Pattern Recognition 71 (2017) 337–348

Contents lists available at ScienceDirect

Pattern Recognition

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

Active learning based on minimization of the expected path-length of random walks on the learned manifold structure

Chin-Chun Chang[*], Bo-Han Liao

Department of Computer Science and Engineering, National Taiwan Ocean University, Keelung, 202, Taiwan

ARTICLE INFO

Article history:
Received 20 January 2017
Revised 8 May 2017
Accepted 1 June 2017
Available online 6 June 2017

Keywords:
Active learning
Locally linear embedding
Random walks
Submodular set functions

ABSTRACT

Active learning algorithms aim at selecting important samples to label for subsequent machine learning tasks. Many active learning algorithms make use of the reproducing kernel Hilbert space (RKHS) induced by a Gaussian radial basis function (RBF) kernel and leverage the geometrical structure of the data for query-sample selection. Parameters for the kernel function and the k-nearest-neighborhood graph must be properly set beforehand. As a tool exploring the structure of data, active learning algorithms with automatic tuning of those parameters are desirable. In this paper, local linear embedding (LLE) with convex constraints on neighbor weights is used to learn the geometrical structure of the data in the RKHS induced by a Gaussian RBF kernel. Automatic tuning of the kernel parameter is based on the assumption that the geometrical structure of the data in the RKHS is sparse and local. With the Markov matrix established based on the learned LLE weight matrix, the total expected path-length of the random walks from all samples to selected samples is proposed to be a criterion for query-sample selection. A greedy algorithm having a guaranteed solution bound is developed to select query samples and a two-phase scheme is also proposed for scaling the proposed active learning algorithm. Experimental results on data sets including hundreds to tens of thousands of samples have shown the feasibility of the proposed approach.

1. Introduction

Labeling every sample in a data set for machine learning tasks is expensive. Active learning algorithms attempt to reduce the labeling cost by selecting a few unlabelled samples to annotate such that the classifier trained based on the annotated sample is accurate. The active learning algorithm considered in this paper selects unlabelled samples in batches and aims at data sets of high data redundancy with no initial labelled samples. Many state-of-the-art active learning algorithms [1–11] leverage the geometrical structure of samples and the reproducing kernel Hilbert space (RKHS) induced by a Gaussian radial basis function (RBF) kernel for query-sample selection, and require to set hyper-parameters for the RBF kernel and the k-nearest-neighborhood (k-NN) graph beforehand. However, for the problem considered in this paper, because of no labelled samples, general users can have difficulty in setting the hyper-parameter. Active learning algorithms with automatic tuning of the hyper-parameter without labelled samples are rarely studied.

In this paper, a kernel version of locally linear embedding (LLE) [12] with convex constraints on the neighbor weights is used to learn the geometrical structure of the unlabelled sample in the RKHS induced by an RBF kernel. Unlabelled samples likely to be met by random walks based on the Markov matrix established by the learned LLE weight matrix are regarded as informative samples and selected for manual annotation. To scale the proposed active learning algorithm, a two-phase scheme is also developed. There exist active learning algorithms based on random walks [13,14]. In [13], samples are selected by considering the maximum probability, in an equilibrium state, that a random walker starts from the unselected sample and reaches the selected sample. This approach requires a user-feedback after selecting every query sample. In [14], the expected path-length of random walks from an unlabelled sample to the labelled sample of each class is used for evaluating the importance of an unlabelled sample. Due to no consideration of the relationship between unlabelled samples, this approach can select similar unlabelled samples as query samples when selecting a batch of samples. In contrast, the proposed approach can select samples in batches and tends to select query samples covering the geometrical structure of data because considering the minimum of the total expected path-length of the random walks from all samples to the selected sample.

In the literature, research on automatic tuning of the hyper-parameter mainly focuses on supervised learning [15–17], where the hyper-parameter optimizing some criterion functions in terms

[*] Corresponding author.
E-mail address: cvml@mail.ntou.edu.tw (C.-C. Chang).

http://dx.doi.org/10.1016/j.patcog.2017.06.001
0031-3203/© 2017 Elsevier Ltd. All rights reserved.

of labelled samples is considered to be appropriate to subsequent tasks of supervised learning. In this paper, the hyper-parameter associated with a spare and local geometrical structure of samples is preferable. Basically, the proposed approach requires to set the hyper-parameters for the RBF kernel and construction of the neighborhood of samples. In [18], it turns out that without construction of the neighborhood of samples, a sparse and local similarity matrix for Laplacian embedding can be learned by considering a ℓ^1-norm penalty on the similarity matrix. In this paper, with the inspiration of [18], the analysis result shows that without construction of the neighborhood of samples, sparse and local LLE weights can be learned with a proper setting for the parameter of the RBF kernel. Thus, only the parameter for the RBF kernel is required to tune and the appropriateness of the setting for this parameter can be evaluated by the sparsity of learned LLE weights.

The contributions of this work are twofold. First, a greedy algorithm is proposed to select query samples by minimization of the expected path-length of random walks. The selected query samples prove to have a guaranteed solution bound. Second, the proposed active learning algorithm can automatically tune the hyper-parameter without labelled samples. The sparsity of learned LLE weights turns out to be a promising criterion for tuning of the hyper-parameter for the proposed approach.

The remaining part of this paper is organized as follows. Section 2 reviews related active learning algorithms. Section 3 introduces the proposed approach and a two-phase scheme for scaling the proposed active learning algorithm. Section 4 presents the step of tuning the kernel parameter. Section 5 shows experimental results. Concluding remarks are drawn at last.

2. Related work

A recent trend in active learning focuses on pool-based settings [19,20]. Pool-based active learning can be further classified into single instance selection and batch selection [21]. Active learning algorithms of single instance selection, such as [22,23], inquire the label of the most uncertain sample with respect to a classification model and require frequent model retraining. In contrast, batch-mode active learning algorithms can choose several query samples simultaneously [21]. In the last decade, batch-mode active learning algorithms making use of manifold learning [1–6] and the RKHS [1,6–11] have shown to be effective to data sets of complicated geometrical structures.

In [1], active learning is based on manifold-regularized D-optimal experimental design, and query samples are selected by minimization of the variance of a Laplacian regularized regression model. In [6], based on manifold-regularized D-optimal experimental design, the Hilbert-Schmidt independence criterion is also considered to strengthen the dependence between sample points and their predictions.

In contrast to classic experimental design, which only evaluates the expected prediction error on selected samples, transductive experimental design (TED) [7] also takes into account the expected prediction error on unselected samples. Extensions of TED making use of the geometrical structure of data have come out. In [8], TED is based on a manifold adaptive kernel [24], which incorporates the manifold structure into the RKHS. In [3], TED is based on the manifold structure learned by LLE. In [4], TED is extended by localizing the reconstruction of a sample.

Query samples can be selected by minimization of the difference in the distribution between the selected and the unselected sample. In [9], the maximum mean discrepancy [25] is used for measuring that distribution difference. In [10,11], informative and representative unlabelled samples are selected as query samples. It turns out that a proper balance of the criteria regarding informativeness and representativeness of samples can boost the active learning performance.

Query samples can also be selected by evaluating the importance of unlabelled samples to the learned manifold structure. In [2], query samples are selected based on the clustering coefficient measure. In [5], query samples are selected by minimizing the total shortest-path length between the unselected and the selected samples in the k-NN graph. In [3], the importance of samples to the learned manifold structure is analyzed more thoroughly by considering all paths in the k-NN graph, and this strategy is also employed in this study.

Properly setting the parameters for nonlinear kernel functions and manifold learning is crucial for the aforementioned active learning algorithms. Automatic tuning of those hyper-parameters for active learning algorithms usually requires initial labelled samples, such as [26,27]. Active learning algorithms are desired to have the capability of automatic tuning of those hyper-parameters without labelled samples.

3. Methodology

Let $\mathcal{X} \triangleq \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be a set of n data points in $\mathbb{R}^d$ and $\tilde{\mathbf{W}} \triangleq [\tilde{w}_{ij}] \in \mathbb{R}^{n\times n}$ be the Markov matrix for random walks, where $\tilde{w}_{ij}$ is the transition probability from point j to point i with $\sum_{i=1}^{n} \tilde{w}_{ij} = 1$. Denote by $\boldsymbol{\pi} \triangleq [\pi_i] \in \mathbb{R}^n$ the vector, where $\pi_i \in [0, 1]$ is the weight for $\mathbf{x}_i$. Define a criterion function $h(S)$ as

$$h(S) = \mathbf{1}^T(\mathbf{I} + \tilde{\mathbf{W}}_S + \tilde{\mathbf{W}}_S^2 + \tilde{\mathbf{W}}_S^3 + \ldots)\boldsymbol{\pi} - n = \mathbf{1}^T(\mathbf{I} - \tilde{\mathbf{W}}_S)^{-1}\boldsymbol{\pi} - n, \quad (1)$$

where $\tilde{\mathbf{W}}_S$ is equal to $\tilde{\mathbf{W}}$ with the ith row and the ith column set to zero for every $\mathbf{x}_i \in S$. If $\boldsymbol{\pi} = \mathbf{1}$, $h(S)$ is the total expected path-length of the random walks beginning from the points in $\mathcal{X}$ and ending at a point in $S \subseteq \mathcal{X}$ by regarding the point in S as the absorbing state in an absorbing Markov chain [28]. For brevity, $h(S)$ is called the random-walk path-length with respect to S. If $h(S) \leq h(S')$, S is more informative than S' because S has more reduction in the uncertainty of random walks. In this study, the set S_ℓ of ℓ points such that $h(S_\ell)$ attains the minimum is selected for manual annotation. Selecting such ℓ points is at least as hard as the vertex cover problem in a general graph, an NP-complete problem, and thus a greedy algorithm for obtaining S_ℓ is proposed in this study.

The proposed algorithm has two main steps. The first main step establishes $\tilde{\mathbf{W}}$ and $\boldsymbol{\pi}$ based on a kernel version of LLE. The second main step selects S_ℓ from $\mathcal{X}$ by minimization of the criterion function $h(S)$. In the sequel, the step of establishing $\tilde{\mathbf{W}}$ is introduced first and followed by the proposed algorithm for selecting query points. Next, the time complexity of the proposed algorithm is analyzed and compared with several state-of-the-art algorithms. A two-phase scheme for scaling the proposed algorithm is presented at last.

3.1. Construction of the Markov matrix $\tilde{\mathbf{W}}$ Based on a kernel version of LLE

Denote by $\mathcal{N}_k(\mathbf{x}_i) \subseteq \mathcal{X}$ the k-neighbor set of $\mathbf{x}_i$, in which for every $\mathbf{x}$ in $\mathcal{N}_k(\mathbf{x}_i)$, either $\mathbf{x}$ is a k-NN of $\mathbf{x}_i$ or $\mathbf{x}_i$ is a k-NN of $\mathbf{x}$. LLE assumes that data lie on a manifold which can be locally linearly approximated. The neighbor weights are learned by solving the problem [12]:

$$\arg\min_{w_{ij}} \sum_{j=1}^{n} \left\| \mathbf{x}_j - \sum_{i=1}^{n} w_{ij}\mathbf{x}_i \right\|_2^2 \quad (2)$$

$$\text{subject to } \sum_{i=1}^{n} w_{ij} = 1, j = 1, \ldots, n,$$