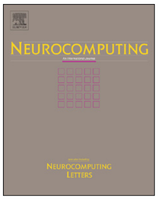
Neurocomputing 171 (2016) 452–462

Contents lists available at ScienceDirect

Neurocomputing

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



# Active learning on anchorgraph with an improved transductive experimental design

Weijie Fu, Shijie Hao*, Meng Wang

*School of Computer and Information, Hefei University of Technology, Hefei 230009, China*

ARTICLE INFO

*Article history:*
Received 19 February 2015
Received in revised form
25 May 2015
Accepted 25 June 2015
Communicated by Jun Yu
Available online 2 July 2015

*Keywords:*
Semi-supervised learning
Anchorgraph
Active learning
Transductive experimental design
Graph spectrum
Data distribution

ABSTRACT

Anchorgraph based learning methods have met with success in modeling the large data for scalable semi-supervised learning. However, like most graph based learning algorithms, they are usually built with a randomly selected labeled set classified in advance. Although many pool-based active learning methods have been proposed, they often require a relatively large computational and storage consumption, which tends to impose extra burden on the learning system. Thus in this paper, we propose a novel active learning method named anchor-based transductive experimental design (ATED). By fully utilizing the representing power of anchors, the improved method efficiently enhances the performance of the original anchorgraph based learning while introduces much less extra cost on computation and storage. Extensive experimental results on real-world datasets have validated our approach in terms of classifying accuracy and computational efficiency.

## 1. Introduction

In many real-world supervised learning tasks, labeling instances to create a training set is time-consuming and labor-intensive. There have been considerable interests in exploring unlabeled data to address this issue in recent years. There are two mainstream groups of such learning methods, i.e. the semi-supervised learning and the active learning. Both of them are extensively applied in fields such as multimedia content analysis [1–3], computer vision [4,5], medical image analysis [6,7].

Given that the labeled data is limited and usually expensive while the unlabeled data is abundant and easy to collect, semi-supervised learning (SSL) methods tend to achieve better performances than supervised ones by fully exploring additional information extracted from larger feature space of unlabeled data [8]. For instance, the Mixture Models with EM [9,10] try to find the maximum likelihood estimation on both the labeled and unlabeled data. The Semi-Supervised Support Vector Machine [11] looks for a low-density gap in unlabeled data to improve the decision boundary.

Graph-based learning [12] is another typical family of SSL, which is based on the assumption: points that can be connected via paths through high-density regions are likely to have the same label. And this kind of learning algorithm typically consists of two main parts: the fitting constraint and the smoothness constraint. These two parts both have clear geometric meanings. The former one means that a good classifying function should not change too much from the initial label assignment, while the latter means that this function should have similar semantic labels among nearby points. Based on the above formulation, these algorithms generally produce satisfying classifying results in the manifold space. However, most traditional graph-based learning methods have relatively expensive computational and storage costs in either constructing the graph or calculating the inverse of graph Laplacian matrix. Recently, Liu et al. [13] proposed a graph-based learning algorithm called Anchorgraph Regularization (AGR), which first finds a small number of anchors to cover the data space and then employs these virtual anchors as transition points to construct a graph in two steps. It reduces the computational cost by subtle low-rank matrix factorization via utilizing the intrinsic structure of this graph. However, like most graph-based learning algorithms, it is generally applied with a randomly selected labeled set classified in advance.

As a different but complementary way to reduce the labeling cost in supervised learning, active learning (AL) [14] has also received much attention. In recently years, various active learning methods have been developed to adapt different kinds of data, mainly including two ideas, i.e. uncertainty sampling and representativeness sampling. Specifically, the methods in the former group are usually associated with classifiers, which are used to

* Corresponding author. Tel.: +86 551 62901392.
*E-mail address:* hfut.hsj@gmail.com (S. Hao).

http://dx.doi.org/10.1016/j.neucom.2015.06.046
0925-2312/© 2015 Elsevier B.V. All rights reserved.

evaluate the uncertainty. For example, SVM based active learning [15] selects unlabeled datapoints, which are closest to the classifying boundary, and obtains labels from users so as to achieve maximal optimization on the hyperplane between two different classes. Graph-based active learning [16] selects the instances with the maximum expected information gain which is evaluated by a graph-based SSL classifier. Differently, methods in the latter group try to select the most representative examples according to data distribution, such as optimal experiment design (OED) [17].

As a typical way of representativeness sampling, OED selects the samples that minimize the variance of a parameterized model and usually works in an iterative process. It analyzes all collected data in order to select new samples in interesting areas of the design space according to some measures. Thus, this kind of experiment design leads to a more efficient distribution of selected samples compared to traditional design of experiments. Recently, Yu et al. [18] proposed the Transductive Experimental Design (TED) which tends to favor instances that are on the one side hard-to-predict and on the other side representative for the rest of the instances. However, due to the relatively large storage and computational cost, i.e. $O(n^2)$, TED has its own limitations to handle large scale databases, especially for the situation where Internet data is rapidly increasing recently.

The aforementioned SSL and AL methods can be naturally incorporated for real-world applications, which is the main focus of our research. In this paper, by fully utilizing the property of anchors in anchorgraph, we propose an improved active learning model called anchor-based transductive experimental design (ATED) and integrate it into the anchorgraph based SSL framework. With the proposed ATED, we can enhance the performance of original anchorgraph based model by introducing additional computational and storage cost as small as possible. Our research highlights itself in the following aspects: First, from the perspectives of the data distribution and the graph spectrum theory, we demonstrate that less-numbered anchors extracted from the original graph still preserve the backbone of the implicit data manifold. Second, based on the former observation, we propose an anchor-based transductive experimental deign (ATED) to efficiently select unlabeled data for manually labeling. Third, we employ the ATED method in a few multi-class classifying tasks and achieve comparable or better accuracy with much less temporal costs than the original TED model.

The rest of this paper is organized as follows. In Section 2, we briefly introduce several related works. In Section 3, we analyze the anchors' properties and review the anchorgraph-based learning model. The proposed ATED and the whole classifying framework are described in Section 4. In Section 5, we conduct experiments on several publicly available datasets to validate our model. Section 6 finally concludes the paper.

## 2. Related works

As the graph-based SSL and the ODE are closely involved in our research, we briefly introduce the related works of these learning methods in this section.

Researchers have been focusing on the two main problems in graph-based SSL, i.e. graph construction and label propagation. On one hand, although the graph can be usually constructed by the traditional kNN strategy, researchers have developed many other ways of building graphs instead of the ordinary kNN, to better capture the data distribution or leverage multimodal information. For instances, Zelnik et al. [19] suggested to employ a local scale in computing the affinity between each pair of points for the edge. Wang et al. [20] proposed a method of measuring neighborhood similarity by exploring both local points and label distributions. Furthermore, since relationships between the objects are more complex than pairwise, Zhou et al. [21] proposed an approach to design hypergraphs which can handle higher-order relationships between elements. Later, Yu et al. [22] designed an adaptive hypergraph for image classification. Liu et al. [23] presented a method that enriches event-based media by constructing an adaptive probabilistic hypergraph. Yu et al. [24] proposed an approach which adopts the high-order distance obtained from the hypergraph in estimating the probability matrix of data distribution. On the other hand, researchers have also designed a few effective label propagating strategies to improve the classification. For instances, Zhu et al. [12] formulated a semi-supervised learning problem with a Gaussian random field model via graph and gave a closed-form solution based on it. Zhou et al. [25] proposed an approach to design a classifying function which is sufficiently smooth with respect to the intrinsic graph structure collectively revealed by labeled and unlabeled datapoints. These graph based SSL methods have been successfully applied in the visual content analysis such as image classification [26–28], image retrieval [29,30] and video content analysis [31,32].

Apart from the above two key issues, with the rapid increasing data size, the computational efficiency has also become another important issue which cannot be ignored. In recent years, much attention has been paid to address the computational issue in the graph based SSL methods. Wang et al. [33] proposed an approach to constructing an approximated kNN graph by two steps, including dividing dataset into subset to build exact neighborhood subgraphs and combining all subgraphs to get an approximated neighborhood graph. Similarly, Wang et al. [34] proposed an iterated approach to construct an approximate kNN graph with two major steps: random hierarchical partitioning and neighborhood propagation. Recently, Liu et al. [13] proposed a novel graph construction approach by employing a small number of virtual anchors (clustering centers) as transition points. It constructs a weight matrix $\mathbf{W}$ of low-rank and then uses this property to achieve impressive computational and storage reduction in the graph-based learning. In conclusion, all these algorithms simplify either the graph construction or the label propagation, so they effectively reduce the computational cost to some extent.

Meanwhile, the ODE methods have also received much attention. Yu et al. [18] proposed an approach named Transductive Experimental Design (TED) which tends to find representative data samples $\mathcal{X}$ that span a linear space to retain the most of information of the whole dataset $\mathcal{V}$. Furthermore, since the pervious transductive experimental design is NP-hard, Yu et al. [35] later formulated it into the continuous optimization and proposed a non-greedy active learning called Convex Transductive Experimental Design which guarantees the global optimum. Cai et al. [36] proposed an approach called manifold adaptive experimental design (MAED) based on the data manifold adaptive kernel. Via employing the Laplacian matrix to reflect the underlying geometry of the data, the MAED method can select the most representative and discriminative datapoints for labeling. ODE based methods have been adopted in many real-world applications. For instance, He et al. [37] proposed a Laplacian Optimal Design for relevance feedback image retrieval which makes efficient use of both user-labeled and unlabeled samples. He et al. [38] proposed a document summarization method based on data reconstruction which finds the summary sentences by minimizing the reconstruction error. Hao et al. [39] employed the TED method in the task of intervertebral MRI image classification. Flaherty et al. [40] applied the robust design in biological experiments with efficient computation.

Different from the above-mentioned methods which aim at selecting representative samples, methods based on uncertainty sampling usually optimize the classifying boundary by finding and