Bayesian consensus clustering

Stephen Coleman

22/10/2019

Notation

N	The number of samples in the data.
L	The number of datasets present.
p_l	The number of measurements for each sample in the lth dataset.
K_l	The number of components present in the lth dataset.
$x = (x_1, \ldots, x_l)$	The datasets.
$x_l = (x_{l1}, \ldots, x_{lN})$	The observed data for the lth dataset.
$c = (c_1, \ldots, c_L)$	The membership vectors for each dataset (our latent variable).
$c_l = (c_{l1}, \ldots, c_{lN})$	The context-specific component membership.
$C = (C_1, \ldots, C_N)$	The global allocation vector.
$\pi_l = (\pi_{l1,} \ldots, \pi_{lK_l})$	The mixture weights in the lth context.
$\Pi = (\Pi_1, \ldots, \Pi_K)$	The mixutre weights for the global components.

If $K_1 = \ldots = K_L$ then we use K as the number of components in each context. We treat p_l in the same way. I denote abbreviations or terms that will be used in place of another in the format "[Full name] ([name hereafter])".

Dirichlet-Multinomial Allocation mixture model

Bayesian Consensus Clustering (BCC) (Lock and Dunson 2013) is based upon a finite approximation of the Dirichlet process known as the Dirichlet-Multinomial Allocation (DMA) mixture model (Green and Richardson 2001), as is Clusternomics (Gabasova, Reid, and Wernisch 2017) and Multiple Dataset Integration (Kirk et al. 2012).

Consider a dataset of N samples, $x = (x_1, \ldots, x_N)$ where each sample is itself a p-vector of measurements for some $p \in \mathbb{N}$. We are interested in uncovering structure in the data. To do this we associate each sample, x_i, with a *cluster*.This moves us from a p-dimensional space to a 1-dimensional space based upon similarity of samples which enables interpretation. We want to understand if there are sub-populations in our sample that are responsible for heterogeneity. By partitioning the data into clusters we hope to have some insight into this underlying structure and improve our understanding of the data. There are a myriad of ways to uncover such partitions. Some methods have more obvious disadvantages than others, a common problem being that the number of clusters allowed in the result is arbitrary. Another common problem is that some of the methods are not model based which gives an ad-hoc nature to the paritioning. A method that does not suffer from these problems is the Dirichlet process. A tractable approximation of this is the DMA mixture model which is a common choice in the Bayesian-model based clustering. In this case we model each sub-population by an individual distribution. We also allow the number of clusters present to be inferred from the data, thus avoiding a heuristic or arbitrary means of selecting the number of clusters allowed.

In this model we use a mixture of K components to model the sub-populations. The components are all modelled by a distribution with a density function $f(\cdot)$. The kth component has associated parameters, θ_k, based on the samples allocated to this component and is thus defined by density $f(\theta_k)$. The proportion of samples within the components define the associated mixture weights, $\pi = (\pi_1, \ldots, \pi_K)$. Each sample

is allocated to a specific component; this component memberhsip is represented by the latent variable $c = (c_1, \ldots, c_n)$, where $c_i \in \{1, \ldots, K\} \forall i \in \{1, \ldots, N\}$. Therefore we have allocation probability:

$$p(x_i|c_i = k) = \pi_k f(x_i \theta_k),$$

and the full model density, which is a weighted sum of the mixture densities:

$$p(x_i) = \sum_{k=1}^{K} \pi_k f(x_1|\theta_k)$$

Note that while there are component-specific parameters, the density function used, $f(\cdot)$, is common across all components.

In the DMA mixture model, the mixture weights are given a Dirichlet prior, normally with a symmetric concentration parameter. The allocation variable is then sampled from a Cateogrical distrbution defined by the component weights. The component parameters are then updated based upon the allocation of samples and the full model is then the weighted sum described above. The full hierarhcical model is described below:

$$\begin{aligned} \pi &\sim \text{Dirichlet}\left(\frac{\alpha_0}{K}, \ldots, \frac{\alpha_0}{K}\right), \\ c_i &\sim \text{Categorical}(\pi), \\ \theta_k &\sim h(\cdot), \\ x_i|c_i = k &\sim f(x_i|\theta_k). \end{aligned}$$

Here $h(\cdot)$ is some distribution, often with additional hyperparameters. If one let's $K \to \infty$ then this model becomes a Dirichlet process. To approximate this K is set to an arbitrarily large number. In this scenario one must have chosen a K sufficiently large that there are empty clusters for the model to be a true approximation of the Dirichlet process. For example, say we set $K = 30$ in our initialisation, and that in the output 30 clusters are occupied (i.e. in our c vector we have 30 unique labels). One must use a higher K then this; if we try again with $K = 50$ and c contains 40 unique values than the number of clusters is learnt from the data and not an arbitrary choice of the user.

Integrative model

BCC (Lock and Dunson 2013) is extension of the DMA mixture model. Consider the case of L datasets describing the same N samples with different measurements or under different experimental conditions (e.g. gene expression data in one dataset, copy number variation of the same genes in another). In this case there are context-specific models similar to that described above with a common number of components, K, in each model. There is then an additional *global clustering*, $C = (C_1, \ldots, C_N)$, where $C_i \in \{1, \ldots, K\} \forall i \in [1, N]$. In this model it is the local allocations that are dependent upon the global membership. For the lth context, given the global clustering C, the context-specific concentration parameter, α_l:

$$P(c_{li} = k|C_i) = \nu(k, C_i, \alpha_l)$$

where α_l changes the magnitude of dependence of the local clustering about the global clustering. The conditional model is:

$$P(c_{li} = k|x_{li}, C_i, \theta_{lk}) \propto \nu(k, C_i, \alpha_l) f_L(x_{li}|\theta_{lk}).$$

Lock and Dunson (2013) assume the following form for $\nu(\cdot)$:

$$\nu(c_{li}, C_i, \alpha_l) = \alpha_l \mathbb{I}(c_{li} = C_i) + \frac{1 - \alpha_l}{K - 1}(1 - \mathbb{I}(c_{li} = C_i))$$

where $\alpha_l \in [\frac{1}{K}, 1]$ controls the similarity of the clustering in context l to the global clustering and $\mathbb{I}(\cdot)$ is the indicator function. If $\alpha_l = 1$ then $c_l = C$. The α_l is a random variable inferred from the data. Let $\alpha = (\alpha_1, \ldots, \alpha_L)$.

Let $\Pi = (\Pi_1, \ldots, \Pi_K)$ be the component weights at the global level. Then the conditional probability of the context-specific clustering given the component weights is defined to be:

$$P(c_{li} = k|\Pi) = \alpha_l \Pi_k + (1 - \Pi_k)\frac{1 - \alpha_l}{K - 1}$$

Note that if *kth* global mixture weight is 0 (and hence the associated component has a total membership of 0) then the probability of assignment to the *kth* local cluster is 0 if and only if $\alpha_l = 1$. This is unlikely to happen (as 1 is the upper bound on α_l), and therefore it is normal that there are *more* local clusters than global clusters in stark contrast to the Clusternomics model.

The probability of being allocated to the *kth* global component is:

$$P(C_i = k|c, \Pi, \alpha) \propto \Pi_k \prod_{l=1}^{L} \nu(c_{li}, k, \alpha_l).$$

References

Gabasova, Evelina, John Reid, and Lorenz Wernisch. 2017. "Clusternomics: Integrative Context-Dependent Clustering for Heterogeneous Datasets." *PLoS Computational Biology* 13 (10). Public Library of Science: e1005781.

Green, Peter J, and Sylvia Richardson. 2001. "Modelling Heterogeneity with and Without the Dirichlet Process." *Scandinavian Journal of Statistics* 28 (2). Wiley Online Library: 355–75.

Kirk, Paul, Jim E Griffin, Richard S Savage, Zoubin Ghahramani, and David L Wild. 2012. "Bayesian Correlated Clustering to Integrate Multiple Datasets." *Bioinformatics* 28 (24). Oxford University Press: 3290–7.

Lock, Eric F, and David B Dunson. 2013. "Bayesian Consensus Clustering." *Bioinformatics* 29 (20). Oxford University Press: 2610–6.