Bayesian Model-based Clustering

Zhuoqun Wang

October 2, 2019

Cover Note

- What is your main topic or purpose in this paper?

  Literature review on Bayesian model-based clustering, mainly about the three key factors of mixture models.

- Who is your audience?

  Field-specific audience, probabily statisticians.

- What organizational format did you use in this paper?

  Introduction: introduce basic mixture model for clustering; Main body: discuss the three key factors respectively; Conclusion; References; Description of research methods.

- What questions do you still have?

  No questions at this point. I'm very sorry that I haven't finished this yet because I talked to the professor working in this field on Tuesday and got many new things to talk about in this literature review.

Introduction

Mixture models are widely used for model-based clustering of complex data. The basic model assumes that each observation is independently drawn from an unknown distribution corresponding to the cluster assignment. From a Bayesian perspective, the generative model is as follows:

$y_i \sim K(\theta_{C_i})$

$Pr(C_i = c) = \pi_c$

$\pi_c \sim P, c = 1, 2, ...$

$\theta_c \sim P_0$, iid

where random variable C_i is cluster assignment, π_c characterizes the prior on C_i, P_0 is the base measure, K is the kernel function. We can write the model as

$$f(y) = \Sigma_{c=1}^{\infty} \pi_c K(y; \theta_c),$$

so y_i is generated from a mixture of distributions parametrized by θ , with the mixing distribution of θ being P.

There are three factors that will have impacts the posterior on clustering: the prior on clustering P, the base measure P_0, and the kernel function K.

Choice of Prior on Clustering

Peter Orbanz pointed out that there are three different types of problems with respect to number of components. Note that under this model, if we set P such that only finite number K of π_c's have positive measure, then the number of mixture components is finite, hence the number of mixture components can be either finite or infinite.

In the finite case, we need to distinguish between "K is finite and known" and "K is finite but unknown". If K is known, we could define P as a probability measure that is only positive on K singletons. One commonly used mixture model with finite known K in microbiome study and natural language processsing is the Dirichlet-multinomial Mixture (DMM) model, which takes advantage of Dirichlet-multinomial conjugacy. If K is unknown, the natural Bayesian approach is to put a prior on the number of components K.

In the infinite case, P is a probability measure that assigns positive probability to countably infinite points in the parameter space Π. The likelihood function can be rewrite as integration with respect to a discrete measure parameterized by π and θ. If we take the Bayesian approach and put priors on the space of these discrete measures, then the dimension of the space is infinite, which makes this problem nonparametric. Commonly used priors include Dirichlet Process, which leads to the Dirichlet Process Mixture (DPM) model. As a special case of models based on stick-breaking process, the DPM model enjoys many theoretical properties (Neal 2000, Ferguson 1973). Sampling algorithms for posterior inference of DPM model are well established, based on Blackwell–MacQueen urn scheme and Gibbs sampling (Blackwell and MacQueen 1973, Neal 2000).

Paralell properties are well established for finite mixture models and DPM, so it has been noted that methods to apply the Gibbs samplers for DPM on the finite mixture models are possible (Green and Richardson 2001). Recently, posterior inference algorithms for finite mixture models are developed paralelling those of DPM, based on the asymptotic results of finite mixture models. (Miller and Harrison 2018).

Choice of Kernel Function

To be added.

Choice of Base Measure

To be added.

Conclusion

To be added.

References

Blackwell, D., & MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *The annals of statistics*, *1*(2), 353-355.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The annals of statistics*, 209-230.

Green, P. J., & Richardson, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scandinavian journal of statistics*, *28*(2), 355-375.

Miller, J. W., & Harrison, M. T. (2018). Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, *113*(521), 340-356.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, *9*(2), 249-265.

Description of Research Process

I'm not currently working on any projects so I only have an idea of the general domain I'm going to write about. Then I talked with a professor working in the domain of Bayesian nonparametrics and he helped me narrowed my domain to a specific topic, and recommended some papers cited (or will be cited) in this draft. The papers cited here are selected because they are closely related to the topic “Bayesian model based clustering”. Some of them are about the state-of-art studies on the three factors discussed in this literature review, the others are about the background or foundation of my topic. Since there are three key factors in this topic, the organization is naturally to discuss them respectively.