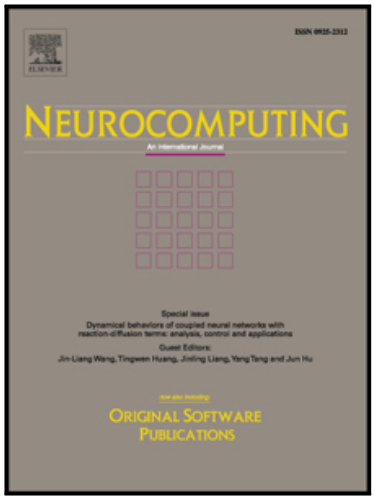
Communicated by Dr Qinfeng Shi

Accepted Manuscript



Supervised Dirichlet Process Mixtures of Principal Component Analysis

Jiangtao Ren, Kang Li, Chaotao Chen

PII: S0925-2312(18)30483-1
DOI: 10.1016/j.neucom.2018.04.047
Reference: NEUCOM 19516

To appear in: *Neurocomputing*

Received date: 1 March 2017
Revised date: 5 December 2017
Accepted date: 1 April 2018

Please cite this article as: Jiangtao Ren, Kang Li, Chaotao Chen, Supervised Dirichlet Process Mixtures of Principal Component Analysis, *Neurocomputing* (2018), doi: 10.1016/j.neucom.2018.04.047

Supervised Dirichlet Process Mixtures of Principal Component Analysis

Jiangtao Ren, Kang Li, Chaotao Chen

School of Data and Computer Science, Sun Yat-sen University, Guangzhou, Guangdong, P.R.China 510006

Abstract

We introduce Probabilistic Principal Component Analysis (PPCA) into Dirichlet Process Mixtures of Generalized Linear Models (DPGLM) and propose a new model called Supervised Dirichlet Process Mixtures of Principal Component Analysis (SDPM-PCA). In SDPM-PCA, we assume covariates and response variable are generated separately through the latent variable of PPCA, and nonparametrically modeled using the Dirichlet process mixture. By jointly learning the latent variable, cluster label and response variable, SDPM-PCA performs locally dimensionality reduction within each mixture component, and learns a supervised model based on the latent variable. In this way, SDPM-PCA improves the performance of both dimensionality reduction and prediction on high-dimensional data with all advantages of DPGLM. We also develop an inference algorithm for SDPM-PCA based on variational inference, which provides faster training speed and deterministic approximation compared with sampling algorithms based on MCMC method. Finally, we instantiate SDPM-PCA in regression problem with a Bayesian linear regression model. We test it on several real-world datasets and compare the prediction performance with DPGLM and other regular regression model. Experiment results show that by setting properly latent dimension number, SDPM-PCA would provide better prediction performance on high-dimensional regression problem and avoid the curse

Email addresses: issrjt@mail.sysu.edu.cn (Jiangtao Ren), likang7@mail2.sysu.edu.cn (Kang Li), chencht3@gmail.com (Chaotao Chen)