Optimal Bayesian clustering using non-negative matrix factorization

Ketong Wang, Michael D. Porter

PII: S0167-9473(18)30185-3
DOI: https://doi.org/10.1016/j.csda.2018.08.002
Reference: COMSTA 6663

To appear in: *Computational Statistics and Data Analysis*

Received date : 8 March 2017
Revised date : 3 August 2018
Accepted date : 4 August 2018

Optimal Bayesian Clustering using Non-negative Matrix Factorization

Ketong Wang[a], Michael D. Porter[a,*]

[a] *Department of Information Systems, Statistics, and Management Science*
The University of Alabama, Tuscaloosa, AL 35401, United States

Abstract

Bayesian model-based clustering is a widely applied procedure for discovering groups of related observations in a dataset. These approaches use Bayesian mixture models, estimated with MCMC, which provide posterior samples of the model parameters and clustering partition. While inference on model parameters is well established, inference on the clustering partition is less developed. A new method is developed for estimating the optimal partition from the pairwise posterior similarity matrix generated by a Bayesian cluster model. This approach uses non-negative matrix factorization (NMF) to provide a low-rank approximation to the similarity matrix. The factorization permits hard or soft partitions and is shown to perform better than several popular alternatives under a variety of penalty functions.

Keywords: Bayesian clustering, Non-negative matrix factorization (NMF), soft clustering, cluster analysis, fuzzy clustering

1. Introduction

The goal of clustering is to discover partitions that assign observations into meaningful groups. A favorable property of Bayesian model-based clustering is that it provides versatile posterior uncertainty assessment on both the model parameters and cluster allocation estimates. However, while inference on model-specific parameters and mixing weights follow standard Bayesian practice, more development on estimating the clustering partition is needed.

An intuitive, yet naive, way to obtain a point estimate of the best partition is to use a maximum a posteriori (MAP) approach which selects the partition (up to label switching) from the MCMC sample that occurs the most frequently. But when the number of observations is large and the generating mixture is complex, the majority of the MCMC clustering samples are likely to be visited, at most, a few times. In this case, the MAP approach will not usually find the best clustering solution.

To better clarify the notion of optimal partitioning, Binder (1978) introduced a loss function approach. This considers optimal clustering as a *Bayesian action* which attempts to minimize the expected loss of

[*] Corresponding author. E-mail address: mdp2u@virginia.edu