proved to be an effective tool to make full use of both labeled and unlabeled data samples to improve learning performance. These methods try to discover the intrinsic structures and discriminative information embedded in the data, by building one or more graphs to model the relationship among the data samples. Consequently, how to build an effective graph is the core problem. In this paper we introduce a novel graph-based classification method, called Supervised clustering-based Regularized Least Squares Classification (SuperRLSC), in which local and global graphs of the data are built by supervised clustering. The motivation is that supervised clustering may discover more actual data structures compared to unsupervised clustering. In our algorithm, we firstly employ supervised k-means to partition the whole training dataset into several meaningful clusters in order to discover the intrinsic and discriminative structures. We then use the discovered structures to build local and global graphs of the data. The local graph reveals local geometric and discriminative structures, while the global graph reveals global discriminative information. Finally a hybrid local/global graph-based regularization term is embedded into supervised classification (i.e., RLSC in this paper). To validate the effectiveness of our algorithm, a series of experiments are performed on several UCI benchmark datasets. The results show that our algorithm can achieve better or at least comparable performance to the other graph-based algorithms and the traditional state-of-the-art supervised classification methods.

© 2018 Elsevier Inc. All rights reserved.

1. Introduction

The past ten years have witnessed a new vitality of machine learning in many practical applications [19,20,31,33]. Among different learning methods, graph-based learning methods have become one of the most widely used approaches and play an important role in the machine learning field. Graph-based learning methods have proved to be an effective tool to make full use of both labeled and unlabeled data samples to improve learning performance. Hence they have recently attracted much attention and can be applied in unsupervised, semi-supervised and supervised scenarios [3,8,16,26,34]. Generally speaking, the existing graph-based learning methods can be cast into two categories which are called *unlabeled sample-based* methods and *labeled sample-based* methods. As the name implies, the *unlabeled sample-based* methods utilize unla-

[*] Corresponding authors.
E-mail addresses: htgan@hdu.edu.cn (H. Gan), ruihuang@cuhk.edu.cn (R. Huang), luo@hdu.edu.cn (Z. Luo), xixugang@hdu.edu.cn (X. Xi), gyy@hdu.edu.cn (Y. Gao).

https://doi.org/10.1016/j.ins.2018.04.080
0020-0255/© 2018 Elsevier Inc. All rights reserved.

beled samples (ignoring the labels for the labeled ones) to construct a p nearest neighbors (p-NN) graph and discover the local geometric structure of the data. In this sense, the *unlabeled sample-based* approaches do not consider the label information. On the contrary, the *labeled sample-based* approaches generally construct intra-class and inter-class graphs to find the discriminative information using the labeled samples.

In the *unlabeled sample-based* approaches, He et al. [17] introduced a graph-based method for clustering and proposed a Laplacian regularized Gaussian Mixture Model (LapGMM) to improve clustering performance. LapGMM constructed a p-NN graph to exploit the local geometric structure of the whole samples. Empirical results on image clustering showed the effectiveness of LapGMM. The graph-based methods can also be applied in the other unsupervised scenarios, such as dimensionality reduction [5], nonnegative matrix factorization [7]. The graph-based methods have been well applied in semi-supervised learning, including clustering [13], classification [3,26], and dimensionality reduction [6]. Gan et al. [13] proposed Semi-Supervised Locally Consistent GMM (Semi-LCGMM) which constructed a local graph to model the manifold structure of the labeled and unlabeled samples. Experimental results illustrated that Semi-LCGMM outperformed traditional unsupervised and semi-supervised clustering on image clustering and segmentation. Zhu et al. [35] also constructed a local graph and tried to learn a semi-supervised classifier formulated in terms of a Gaussian random field on this graph, where the mean of the field was characterized in terms of harmonic functions and computed using matrix methods or belief propagation. Belkin et al. [3] incorporated a graph-based regularization term into the model-based supervised classification methods. Laplacian Regularized Least Squares Classification (LapRLSC) and Laplacian Support Vector Machine (LapSVM) were proposed in the literature [3]. Likewise, Cai et al. [6] proposed Semi-supervised Discriminant Analysis (SDA) which embedded a graph-based regularization term into LDA for dimensionality reduction. SDA alleviated the performance degeneration of LDA with small training size. Overall, the *unlabeled sample-based* approaches can achieve promising performance in the graph-based unsupervised and semi-supervised learning.

In the *labeled sample-based* approaches, Peng et al. [22] proposed a discriminative Graph regularized Extreme Learning Machine (GELM) in which the label information of training samples was used to construct an adjacent graph. The edge weight of the graph was set to $\frac{1}{N_i}$ if two samples belonged to the same class and 0 otherwise where N_i was the sample number in the ith class. Therefore, the samples from the same class should have the similar outputs. The experimental results on face recognition verified the effectiveness of GELM. Furthermore, Xue et al. [29] proposed Discriminatively Regularized Least-squares Classification (DRLSC) in which a discriminative graph-based regularization term was embedded into RLSC. The discriminative information was exploited by two graphs (i.e., intra-class and inter-class). The intra-class and inter-class graphs were constructed through local p nearest samples belonging to the same and different classes, respectively. The two graphs discovered not only the local discriminative information, but the local manifold structure to some extent.

However, on the one hand, the *unlabeled sample-based* approaches do not consider the discriminative information of labeled samples which is crucial for classification. On the other hand, the *labeled sample-based* approaches do not make full use of the local or global geometric structures of the whole samples. Specifically, DRLSC ignored the global discriminative structure covered by the labeled samples. In order to discover the intrinsic and discriminative information, there are several feasible ways, such as unsupervised and semi-supervised clustering. Wang et al. [26] proposed Discrimination-Aware LapRLSC (DA_LapRLSC) which exploited the intrinsic structure of the labeled and unlabeled samples using unsupervised clustering. The discovered structure was used to build a discrimination-aware graph and a graph-based regularization term was then embedded into LapRLSC to improve the classification performance.

Compared to unsupervised clustering, semi-supervised clustering can make better use of the labeled information [2,21,27,30,32]. The methods in semi-supervised clustering can generally be divided into the following categories: (1) metric-based approaches; (2) constraint-based approaches. The metric-based approaches mainly focus on learning a distance metric which should satisfy the given prior information. Yin et al. [32] proposed an adaptive Semi-supervised Clustering Kernel Method based on Metric learning (SCKMM) which utilized the pair-wise constraints. Yan et al. [30] developed a novel search-based semi-supervised clustering method which learned the multi-viewpoint based similarity metric. The constraint-based approaches usually revise the objective function or initialize the cluster centers to guide the clustering process through the label information. Basu et al. [2] invented a semi-supervised framework of k-means which used the labeled samples to compute the initial cluster centers. Pedrycz and Waletzky [21] presented Semi-Supervised Fuzzy C-Means (SSFCM) by introducing a fidelity term derived from the label information. SSFCM implemented a tradeoff between the unsupervised outputs and the given labels.

Since semi-supervised clustering can generally yield better clustering performance than unsupervised clustering, Gan et al. [12] employed semi-supervised fuzzy clustering to discover the intrinsic structure in data. The data structure was then used to model the local discriminative capacity of the unlabeled samples. Experimental results on face recognition illustrated that the unlabeled samples might cover the useful discriminative information to improve the classification performance.

Although unsupervised [26] and semi-supervised [12] clustering have shown the effectiveness in revealing the geometric and discriminative information, they have blindness in the clustering process to some extent and do not exploit the actual data structure. Additionally, the obtained clusters do not contain class-label information. Moreover, if the whole samples are all labeled, unsupervised and semi-supervised will not be the optimal strategy in discovering the useful information. In this case, supervised clustering [10,11,24] will be an effective tool to discover the meaningful and actual data structure by making full use of the labeled samples.

Download English Version:

https://daneshyari.com/en/article/6856382

Download Persian Version:

https://daneshyari.com/article/6856382

Daneshyari.com