Application of sequential Monte Carlo methods for the clustering of multiple datasets

Nathan Cunningham

September 1, 2016

Abstract

0.1 Introduction

Bit on cluster analysis, bit on multiple datasets, applications of this to e.g. gene expression data.

A little bit on sequential monte carlo methods.

0.2 Literature

In their paper [2], Kirk et al. propose an unsupervised method for the integration of multiple datasets. Their work is applicable to a number of data types: gaussian, gaussian processes, time series data, multinomial data. MDI paper summary [2]

SMC paper summary [1]

Maybe bit on Sarah Wade's paper? [3]

0.3 Methods

The algorithm presented here is a combination of the work by Griffin and Kirk et al.. The aim of the algorithm is to allocate observations from a number of datasets to clusters, and discover a shared clustering across the data sets.

It is assumed that for each of K datasets we make Q_k observations on each of n genes. Each dataset is assumed to share a common clustering structure. The prior probability of assigning an observation to a cluster c_{ij} is denoted π_j. The parameter ϕ_{kl} models the dependence between the component allocations of observations in dataset k and l.

$x_{i,k}$ denotes the i^{th} observaiton in dataset k. c_{ik} denotes the component allocation variable for the i^{th} observation in dataset k. The prior pribability of the component allocation variables $[c_{1k}, \ldots, c_{nk}]$ are given in the vector $\pi_{\mathbb{k}}$ which in turn are given a $Dirichlet(\alpha_k)$ prior.

The algorithm assumes the data arises as a structure as presented in Figure 1

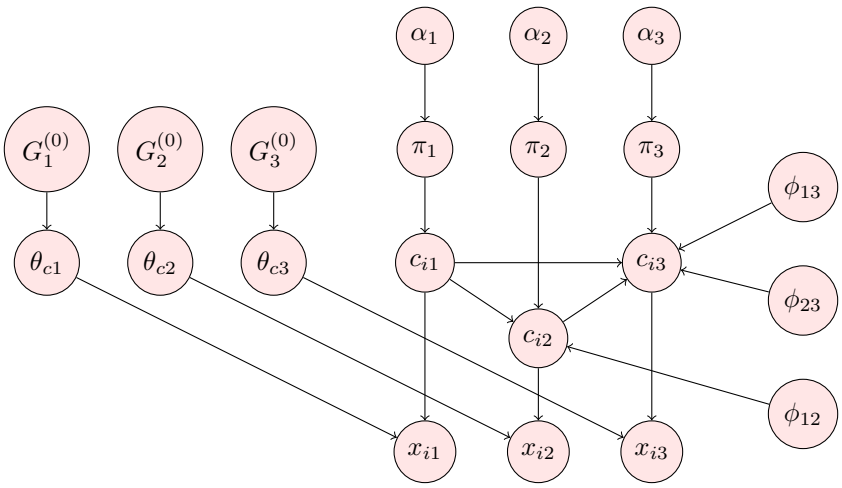

Figure 1: Graphical representation of the model

Algorithm 1 Gibbs sampler

1: Initialise Γ matrix of prior allocation weights and Φ matrix of dataset concordance values
2: **for** i = 1, ..., number of iterations **do**
3: Conditional on Γ_{i-1} and Φ_{i-1} update the cluster labels, c_i, using alg. 2
4: Conditional on c_i update Γ_i and Φ_i
5: **end for**

Algorithm 2 Particle filter to update cluster allocations

for i = 1, ..., n **do** ▷ Loop over observations
  for m = 1, ..., M **do** ▷ Loop over particles
    for j = 1, ..., d **do** ▷ Loop over datasets
      Sample $c_{i,j}^{(m)}$ ▷ Propose a cluster for each datum
      $q(c_{i,j}^{(m)} = k) \propto k^*(y_{i,j}|c_{i,j}^{(m)} = k) \times \gamma_{i,k,j}$
      $\xi^{(m)} = \xi^{(m)} \times \gamma_{i,k,j} \times (1 + \phi_i) \times k^*(y_{i,k}|c_{i,j}^{(m)} = k)$
    end for
  end for
  Resample particles according to $\xi^{(m)}$
end for
Update cluster labels using allocation in particle with largest $\xi^{(m)}$

Where

$$k^*(y_{i,k}|c_{i,k}^{(m)} = k) = (\mathbf{y_{i,k}} - \mu_\mathbf{k})\mathbf{\Sigma^{-1}}(\mathbf{y_{i,k}} - \mu_\mathbf{k})^\top \tag{1}$$

$$\Phi \text{ is a measure of cluster label correspondence across datasets} \tag{2}$$

$$\gamma_{i,k,j} \text{ is a prior weight for assigning observation i, in dataset k to cluster j} \tag{3}$$

The sequential nature of particle filter methods mean that clustering allocations are only based on all prior observations. This would suggest that the quality of clustering improves for later observations as they are based on more data. In typical applications, the observations would be time-indexed and, as such, the order of observations is important. This is not the case here, however, and between runs of the particle filter the observations are shuffled.

0.4 Example application

Comparison of this versus independent clustering of the datasets

Use on multinomial data and gaussian data

0.5 Conclusions and proposals for future work

Some success. Improvements over independent clustering...?

Future work needed: Updating of concordance values by particle? Outputting of more than one particle?

Parallelisation of the code to speed things up.

Feature selection in cluster analysis

Application to real-life data (genomics England)

Bibliography

[1] JE Griffin. Sequential monte carlo methods for mixtures with normalized random measures with independent increments priors. *Statistics and Computing*, pages 1–15, 2014.

[2] Paul Kirk, Jim E Griffin, Richard S Savage, Zoubin Ghahramani, and David L Wild. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, 28(24):3290–3297, 2012.

[3] Sara Wade and Zoubin Ghahramani. Bayesian cluster analysis: Point estimation and credible balls. *arXiv preprint arXiv:1505.03339*, 2015.