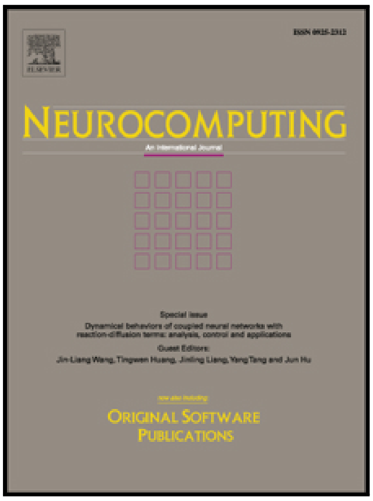
ISSN 0925-2312
ELSEVIER
NEUROCOMPUTING
An International Journal
Special issue
Dynamical behaviors of coupled neural networks with reaction-diffusion terms: analysis, control and applications
Guest Editors:
Jin-Liang Wang, Tingwen Huang, Jinling Liang, Yang Tang and Jun Hu
now also including
ORIGINAL SOFTWARE PUBLICATIONS

Streaming Clustering with Bayesian Nonparametric Models

Viet Huynh, Dinh Phung

PII: S0925-2312(17)30425-3
DOI: 10.1016/j.neucom.2017.02.078
Reference: NEUCOM 18175

To appear in: *Neurocomputing*

Received date: 8 June 2016
Revised date: 31 January 2017
Accepted date: 14 February 2017

Please cite this article as: Viet Huynh, Dinh Phung, Streaming Clustering with Bayesian Nonparametric Models, *Neurocomputing* (2017), doi: 10.1016/j.neucom.2017.02.078

Streaming Clustering with Bayesian Nonparametric Models

Viet Huynh[a,*], Dinh Phung[a],

[a] *Centre for Pattern Recognition and Data Analytics, Deakin University, Waurn Ponds, Victoria, 3216, Australia*

Abstract

Bayesian nonparametric (BNP) models are theoretically suitable for learning streaming data due to their complexity relaxation to growing data observed over time. There is a rich body of literature on developing efficient approximate methods for posterior inferences in BNP models, typically dominated by MCMC. However, very limited work has addressed posterior inference in a streaming fashion, which is important to fully realize the potential of BNP models applied to real-world tasks. The main challenge resides in developing one-pass posterior update which is consistent with the data streamed over time (i.e., data is scanned only once), for which general MCMC methods will fail to address. On the other hand, Dirichlet process-based mixture models are the most fundamental building blocks in the field of BNP. To this end, we develop in this paper a class of variational methods suitable for posterior inference of the Dirichlet process mixture (DPM) models where both the posterior update and data are presented in a streaming setting. We first propose new methods to advance existing variational based inference approaches for BNP to allow the variational distributions growing over time, hence overcoming an important limitation of current methods in imposing parametric, truncated restrictions on the variational distributions. This results in two new methods namely *truncation-free variational Bayes* (TFVB) and *truncation-free maximization expectation* (TFME) respectively where the latter further supports hard clustering. These inference methods form the foundation for our streaming inference algorithm where we further adapt the recent Streaming Variational Bayes proposed in [1] to our task. To demonstrate our framework for real-world tasks whose datasets are often heterogeneous, we develop one more theoretical extension for our model to handle assorted data where each observation consists of different data types. Our experiments with automatically learning the number of clusters demonstrate the comparable inference capability of our framework in comparison with truncated version variational inference algorithms for both synthetic and real-world datasets. Moreover, an evaluation of streaming learning algorithms with text corpora reveals both quantitative and qualitative efficacy of the algorithms on clustering documents.

Keywords: streaming learning, Bayesian nonparametric, variational Bayes inference, Dirichlet process, Dirichlet process mixtures, heterogeneous data sources

1. Introduction

We are at the dawn of a new revolution in the Information Age: *data*. "Every animate and inanimate object on Earth will soon be generating data" [2]. While we collectively are tweeting 8,000 messages around the world every second, our homes, cars, cities and even our bodies are also constantly generating terabytes of signals. A common setting is that these data are collected sequentially in time and our modern machine learning tools need algorithms to learn from data stream without the need of revisiting past data. More importantly, not only are data getting bigger in size, but also they are growing complexity, structure, and geometry. Hence, dealing with streaming data requires flexible models that can expand with data size and complexity. Bayesian nonparametric (BNP) models naturally fit this purpose since their complexity, e.g., the number of mixture components, can grow as new data appear. One challenge, however, for Bayesian models in general and Bayesian nonparametric models in particular is that it lacks efficient inference methods to deal with large scale and streaming data.

Two main inference approaches for BNP models are simulation methods such as Markov Chain Mote Carlo (MCMC) and deterministic variational methods. To deal with streaming data, sequential and particle MCMC methods were developed. However, MCMC algorithms are often unable to cope with large-scale data sets due to their slow convergence and unpredictable convergence diagnosis [3]. On the other pillar, deterministic variational inference methods are preferred in large-scale settings. The underlying idea of variational inference is to cast the posterior distribution of the model to an optimization problem by introducing an approximate (and tractable) variational distribution. The optimization problem is obtained by formulating the distance between the variational distribution and the posterior which usually is Kullback–Leibler

[*] Corresponding author
Email address: hvhuynh@deakin.edu.au (Viet Huynh)