International Journal of Approximate Reasoning 93 (2018) 153–177

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Fast approximation of variational Bayes Dirichlet process mixture using the maximization–maximization algorithm

Kart-Leong Lim *, Han Wang

Nanyang Technological University, 50 Nanyang Ave, 639798, Singapore

ARTICLE INFO

Article history:
Received 27 May 2017
Received in revised form 1 November 2017
Accepted 1 November 2017
Available online 7 November 2017

Keywords:
Fast Bayesian nonparametrics
Dirichlet process mixture
Gaussian mixture model
Variational Bayes
Variational maximization–maximization algorithm

ABSTRACT

In Bayesian nonparametrics model such as Dirichlet process mixture (DPM), learning is almost exclusive to either variational inference or Gibbs sampling. Yet variational inference is seldom mainstream in fast algorithms for DPM mainly due to high computational cost. Instead, most fast algorithms are largely based on MAP estimation of Gibbs sampling probabilities. However, they usually face intractable posterior and typically degenerate the conditional likelihood to overcome the inefficiency. Scalable variational inference such as stochastic variational inference exist but these works rely on the same two-step learning approach that involves hyperparameters and expectations update. This constitutes to the high cost often associated with variational inference. Inspired by fast DPM algorithms, we propose using MAP estimation of variational posteriors for approximating expectations. As a result, learning can be completed in a single step. However, we encounter undefined variational posteriors of log expectation. We overcome this problem by the use of lower bounds. When our cluster assignment also uses a MAP estimation, we have a global objective known as the maximization–maximization algorithm. We revisit the concepts of variational inference and observe that some of the analytical solutions obtained by our proposed method are very similar to variational inference. Lastly, we compare our fast approach to variational inference and fast DPM algorithms on some UCI and real datasets. Experimental results showed that our proposed method obtained comparable clustering accuracy and model selection but significantly faster convergence than variational inference.

1. Introduction

In clustering, we require beforehand the number of clusters K to use. Bayesian Nonparametric (BNP) does not have this issue as it can learn K directly from the dataset. However, one limiting factor of BNP is the expensive cost to compute. There are existing works that addressed this issue which we name as fast algorithms of BNP. These works are dedicated to achieving fast computation of BNP and in most cases are willing to trade accuracy for efficiency. Also, for mathematical convenience they mostly restricted to conjugate distributions from the exponential family, notably the DPM.

Several works dedicated to fast algorithms of Dirichlet process mixture (DPM) have been proposed in the past. A common trait in these methods is that MAP estimate is preferred over the slower Gibbs sampling for cluster assignment. An early work is DPsearch [1] whereby the authors use MAP estimate on the true posterior. Due to intractable MAP solution, a tighter

* Corresponding author.
E-mail address: lkartl@yahoo.com.sg (K.-L. Lim).

https://doi.org/10.1016/j.ijar.2017.11.001
0888-613X/

bound is used on the conditional likelihood for efficiency. Similarly, SUGS [2,3] use MAP estimation for cluster assignment by sequentially allocating new samples to clusters that locally maximizes the posterior of DPM. However, they replaced the DPM posterior with the posterior of partition model to avoid intractable solution. Inspired by K-means, the authors in DP-means [4] applied small variance asymptotics (SVA) to the prior model of DPM. MAP estimate on the probabilities of Gibbs sampling for cluster assignment then lead to an objective function that resembles K-means. In [5], a direct posterior approach is used but because MAP estimate on the cluster assignment posterior is intractable, by assuming SVA the authors similarly arrive at an identical objective function to DP-means. While SVA greatly reduce the complexity in [5,4], it also rob away the "rich get richer" property in DPM. MAP-DPM [6] overcame the SVA reliance in [5] by observing that the conjugacy leads to the Student-T distribution for the cluster assignment posterior. Cluster assignment probabilities for Gibbs sampling is then computed using a simple MAP estimate on the Student-T distribution. Conversely, we can improve the speed of DPM by making variational inference based DPM more efficient. Stochastic variational inference [7,8] looked at improving the scalability of variational inference by defining a new set of rules that allow local learning from sampled batches of the full dataset. In memorized online variational inference for DPM [9], the authors mainly aimed at improving the shortcomings of the stochastic variational inference such as requiring careful choices of batch size and learning rate. More crucially, the learning approach to variational posteriors [10–16] still remained largely unchanged from past works [17,18].

There are two key observations on fast DPM algorithms:

i) Most works do not consider variational inference as mainstream due to the difficulty in implementation and high computation cost. Their works are largely based on MAP estimate of the true posterior for cluster assignment. However, they face intractable posterior and in return they usually degenerate the conditional likelihood (e.g. SVA) to overcome the inefficiency.

ii) Most works using variational inference rely on a two-step learning approach (i.e. first compute hyperparameter updates then compute expectation of variational posteriors e.g. Algorithm 1). Also, expressions for the expectation of the variational posteriors cannot be obtained independently. Instead, they must rely on the mathematical convenience of conjugate prior where the expectation of the variational posteriors is known to have a similar expression as the statistical moments of their conjugate prior counterparts. By restricting to prior models that already have predefined closed-form expressions for their statistical moments, only then we can obtain expressions for the expectation of the variational posteriors. Moreover, computing the expectation for the cluster assignment variational posterior is more expensive than a MAP estimate. This two-step approach mainly constitutes to the inflexibility and high cost involved.

Inspired by the challenges of fast DPM algorithms, this paper contributes in the followings:

i) We no longer need to refer to any predefined closed-form expressions from the statistical moments of the conjugate prior. Instead, we directly approximate expectations of variational posteriors by using MAP estimation. We also use MAP estimation on the posterior of cluster assignment but in a variational framework. Thus, variational learning is now completed in a single step in Algorithm 2.
ii) We show that the use of lower bounds for computing expectation functions greatly reduce the analytical complexity involved in our approach.
iii) We revisit the concepts of variational inference and observe that some analytical solution of expectations obtained by our proposed method have very similar expressions to variational inference. Convergence of our method is also briefly discussed.
iv) Lastly, we compare our work mainly with DP-means and variational inference on synthetic dataset, UCI datasets and real datasets. Empirical results showed that our proposed method obtained comparable clustering accuracy and model selection but significantly faster convergence than variational inference.

2. Related work

In DPM-EM [19,20], Heinzl et al. used the EM algorithm for the inference of DPM with linear mixture model. Our work mainly differs from DPM-EM in several ways. Firstly, they only considered a maximum likelihood approach to the mixture model whereas we used a Bayesian approach. Secondly, their proposed solution is based on the EM algorithm, whereas we used the MM algorithm within a variational inference framework. Lastly, while linear mixture model is more expressive than GMM is used to compute longitudinal data, it comes at an extra computational cost. Thus, linear mixture model may be less desirable when it comes to application that require fast DPM computation. One key similarity we share is that both our expression for the cluster weight variable, v have identical closed-form expression despite both method employing different mixture models. This is possible since the cluster assignment path is disjointed from the mixture model path in the graph model of DPM in Fig. 1. Thus, inference of v is essentially the same problem for both DPMs despite using different mixture models and approach.

Recently, in [21,22] variational Maximization–Maximization was proposed to successfully perform learning for the Bayesian inference of GMM [22] and the Sparse Coding based GMM [21]. The main difference between this work and MM-GMM [22] is that the theoretical comparison between variational Maximization–Maximization (MM) and variational Expectational–Expectation (EE) is now extended to stochastic process. Due to having a more complex model, there is greater difficulty computing expectations of variational posteriors here.