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

Nathan Cunningham

September 1, 2016

Abstract

In this paper I

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.

Algorithm 1 Gibbs sampler

Initialise Γ matrix of prior allocation weights and Φ matrix of dataset concordance values
for i = 1, ..., number of iterations **do**
  Conditional on Γ_{i-1} and Φ_{i-1} update the cluster labels, c_i, using alg. 2
  Conditional on c_i update Γ_i and Φ_i
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
      $q(c_{i,j}^{(m)} = k) \propto k^*(y_{i,j}|c_{i,j}^{(m)} = k)\gamma_{i,k,j}$
      $\xi^{(m)} = \xi^{(m)} \times \gamma_{i,k,j}(1 + \phi_i)k^*(y_{i,k}|c_{i,j}^{(m)} = k)$
    end for
  end for
  Resample particles according to $\xi^{(m)}$
end for
Update cluster labels using allocation in particle with largest $\xi^{(m)}$

Where

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

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

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

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.