Pattern Recognition 52 (2016) 17–32

Contents lists available at ScienceDirect

Pattern Recognition

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

Infinite max-margin factor analysis via data augmentation

Xuefeng Zhang [a], Bo Chen [a,b,*], Hongwei Liu [a,b], Lei Zuo [a], Bo Feng [a]

[a] National Lab of Radar Signal Processing, Xidian University, Xi'an, Shaanxi, China, 710071
[b] Collaborative Innovation Center of Information Sensing and Understanding at Xidian University, Xi'an, Shaanxi, China, 710071

ARTICLE INFO

Article history:
Received 20 September 2014
Received in revised form
2 September 2015
Accepted 29 October 2015
Available online 10 November 2015

Keywords:
Latent variable support vector machine
Factor analysis
Dirichlet process mixture
Classification and rejection performance

ABSTRACT

This paper addresses the Bayesian estimation of the discriminative probabilistic latent models, especially the mixture models. We develop the max-margin factor analysis (MMFA) model, which utilizes the latent variable support vector machine (LVSVM) as the classification criterion in the latent space to learn a discriminative subspace with max-margin constraint. Furthermore, to deal with multimodally distributed data, we further extend MMFA to infinite Gaussian mixture model and develop the infinite max-margin factor analysis (iMMFA) model, via the consideration of Dirichlet process mixtures (DPM). It jointly learns clustering, max-margin classifiers and the discriminative latent space in a united framework to improve the prediction performance. Moreover, both of MMFA and iMMFA are natural to handle outlier rejection problem, since the observations are described by a single or a mixture of Gaussian distributions. Additionally, thanks to the conjugate property, the parameters in the two models can be inferred efficiently via the simple Gibbs sampler. Finally, we implement our models on synthesized and real-world data, including multimodally distributed datasets and measured radar echo data, to validate the classification and rejection performance of the proposed models.

1. Introduction

Probabilistic latent models have been widely utilized to discover latent structures and reveal hidden explanatory factors for complex data in statistics and machine learning [1–6]. These models, which project an observation into a low dimensional space, have the ability of exploring and keeping some useful data information in the new space. Factor analysis (FA) is a typical example of the probabilistic latent models. It has been used extensively as a data analytic technique to examine patterns of interrelationship, data reduction, classification and description of data [1–3,5,6]. In FA, latent factors can be regarded as a low dimensional representation of the observations in a latent subspace. Nevertheless, FA is an unsupervised model without utilization of any label information and only focuses on the observations [1–3]. Recently, considering supervising information for learning predictive latent features has attracted a lot of attentions, where the inferred discriminant latent features are considered as input features [7–12]. Yu et al. [7] propose a linear supervised probabilistic PCA, however, it is developed for real outputs and only considers the classification problem as the imputation of missing values without using any classification criterion. Lacoste-Julien [8] proposes a discriminative variation on Latent Dirichlet Allocation (LDA), called DiscLDA, which is trained by maximizing the conditional likelihood of response variables. The Maximum Entropy Discrimination Latent Dirichlet Allocation (MedLDA) model seeks a regularized posterior distribution of the predictive function in a feasible space [9]. In MedLDA, the predictive function is defined by a set of expected margin constraints generalized from the SVM-style margin constraints. Furthermore, by employing the latent variable representation of SVM (LVSVM) [13], Gibbs MedLDA [10,11] is proposed with an efficient inference algorithm-Gibbs sampling [14,15]. Those three supervised LDA models focus on the supervised probabilistic topic models for dimensionality reduction in collections of text documents or images (represented by bag of words model) rather than the continuous data. In [12], Zhu et al. develop the regularized Bayesian inference (RegBayes) principle and present two concrete examples of RegBayes, infinite latent support vector machines (iLSVM) and multi-task infinite latent support vector machines (MT-iLSVM), where max-margin constraints are introduced to improve the discriminative power of a Bayesian model.

For multimodally distributed database in many real-world problems, simple linear classifier cannot provide a well discriminative boundary. Though kernel method classifiers can handle linearly inseparable data, they need to build the kernel matrix with all training data, which leads to computational and storage

* Corresponding author.
E-mail address: bchen@mail.xidian.edu.cn (B. Chen).

http://dx.doi.org/10.1016/j.patcog.2015.10.020
0031-3203/

burden. Therefore, some approximate realizations of SVM, like SimpleSVM [16] and the core vector machines (CVM)[17], have been proposed to reduce complexities caused by kernel matrix. As an alternative way, the mixture-of-expert (ME) strategy is proposed to discover underlying descriptive patterns and improve efficiency, which partitions the input data into finite clusters and then learns a linear classifier within each cluster [18,19]. Moreover, to deal with the model selection problem, nonparametric Bayesian technique, especially the DP mixture (DPM) model, has been introduced into ME models [20–24]. For example, Shahbaba et al. [21] build a nonlinear model based on a DP mixture of multinomial logit (MNL), which is denoted as dpMNL. In [22], Hannah et al. propose Dirichlet process mixtures of generalized linear models (DP-GLM), a new class of methods for nonparametric regression. Zhu and his colleagues propose infinite SVM (iSVM), a DP mixture of kernel SVM [23] also with the Gibbs sampler [24]. Nevertheless, those methods work in original space, which is not practical to handle high dimensional data without feature extraction procedure.

In this paper, we first develop max-margin factor analysis (MMFA), which jointly learns the discriminative subspace and max-margin classifier. The interplay between the likelihood function of the observation modeled by FA and maximum margin constraint induced by LVSVM can yield latent representations that are more discriminative and reasonable for supervised prediction tasks. Furthermore, to handle multimodally distributed databases, we develop infinite max-margin factor analysis (iMMFA) by introducing DPM into MMFA, which divides the dataset into 'infinite' clusters in the learned latent space and learns a LVSVM classifier on each cluster jointly. Both MMFA and iMMFA capture the underlying structure of the observations in the subspace and employ LVSVM as the classifier with the low-dimensional latent representations as the input feature. Additionally, the observations in MMFA and iMMFA are modeled by a single or a group of Gaussian distributions. Thus, outlier rejection, which is important in many real recognition tasks, can be implemented in the prediction phase. The parameters of MMFA and iMMFA have good conjugacy conditioned on augmented variables and can be effectively inferred via the simple and efficient Gibbs sampler.

The remainder of this paper is structured as follows. In Section 2 we introduce the latent variable SVM and FA model, and then present MMFA. Further, we introduce DP and DPM, and then present Gibbs iMMFA in Section 3. In Section 4, experiments are conducted on synthetic, benchmark, and measured radar HRRP dataset to evaluate the effectiveness and efficiency of our models. Finally, the paper is concluded.

2. Max-margin factor analysis

The goal of the supervised probabilistic latent models is to learn a discriminative subspace guided by a given classifier and classification strategy. In classification, SVM, the best known example, is arguably more discriminative and has achieved a great success. Considering the conjugacy property in the Bayesian models, we would like to formulate a joint probabilistic model of latent models and supervised learning with a fully Bayesian treatment. However, the hinge loss makes SVM difficult to be modeled under the traditional Bayesian framework. Fortunately, Polson et al. [13] reformulate the SVM optimality criterion with the parameter regularization penalty as a mixture of normal pseudo-posterior distributions, which allow SVM to be analyzed with Bayesian treatments. In this case, Bayesian models and SVM can be jointly learned in a united framework. Based on Polson's work [13], we develop MMFA in this section. And we review LVSVM first.

2.1. Latent variable support vector machine (LVSVM)

Given a labeled training dataset $\{(\mathbf{x}_n, y_n)|\,\mathbf{x}_n \in \mathbb{R}^P, y_n \in \{-1,+1\}\}_{n=1}^N$, SVM describes a binary linear classification with the decision function $\hat{y}_n = \mathrm{sign}(\boldsymbol{\eta}^{\mathrm{T}}\tilde{\mathbf{x}}_n)$, where $\tilde{\mathbf{x}}_n = [\mathbf{x}_n; 1]$ is augmented feature vector and $\boldsymbol{\eta}$ is the weighted coefficient. If $\hat{y}_n \geq 0$, $\mathbf{x}_n$ is classified as a positive sample $(+1)$, else as a negative sample (-1). SVM is a max-margin method, where the margin is defined as the smallest distance between the decision boundary and any of the samples. In order to maximum the margin, SVM solves the problem

$$\begin{aligned} &\min_{\boldsymbol{\eta},\xi_n} \tfrac{1}{2}\|\boldsymbol{\eta}\|_2^2 + C_0\sum_{n=1}^{N}\xi_n \\ &\quad s.t. \quad y_n\boldsymbol{\eta}^{\mathrm{T}}\tilde{\mathbf{x}}_n \geq 1-\xi_n \\ &\qquad \xi_n \geq 0, n = 1,\ldots,N \end{aligned} \tag{1}$$

where C_0 is a positive tuning parameter. The underlying discriminative objective is a linear hinge loss function, $\max(1-y_n\boldsymbol{\eta}^{\mathrm{T}}\tilde{\mathbf{x}}_n, 0)$, which seems to make traditional Bayesian analysis difficult to model.

Conventionally, problem (1) can be solved by convex optimization algorithm. Unlike the conventional way, Polson et al. [13] present a latent variable representation of SVM. They present the pseudo-likelihood contribution from observation y_n,$\phi_n(y_n|\,\boldsymbol{\eta})$, as a location-scale mixture of normal to deal with hinge loss function. The pseudo-likelihood contribution can be expressed as [13]

$$\begin{aligned} \phi_n(y_n|\,\boldsymbol{\eta}) &= \exp\{-2C_0\max(1-y_n\boldsymbol{\eta}^T\tilde{\mathbf{x}}_n, 0)\} \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi\lambda_n}}\exp\left(-\frac{(\lambda_n + C_0(1-y_n\boldsymbol{\eta}^{\mathrm{T}}\tilde{\mathbf{x}}_n))^2}{2\lambda_n}\right)d\lambda_n \end{aligned} \tag{2}$$

$\phi_n(y_n|\,\boldsymbol{\eta})$ can be regarded as the marginal from a joint distribution $\phi_n(y_n, \lambda_n|\,\boldsymbol{\eta})$, which is conjugate to multivariate normal prior distribution.

In this paper, we employ a Student-t prior, implemented via the hierarchical construction of normal-gamma distribution on $\boldsymbol{\eta}$:

$$\boldsymbol{\eta} \sim \mathbf{N}(0, \sigma^{-1}\mathbf{I}), \quad \sigma \sim \mathrm{Ga}(a_0, b_0) \tag{3}$$

According to (2) and (3), the SVM pseudo-posterior distribution can be expressed as the marginal distribution of a higher dimensional distribution with the augmented variables $\boldsymbol{\lambda}$. Then we can write down the complete data pseudo-posterior distribution as

$$\begin{aligned} p(\boldsymbol{\eta}, \boldsymbol{\lambda}, \sigma|\,\mathbf{y}) &\propto \prod_{n=1}^{N}\phi_n(y_n, \lambda_n|\,\boldsymbol{\eta})p(\boldsymbol{\eta}|\,\sigma)p(\sigma) \\ &\propto \prod_{n=1}^{N}\lambda_n^{-\frac{1}{2}}\exp\left(-\frac{(\lambda_n + C_0(1-y_n\boldsymbol{\eta}^{\mathrm{T}}\tilde{\mathbf{x}}_n))^2}{2\lambda_n}\right)\mathbf{N}(\boldsymbol{\eta}; \mathbf{0}, \sigma^{-1}\mathbf{I})\mathrm{Ga}(\sigma; a_0, b_0) \end{aligned} \tag{4}$$

Consequently, Gibbs sampling algorithm [14,15] can be implemented to repeatedly sample each random variable from its conditional distribution. The augmented data space allows the SVM optimality criterion to be expressed as a conditionally Gaussian linear model that is familiar to most Bayesian statisticians.

2.2. Max-margin factor analysis

Factor analysis, one of the most popular probabilistic latent models, projects an observation into a low dimensional space that captures the latent feature of data, which, specifically, assumes that an observed P-dimensional variable $\mathbf{x}$ is generated as a linear transformation of some lower K-dimensional latent variable $\mathbf{s}$ plus additive Gaussian noise $\boldsymbol{\varepsilon}$. The transformation matrix $\mathbf{D}$ is the loading matrix with each column $\mathbf{d}_k, k = 1,\cdots,K$. Then the