Neurocomputing 142 (2014) 429–437

Contents lists available at ScienceDirect

Neurocomputing

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

Manifold optimal experimental design via dependence maximization for active learning

Ping Li [a,*], Jiajun Bu [b], Chun Chen [b], Deng Cai [b]

[a] School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China
[b] College of Computer Science, Zhejiang University, Hangzhou 310027, China

ARTICLE INFO

Article history:
Received 29 May 2013
Received in revised form
7 November 2013
Accepted 6 April 2014
Communicated by Qingshan Liu
Available online 17 May 2014

Keywords:
Optimal experimental design
Dependence maximization
Manifold learning
Hilbert–Schmidt independence criterion
Image retrieval

ABSTRACT

Naturally occurring data have been growing in a huge volume size, which poses a big challenge to give them high-quality labels to learn a good model. Therefore, it is critical to only select the most informative data points for labeling, which is cast into the framework of active learning. We study this problem in a regression model from *optimal experimental design* (OED). To this end, several OED based methods have been developed, but the relations between the data points and their predictions are still not fully explored. Inspired by this, we employ the *Hilbert–Schmidt independence criterion* (HSIC) to maximize the dependence between the samples and their estimations in a global view. Thus, we present a novel active learning method named *manifold optimal experimental design via dependence maximization* (MODM). Specifically, those points having maximum dependence with their predictions are expected to be included for labeling. Besides, it utilizes the graph Laplacian to preserve the locally geometrical structure of the data. In this way, the most informative data points can be better selected. Moreover, we adopt a sequential strategy to optimize the objective function. The effectiveness of the proposed algorithm has been experimentally verified in content-based image retrieval.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

During the past decades, hundreds of thousands of data have emerged in an extensive range of fields and have been applied to numerous real-world tasks. Nevertheless, the majority of them has no access to labels, which require heavy loads and costly expert knowledge. In this regard, it becomes a crucial demand to select a much smaller subset of points characterizing the most information from the data collection. In the machine learning community, it is treated as an active learning problem [7,23], which has received lots of interests from both academia and industry. For example, the merits of active learning in multimedia annotation, image retrieval [28] and video indexing [31] have been empirically demonstrated.

To this end, the popular principles adopted in active learning include uncertainty sampling, query by committee, error reduction and variance reduction. Typically, the uncertainty sampling rule has been applied to support vector machine (SVM) [27], nearest neighbor classifier [19], *etc*. With this method, the most uncertain samples are queried for labeling. The variance reduction criterion originates from optimal experimental design (OED) [1], which refers to the problem of selecting samples to label in statistics. In the experimental design, the sample and its label are respectively seen as experiment and measurement. OED aims to minimize variances of a parameterized model, *e.g.*, minimizing the variance of the model parameters leads to A-, D- and E-optimal design while minimizing the variance of the estimated value leads to I- and G-optimal design [1]. However, these methods belong to the supervised paradigm, which does not consider the unmeasured (*i.e.*, unlabeled) samples. To overcome this drawback, some methods utilize both measured and unmeasured samples to actively select the most informative points, *e.g.*, Transductive Experimental Design (TED) [30] evaluates the average prediction variance on the pre-given unseen data based on I-optimal design. Nevertheless, TED does not consider the local manifold structure of the data space, which is of vital importance in active learning, since naturally occurring data often reside on a lower dimensional sub-manifold of the ambient Euclidean space [3,16,18]. To handle this deficit, Laplacian regularized D-optimal design (LapRDD) [13] was proposed, where the loss function is defined on both labeled and unlabeled points with an imposed locality preserving regularizer, which has been adopted in several learning methods to improve the performance [15,17]. Overall, the above methods do not fully consider the correlation between the unlabeled data and their estimated predictions, *i.e.*, existing models only well respect the labeled data whereas the dependence between the unlabeled data and their predictions has not been explored in the context of OED.

Recently, the *Hilbert–Schmidt independence criterion* (HSIC) [11], which measures the dependence between two random variables, has been successfully applied to many real-world applications,

* Corresponding author.
E-mail address: patriclouis.lee@gmail.com (P. Li).

http://dx.doi.org/10.1016/j.neucom.2014.04.018
0925-2312/© 2014 Elsevier B.V. All rights reserved.

such as feature selection [24], dimensionality reduction [34], classification [2] and clustering [25]. These methods take advantage of HSIC to maximize the dependence between the input data (*e.g.*, the feature) and the output (*e.g.*, the label), leading to the improved performance. However, they do not consider the local geometry that reflects the intrinsic structure of the data space, which is able to refine the learning performance. Essentially, HSIC is an empirical estimate of the Hilbert–Schmidt norm of the cross-covariance operator and has several advantages. First of all, it has a simple formulation as the trace of the product of Gram matrices. Besides, its rate of converging to the population estimate is conversely proportional to the square root of the number of samples. In addition, if the sample size is large, any existing dependence between the random variables is guaranteed to be revealed with a high probability [11]. Naturally, these merits can be sufficiently employed in OED to better model the correlation between unlabeled data and their estimated predictions. But such a kind of the correlation cannot be revealed by the existing LapRDD, which uses the linear regression to model the relation only between the labeled data and their labels.

Motivated by this, we adopt HSIC in Laplacian regularized OED to improve the performance of active learning. In this way, we propose a novel active learning method named manifold optimal experimental design via dependence maximization (MODM). The central idea is to take advantage of HSIC to measure the dependence between the input data and their estimated outputs. Particularly, in virtue of the regression model, we maximize the inherent dependence between the feature vectors and the corresponding predictions under the OED framework. In some sense, the dependence maximization reflects the relation between the inputs and the outputs globally. Furthermore, since MODM is developed upon LapRDD, thus inheriting the locality preserving property by the graph Laplacian. On the whole, a significant merit of MODM is that both the dependence maximization and the locally geometrical structure of the data are well respected in a unified model. This way, the most informative data points can be better selected for labeling, thus yielding an improved model. To investigate the performance of MODM, we apply it to a natural application of active learning, *i.e.*, relevance feedback in image retrieval [22,35]. Empirical studies have demonstrated the superiority of the proposed method compared to other alternatives.

It is worthwhile to highlight the main contributions of this work as follows:

- A novel active learning named MODM is presented by incorporating HSIC regularizer to manifold optimal experimental design, which enables maximizing the dependence between the unlabeled samples and their estimations. Thus, not only the relations between labeled samples and their labels are considered by a linear regression model, but also the dependence between unlabeled samples and their estimations, in addition to the manifold structure of the data space, is together respected for OED in a unified framework.
- Detailed derivations of the proposed method including a sequential optimization method are given with the time complexity analysis. Moreover, we generalize it to the nonlinear situation, *i.e.*, in Reproducing Kernel Hilbert Spaces (RKHS), thus being performed for linearly nonseparable data points.
- We have applied our method to content-based image retrieval (CBIR) on two real-world databases to investigate its performance. Experimental results have demonstrated the superiority of the proposed approach in terms of several evaluation metrics.

The remainder of this paper is organized as follows. We briefly review the related works in Section 2. Section 3 introduces the proposed manifold optimal experimental design via a dependence maximization algorithm as well as its nonlinear extension described in Section 4. Section 5 reports comprehensively the experimental results and finally we reach a conclusion in Section 6.

2. Related works

In this work, we focus on active learning, which is a hot topic in the machine learning community [8,12,23,32]. The relationship between active learning and semi-supervised learning [5,37] is just like one coin has two sides. They share the goal of relieving the boring tasks of labeling the unlabeled data. Here, we discuss active learning under the framework of OED. The problem setting can be stated as follows. Given a data collection with n samples in $\mathbb{R}^d$, *i.e.*, $\mathcal{X}=\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$, we aim to find a subset $\mathcal{Z}=\{\mathbf{z}_1,\ldots,\mathbf{z}_k\}\subset\mathcal{X}$ which covers the most informative points. The selected points are expected to improve the classifier the most if they are labeled for training the model.

Recently, OED has attracted considerable interests due to its solid theory foundation and practical successes [1,4,31]. For example, Zha et al. [31] proposed an active learning approach based on OED for video indexing, where they exploit the local structure of the data and also the sample density, relevance and diversity information, as well as both the labeled and the unlabeled data, resulting in promising retrieval performance. Active learning is often referred to experimental design in statistics, and many OED based methods have been developed, such as CLapRID [33], LapGOD [6], LapRDD [13], and HOD [20]. Given a data point $\mathbf{x}$, these methods usually consider a linear regression model $y=\mathbf{w}^T\mathbf{x}+\epsilon$, where y is the observation, $\mathbf{w}$ is the weight vector, ϵ is an independent Gaussian random variable with zero mean and constant variance σ^2. They attempt to learn a linear function $f(\mathbf{x})=\hat{\mathbf{w}}^T\mathbf{x}$. Assuming that a set of labeled samples $\{(\mathbf{z}_i,y_i)\}_{i=1}^k$ are available, the least squares is always used as the cost function to estimate $\mathbf{w}$ by minimizing the residual sum of squares, *i.e.*,

$$RSS(\mathbf{w})=\sum_{i=1}^{k}(y_i-\mathbf{w}^T\mathbf{z}_i)^2. \tag{1}$$

Its optimal solution is $\hat{\mathbf{w}}=(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{Z}\mathbf{y}$, where $\mathbf{Z}=[\mathbf{z}_1,\ldots,\mathbf{z}_k]\in\mathbb{R}^{d\times k}$ and $\mathbf{y}=[y_1,\ldots,y_k]^T$. For an unseen data point, its output can be estimated by $\hat{y}=f(\mathbf{x})=\hat{\mathbf{w}}^T\mathbf{x}$. By Gauss–Markov theorem [9], it has been proved that $\hat{\mathbf{w}}$ is an unbiased estimation of $\mathbf{w}$ with the covariance matrix, namely

$$Cov(\hat{\mathbf{w}})=\sigma^2(\mathbf{Z}\mathbf{Z}^T)^{-1}, \tag{2}$$

where $\mathbf{Z}\mathbf{Z}^T$ is the Hessian of $RSS(\mathbf{w})$.

Roughly speaking, there exist two types of criteria of OED [1]. The first one is to minimize the confidence region of the estimated parameter $\hat{\mathbf{w}}$, which results in D-optimal design (determinant of H_{RSS}), A-optimal design (trace of H_{RSS}) and E-optimal design (maximum eigenvalue of H_{RSS}). The second one is to minimize the variance of the predicted value over some region of interest (ROI), which leads to I-optimal design (average predictive variance) and G-optimal design (maximum predictive variance).

In addition, some active learning methods are designed upon SVM [26,14], which select those data points closest to the boundary for labeling by considering that the uncertainty of the points near the boundary is larger than those far apart from the hyperplane. Nonetheless, there are some drawbacks of the SVM based active learning methods, *i.e.*, the boundary of different classes is difficult to estimate and it probably causes a failure without labeled points in the beginning. Thus, we present the novel method in the context of experimental design, which does not suffer from these limitations.

Download English Version:

https://daneshyari.com/en/article/412297

Download Persian Version:

https://daneshyari.com/article/412297

Daneshyari.com