Neurocomputing 139 (2014) 397–407

Contents lists available at ScienceDirect

Neurocomputing

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

An effective framework for supervised dimension reduction

Khoat Than [a,*,1], Tu Bao Ho [b,d], Duy Khuong Nguyen [b,c]

[a] Hanoi University of Science and Technology, 1 Dai Co Viet road, Hanoi, Vietnam
[b] Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
[c] University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam
[d] John von Neumann Institute, Vietnam National University, HCM, Vietnam

ARTICLE INFO

Article history:
Received 15 April 2013
Received in revised form
23 September 2013
Accepted 18 February 2014
Communicated by Steven Hoi
Available online 8 April 2014

Keywords:
Supervised dimension reduction
Topic models
Scalability
Local structure
Manifold learning

ABSTRACT

We consider supervised dimension reduction (SDR) for problems with discrete inputs. Existing methods are computationally expensive, and often do not take the local structure of data into consideration when searching for a low-dimensional space. In this paper, we propose a novel framework for SDR with the aims that it can inherit scalability of existing unsupervised methods, and that it can exploit well label information and local structure of data when searching for a new space. The way we encode local information in this framework ensures three effects: preserving inner-class local structure, widening inter-class margin, and reducing possible overlap between classes. These effects are vital for success in practice. Such an encoding helps our framework succeed even in cases that data points reside in a nonlinear manifold, for which existing methods fail.

The framework is general and flexible so that it can be easily adapted to various unsupervised topic models. We then adapt our framework to three unsupervised models which results in three methods for SDR. Extensive experiments on 10 practical domains demonstrate that our framework can yield scalable and qualitative methods for SDR. In particular, one of the adapted methods can perform consistently better than the state-of-the-art method for SDR while enjoying 30–450 times faster speed.

1. Introduction

In supervised dimension reduction (SDR), we are asked to find a low-dimensional space which preserves the predictive information of the response variable. Projection on that space should keep the discrimination property of data in the original space. While there is a rich body of researches on SDR, our primary focus in this paper is on developing methods for discrete data. At least three reasons motivate our study: (1) current state-of-the-art methods for continuous data are really computationally expensive [1–3], and hence can only deal with data of small size and low dimensions; (2) meanwhile, there are excellent developments which can work well on discrete data of huge size [4,5] and extremely high dimensions [6], but are unexploited for supervised problems; (3) further, continuous data can be easily discretized to avoid sensitivity and to effectively exploit certain algorithms for discrete data [7].

Topic modeling is a potential approach to dimension reduction. Recent advances in this new area can deal well with huge data of very high dimensions [4–6]. However, due to their unsupervised nature, they do not exploit supervised information. Furthermore, because the local structure of data in the original space is not considered appropriately, the new space is not guaranteed to preserve the discrimination property and proximity between instances. These limitations make unsupervised topic models unappealing to supervised dimension reduction.

Investigation of local structure in topic modeling has been initiated by some previous researches [8–10]. These are basically extensions of *probabilistic latent semantic analysis* (PLSA) by Hoffman [11], which take local structure of data into account. Local structures are derived from nearest neighbors, and are often encoded in a graph. Those structures are then incorporated into the likelihood function when learning PLSA. Such an incorporation of local structures often results in learning algorithms of very high complexity. For instances, the complexity of each iteration of the learning algorithms by Wu et al. [8] and Huh and Fienberg [9] is *quadratic* in the size M of the training data; and that by Cai et al. [10] is *triple* in M because of requiring a matrix inversion. Hence these developments, even though often being shown to work well, are very limited when the data size is large.

Some topic models [12–14] for supervised problems can do simultaneously two nice jobs. One job is derivation of a meaningful space which is often known as "*topical space*". The other is that supervised information is explicitly utilized by max-margin approach [14] or likelihood maximization [12]. Nonetheless, there are two

[*] Corresponding author.
E-mail addresses: khoattq@soict.hust.edu.vn (K. Than), bao@jaist.ac.jp (T.B. Ho), khuongnd@jaist.ac.jp (D.K. Nguyen).

[1] This work was done when the author was at JAIST.

http://dx.doi.org/10.1016/j.neucom.2014.02.017
0925-2312/© 2014 Elsevier B.V. All rights reserved.

common limitations of existing supervised topic models. First, the local structure of data is not taken into account. Such an ignorance can hurt the discrimination property in the new space. Second, current learning methods for those supervised models are often very expensive, which is problematic with large data of high dimensions.

In this paper, we approach to SDR in a novel way. Instead of developing new supervised models, we propose the *two-phase* framework which can inherit scalability of recent advances for unsupervised topic models, and can exploit label information and local structure of the training data. The main idea behind the framework is that we first learn an unsupervised topic model to find an initial topical space; we next project documents on that space exploiting label information and local structure, and then reconstruct the final space. To this end, we employ the Frank–Wolfe algorithm [15] for fast doing projection/inference.

The way of encoding local information in this framework ensures three effects: preserving inner-class local structure, widening inter-class margin, and reducing possible overlap between classes. These effects are vital for success in practice. We find that such encoding helps our framework succeed even in cases that data points reside in a nonlinear manifold, for which existing methods might fail. Further, we find that ignoring either label information (as in [9]) or manifold structure (as in [14,16]) can significantly worsen quality of the low-dimensional space. This finding complements a recent theoretical study [17] which shows that, for some semi-supervised problems, using manifold information would definitely improve quality.

Our framework for SDR is general and flexible so that it can be easily adapted to various unsupervised topic models. To provide some evidences, we adapt our framework to three models: *probabilistic latent semantic analysis* (PLSA) by Hoffman [11], *latent Dirichlet allocation* (LDA) by Blei et al. [18], and *fully sparse topic models* (FSTM) by Than and Ho [6]. The resulting methods for SDR are respectively denoted as PLSA^c, LDA^c, and FSTM^c. Extensive experiments on 10 practical domains show that PLSA^c, LDA^c, and FSTM^c can perform substantially better than their unsupervised counterparts.[2] They perform comparably or better than existing methods that base either on max-margin principle such as MedLDA [14] or on manifold regularization without using labels such as DTM [9]. Further, PLSA^c and FSTM^c consume significantly less time than MedLDA and DTM to learn good low-dimensional spaces. These results suggest that the two-phase framework provides a competitive approach to supervised dimension reduction.

Organization: In the next section, we describe briefly some notations, the Frank–Wolfe algorithm, and related unsupervised topic models. We present the proposed framework for SDR in Section 3. We also discuss in Section 4 the reasons why label information and local structure of data can be exploited well to result in good methods for SDR. Empirical evaluation is presented in Section 5. Finally, we discuss some open problems and conclusions in the last section.

2. Background

Consider a corpus $\mathcal{D} = \{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_M\}$ consisting of M documents which are composed from a vocabulary of V terms. Each document $\boldsymbol{d}$ is represented as a vector of term frequencies, i.e. $\boldsymbol{d} = (d_1, \ldots, d_V) \in \mathbb{R}^V$, where d_j is the number of occurrences of term j in $\boldsymbol{d}$. Let $\{y_1, \ldots, y_M\}$ be the class labels assigned to those documents. The task of *supervised dimension reduction* (SDR) is to find a new space of K dimensions which preserves the predictiveness of the response/label variable Y. Loosely speaking, predictiveness preservation requires that projection of data points onto the new space should preserve separation (discrimination) between classes in the original space, and that proximity between data points is maintained. Once the new space is determined, we can work with projections in that low-dimensional space instead of the high-dimensional one.

2.1. Unsupervised topic models

Probabilistic topic models often assume that a corpus is composed of K topics, and each document is a mixture of those topics. Example models include PLSA [11], LDA [18], and FSTM [6]. Under a model, each document has another latent representation, known as *topic proportion*, in the K-dimensional space. Hence topic models play a role as dimension reduction if $K < V$. Learning a low-dimensional space is equivalent to learning the topics of a model. Once such a space is learned, new documents can be projected onto that space via *inference*. Next, we describe briefly how to learn and to do inference for three models.

2.1.1. PLSA

Let $\theta_{dk} = P(z_k|\boldsymbol{d})$ be the probability that topic k appears in document $\boldsymbol{d}$, and $\beta_{kj} = P(w_j|z_k)$ be the probability that term j contributes to topic k. These definitions basically imply that $\sum_{k=1}^{K} \theta_{dk} = 1$ for each $\boldsymbol{d}$, and $\sum_{j=1}^{V} \beta_{kj} = 1$ for each topic k. The PLSA model assumes that document $\boldsymbol{d}$ is a mixture of K topics, and $P(z_k|\boldsymbol{d})$ is the proportion that topic k contributes to $\boldsymbol{d}$. Hence the probability of term j appearing in $\boldsymbol{d}$ is $P(w_j|\boldsymbol{d}) = \sum_{k=1}^{K} P(w_j|z_k)P(z_k|\boldsymbol{d}) = \sum_{k=1}^{K} \theta_{dk}\beta_{kj}$. Learning PLSA is to learn the topics $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$. Inference of document $\boldsymbol{d}$ is to find $\boldsymbol{\theta}_d = (\theta_{d1}, \ldots, \theta_{dK})$.

For learning, we use the EM algorithm to maximize the likelihood of the training data:

$$\text{E-step}: \quad P(z_k|\boldsymbol{d}, w_j) = \frac{P(w_j|z_k)P(z_k|\boldsymbol{d})}{\sum_{l=1}^{K} P(w_j|z_l)P(z_l|\boldsymbol{d})}, \tag{1}$$

$$\text{M-step}: \quad \theta_{dk} = P(z_k|\boldsymbol{d}) \propto \sum_{v=1}^{V} d_v P(z_k|\boldsymbol{d}, w_v), \tag{2}$$

$$\beta_{kj} = P(w_j|z_k) \propto \sum_{\boldsymbol{d} \in \mathcal{D}} d_j P(z_k|\boldsymbol{d}, w_j). \tag{3}$$

Inference in PLSA is not explicitly derived. Hoffman [11] proposed an adaptation from learning: keeping topics fixed, iteratively do the steps (1) and (2) until convergence. This algorithm is called *folding-in*.

2.1.2. LDA

Blei et al. [18] proposed LDA as a Bayesian version of PLSA. In LDA, the topic proportions are assumed to follow a Dirichlet distribution. The same assumption is endowed over topics $\boldsymbol{\beta}$. Learning and inference in LDA are much more involved than those of PLSA. Each document $\boldsymbol{d}$ is independently inferred by the variational method with the following updates:

$$\phi_{djk} \propto \beta_{kw_j} \exp \Psi(\gamma_{dk}), \tag{4}$$

$$\gamma_{dk} = \alpha + \sum_{d_j > 0} \phi_{djk}, \tag{5}$$

where ϕ_{djk} is the probability that topic i generates the jth word w_j of $\boldsymbol{d}$; γ_d is the variational parameters; Ψ is the digamma function; α is the parameter of the Dirichlet prior over $\boldsymbol{\theta}_d$.

Learning LDA is done by iterating the following two steps until convergence. The E-step does inference for each document. The M-step maximizes the likelihood of data w.r.t. $\boldsymbol{\beta}$ by the following

[2] Note that due to being dimension reduction methods, PLSA, LDA, FSTM, PLSA^c, LDA^c, and FSTM^c themselves cannot directly do classification. Hence we use SVM with a linear kernel for doing classification tasks on the low-dimensional spaces. Performance for comparison is the accuracy of classification.