Neurocomputing 123 (2014) 398–405

Contents lists available at ScienceDirect

Neurocomputing

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

Active learning on manifolds

Cheng Li [a], Haifeng Liu [b], Deng Cai [a,*]

[a] The State Key Lab. of CAD&CG, Zhejiang University, No. 388 Yu Hang Tang Road, Hangzhou 310058, China
[b] College of Computer Science, Zhejiang University, No. 388 Yu Hang Tang Road, Hangzhou 310058, China

ARTICLE INFO

Article history:
Received 26 February 2013
Received in revised form
15 May 2013
Accepted 3 August 2013
Communicated by Qingshan Liu
Available online 17 August 2013

Keywords:
Manifold
Active learning

ABSTRACT

Due to the rapid growth of the size of the digital information available, it is often impossible to label all the samples. Thus, it is crucial to select the most informative samples to label so that the learning performance can be most improved with limited labels. Many active learning algorithms have been proposed for this purpose. Most of these approaches effectively discover the Euclidean structure of the data space, whereas the geometrical (manifold) structure is not well respected. In this paper, we propose a novel active learning algorithm which explicitly considers the case that the data are sampled from a low dimensional sub-manifold embedded in the high dimensional ambient space. The geodesic distance of two data points on the manifold is estimated by the shortest-path distance between the two corresponding vertices in the nearest neighbor graph. By selecting the most representative points with respect to the manifold structure, our approach can effectively decrease the number of training examples the learner needs in order to achieve good performance. Experimental results on visual objects recognition and text categorization have demonstrated the effectiveness of our proposed approach.

1. Introduction

In many image processing and computer vision tasks, there are no short of unlabeled data but labels are expensive. In order to reduce the efforts in collecting labels, many researchers studied to use *active learning* [6] for various tasks. The key problem in active learning is determining which unlabeled examples would be the most informative, *i.e.*, improve the classifier the most if they were labeled and used as training examples.

There have been many heuristics proposed for active learning, including choosing the most uncertain data given previously trained models (SVM_{Active} [31]); choosing the data which are expected to change the trained model [5]; exploiting the cluster structure of the data [8]. In statistics, a closely related concept is Optimal Experimental Design (OED) [1] which aims at finding a set of points such that the variance of the estimation is minimized. Classical experimental design approaches include A-optimal design, D-optimal design, E-optimal design and I-optimal design.

Recently, Yu et al. have proposed Transductive Experimental Design (TED) [33] which has yielded impressive results. TED is fundamentally based on the experimental design but evaluates the expected variance on both labeled and unlabeled examples. It has been shown that finding those points which minimize the average predictive variance of the estimated function is equivalent to find those points such that other points can be best approximated by linear combination of the selected points [33]. In other words, TED tries to find the most representative data points (all the other points can be linearly reconstructed from the selected points).

Standard learning systems operate on input data after they have been transformed into feature vectors living in a high dimensional space. In such a space, standard learning tasks like classification, clustering, data selection (active learning) can be performed. The resulting hypothesis will then be applied to test points in the same vector space, in order to make predictions. Many previous studies [2,22,29] have shown that naturally occurring data cannot possibly fill up the high dimensional space uniformly, rather they must concentrate around lower dimensional structure. However, all the above-mentioned active learning heuristics fail to take into account the intrinsic manifold structure.

In this paper, we propose a novel active learning algorithm which explicitly considers the case that the data are sampled from a low dimensional sub-manifold embedded in the high dimensional ambient space. We follow the idea behind TED [33] which finds the most representative data points while the representativeness in our approach is defined explicitly on the data manifold. Inspired by the pioneering manifold learning work [29], we construct a nearest neighbor graph from the data and use the shortest path distance of the two corresponding vertices to approximate the geodesic distance of two data points along the manifold. With the estimated pairwise geodesic distances, we select the data points which can best "cover" the entire data set. It

* Corresponding author.
E-mail addresses: licheng@zju.edu.cn (C. Li), haifengliu@zju.edu.cn (H. Liu), dengcai@cad.zju.edu.cn, dengcai@gmail.com (D. Cai).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.08.002

is worthwhile to point out that the input of our algorithm can be only a graph. In reality, many real world applications have graph represented data and our algorithm suggests an approach for active learning on graphs.

The rest of the paper is organized as follows: in Section 2, we provide a brief review of the related work. Our manifold based active learning algorithm is introduced in Section 3. The experimental results are presented in Section 4. Finally, we provide the concluding remarks in Section 5.

2. Related work

There has been extensive research on the subject of active learning [4,6,11,12,14,16–19,32]. The generic problem of active learning is the following. Given a set of points $\mathbf{X}=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\}$ in $\mathbb{R}^m$, find a subset $\mathbf{Z}=\{\mathbf{z}_1,\mathbf{z}_2,\ldots,\mathbf{z}_k\}\subset\mathbf{X}$ which contains the most informative points.

Existing approaches for active learning can roughly be divided into two groups. The first group of algorithms selects the most uncertain data given previously trained models [9,30]. One representative algorithm in this group is SVM_{Active} [30,31]. This method selects the points that can reduce the size of the version space as much as possible. Since it is difficult to measure the version space, the authors provide three approximations. One of them which selects the points closest to the current decision boundary is called SimpleMargin. This method was also proposed by [24] and has been very popular. The second group of algorithms exploits the cluster structure of the data and selects the most representative points. For example, Dasgupta and Hsu [8] use hierarchical clustering to select the representative points. Some other methods include query-by-committee [28], density-weighted methods [21,27], and explicit error-reduction techniques [23,34]. Refer [26] for a comprehensive treatment of active learning approaches.

In statistics, the problem of selecting samples to label is typically referred to as experimental design. The sample $\mathbf{x}$ is referred to as experiment, and its label y is referred to as measurement. The study of optimal experimental design (OED) [1] is concerned with the design of experiments that are expected to minimize variances of a parameterized model. Classical experimental design approaches include A-optimal design, D-optimal design, E-optimal design. One limitation of these approaches is that there is no guarantee that these approaches will select either most uncertain points or most representative points.

In [12], He et al. have proposed a new approach called Graph Regularized Experimental Design (GRED). GRED is fundamentally based on the D-optimal design. Instead of using least squares classifier (which is the foundation of A-, D-, E-optimal design), GRED uses Laplacian Regularized Least Squares (LapRLS) classifier [3]. Since LapRLS considers the local geometric structure of the data, GRED is also expected to capture the local geometric information. Since GRED is based on the D-optimal design, it has the same limitation as D-optimal design.

Recently, Yu et al. have proposed Transductive Experimental Design (TED) [33] which is closely related to the I-optimal design [1]. TED tries to minimize the average predictive variance of the estimated function, which has been showed that is equivalent to finding a set of points such that other points can be best approximated by linear combination of the selected points [33]. The equivalent objective function is as follows:

$$\min_{\mathbf{Z},\mathbf{A}} \sum_{i=1}^{n}(\|\mathbf{x}_i-\mathbf{Z}\mathbf{a}_i\|^2+\mu\|\mathbf{a}_i\|^2)$$

$$\text{s.t.}\quad \mathbf{Z}=[\mathbf{z}_1,\ldots,\mathbf{z}_k]\subset=[\mathbf{x}_1,\ldots,\mathbf{x}_n]$$

$$\mathbf{A}=[\mathbf{a}_1,\ldots,\mathbf{a}_n]\in\mathbb{R}^{k\times n}$$

where $\mathbf{X}=[\mathbf{x}_1,\ldots,\mathbf{x}_n]$ are n data points, $\mathbf{Z}=[\mathbf{z}_1,\ldots,\mathbf{z}_k]$ are k points selected from $\mathbf{X}$ and $\mathbf{a}_i$ is the coefficient for reconstructing $\mathbf{x}_i$ using $\mathbf{Z}$.

This observation connects TED and those methods which exploit the cluster structure of the data and makes TED significantly different from the traditional A-, D-, E-optimal design.

3. Active learning on manifold

Recent studies [2,22,29] have shown that naturally occurring data cannot possibly fill up the high dimensional space uniformly, rather they must concentrate around lower dimensional structure. In this section, we introduce our novel active learning approach which explicitly considers the data manifold in question. Since our approach is fundamentally based on the differential geometry, we begin with a brief description of the basic geometrical concepts. See [20] for a detailed treatment.

3.1. Riemannian manifolds

Manifolds are generalizations of curves and surfaces to arbitrarily many dimensions. The formal definition of manifold is as follows.

Definition 3.1. A d-dimensional manifold (denoted by $\mathcal{M}^d$) is a topological space that is locally Euclidean. That is, around every point, there is a neighborhood that is topologically the same as the open unit ball in $\mathbb{R}^d$.

In order to compute distances on the manifold, one needs to equip a metric to the topological manifold. A manifold possessing a metric is called *Riemannian Manifold*, and the metric is commonly referred to as *Riemannian Metric*.

Definition 3.2. Suppose for every point $\mathbf{x}$ in a manifold $\mathcal{M}$, an inner product $\langle\cdot,\cdot\rangle_x$ is defined on a tangent space $T_x\mathcal{M}$ of $\mathcal{M}$ at $\mathbf{x}$. Then the collection of all these inner products is called the *Riemannian metric*.

Note that a Riemannian metric is not a distance metric on $\mathcal{M}$. However, for a connected manifold, it is the case that every Riemannian metric induces a distance matric on $\mathcal{M}$, *i.e. Geodesic Distance*.

Definition 3.3. The geodesic distance $d_{\mathcal{M}}(a,b)$ is defined as the length of the shortest curve connecting a and b.

In the plane, the geodesics are straight lines. On the sphere, the geodesics are great circles (like the equator).

3.2. Formulation

Given a set of points $\mathcal{X}=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\}$ in $\mathbb{R}^m$ sampled from the underlying manifold $\mathcal{M}$, the problem of active learning on manifold can be defined as finding a most "representative" (with respect to the manifold) subset $\mathcal{Z}=\{\mathbf{z}_1,\mathbf{z}_2,\ldots,\mathbf{z}_k\}\subset\mathcal{X}$.

If the underlying manifold is known, we can compute the geodesic distance between each pair of the sample points. Given a set of points $\mathcal{A}=\{\mathbf{x}_1,\ldots,\mathbf{x}_k\}$, we define the geodesic distance from a point $\mathbf{x}_i$ to the set $\mathcal{A}$ as the shortest geodesic distance from the point $\mathbf{x}_i$ to any point belongs to $\mathcal{A}$. Then we define the most "representative" subset $\mathcal{Z}$ as the subset which has minimized average geodesic distance to all the remaining points in $\mathcal{X}$.

In real life data sets, the underlying manifold $\mathcal{M}$ is often unknown. One hopes then to estimate geometrical and topological properties of the manifold from random samples ("scattered data") lying on this unknown manifold. Inspired by the pioneering