Dirichlet processes, posterior similarity and graph clustering

Processo di Dirichlet, similarità a posteriori e classificazione su grafi

Stefano Tonellato

Abstract This paper proposes a clustering method based on the sequential estimation of the random partition induced by the Dirichlet process. Our approach relies on the Sequential Importance Resampling (SIR) algorithm and on the estimation of the posterior probabilities that each pair of observations are generated by the same mixture component. Such estimates do not require the identification of mixture components, and therefore are not affected by label switching. Then, a similarity matrix can be easily built, allowing for the construction of a weighted undirected graph. A random walk can be defined on such a graph, whose dynamics is closely linked to the posterior similarity. A community detection algorithm, the map equation, can then be implemented in order to achieve a clustering by minimising an information theoretic criterion.
Abstract *Si propone un metodo di classificazione basato sulla stima sequenziale della partizione indotta dal processo di Dirichlet. Il metodo si basa su un algoritmo stocastico sequenziale (SIR) e sulla probabilità a posteriori che ciascuna coppia di osservazioni sia generata da una componente della mistura. Il metodo non richiede l'identificazione delle singole componenti e non risente degli inconvenienti del label switching. Una matrice di similarità può quindi essere stimata a posteriori. Questo consente la costruzione di un grafo pesato e la definizione, su di esso, di una passeggiata aleatoria. Un algoritmo utilizzato nella classificazione di reti, chiamato map equation, che ricerca la partizione che minimizza un criterio di informazione, può quindi essere facilmente implementato.*

Key words: Dirichlet process, sampling importance resampling, community detection, map equation

Stefano Tonellato
Ca' Foscari University of Venice, e-mail: stone@unive.it

1 Dirichlet process mixtures and clustering

A very important class of models in Bayesian nonparametrics is based on the Dirichlet process and is known as Dirichlet process mixture [1]. In this model, the observable random variables, X_i, $i = 1, \dots, n$, are assumed to be exchangeable and generated by the following hierarchical model:

$$\begin{aligned} X_i|\theta_i &\overset{ind}{\sim} p(\cdot|\theta_i),\ \theta_i \in \Theta \\ \theta_i|G &\overset{iid}{\sim} G \\ G &\sim DP(\alpha, G_0), \end{aligned}$$

where $DP(\alpha, G_0)$ denotes a Dirichlet process (DP) with base measure G_0 and precision parameter $\alpha > 0$. Since the DP generates almost surely discrete random measures on the parameter space Θ, ties among the parameter values have positive probability, leading to a batch of clusters of the parameter vector $\theta = [\theta_1, \dots, \theta_n]^T$. Exploiting the Pólya urn representation of the DP, the model can be rewritten as

$$X_i|s_i, \theta^*_{s_i} \overset{iid}{\sim} p(\cdot|\theta^*_{s_i}),\ \theta^*_{s_i} \in \Theta \tag{1}$$

$$\theta^*_{s_i} \overset{iid}{\sim} G_0 \tag{2}$$

$$p(s_i = j|\mathbf{s}_{<i}) = \begin{cases} \frac{\alpha}{\alpha+i-1} & j = k \\ \frac{n_j}{\alpha+i-1} & j \in \{k-1\}, \end{cases} \tag{3}$$

$$s_i \perp \theta^*_j \qquad \forall i, j, \tag{4}$$

where $\{k\} = \{1, \dots, k\}$, $\mathbf{s}_{<i} = \{s_j,\ j \in \{i-1\}\}$ (in the rest of the paper, the subscript $< i$ will refer to those quantities that involve all the observations $X_{i'}$ such that $i' < i$), $s_j \in \{k\}$ for $j \in \{k-1\}$, and n_j is the number of θ_i's equal to θ^*_j. In this model representation, the parameter θ can be expressed as $(\mathbf{s}, \theta^*)$, with $\mathbf{s} = \{s_i : s_i \in \{k\},\ i \in \{n\}\}$, $\theta^* = [\theta^*_1, \dots, \theta^*_k]^T$ with $\theta^*_j \overset{iid}{\sim} G_0$, and $\theta_i = \theta^*_{s_i}$. Consequently, the marginal distribution of X_i is a mixture with k components, where k is an unknown random integer.

In the case of finite mixtures with k components, with k fixed and known, under a frequentist perspective it would be quite straightforward to cluster the data by maximising the probability of the allocation of each datum to one of the k components, conditionally on the observed sample [6]. Under a Bayesian perspective, the same results can be achieved, provided that either some identifiability constraints on the parameters are introduced, or a suitable risk function is minimised [12]. Unfortunately, under the assumptions we made, such computations are not feasible even numerically, due to the well known label switching problem [3] that persists when the number of mixture components is not known, nor finite, as in the case of Dirichlet process mixtures. Nevertheless, equations (1)–(4) are very helpful in estimating posterior pairwise similarities and building hierarchical clustering algorithms as in [7, 8]. In section 2, a sequential estimation algorithm analogous to the one in [5] is developed. In section 3, individuals are represented as nodes of a weighted undirected graph on which a random walk is built, with transition probabilities proportional to the posterior similarities. Nodes can then be classified by minimising the entropy through the map algorithm introduced in [10, 11]. The approach proposed in sections 2 and 3 has a double benefit. On one hand, the sequential estimation algorithm guarantees a fast estimation of pairwise similarities. On the other hand, the construction of the random walk on the graph mentioned above, allows us to choose the optimal partition by a minimum description length algorithm, so avoiding the subjective choice of a cut of the dendrogram usually associated to hierarchical clustering algorithms. Furthermore, as a byproduct, the entropy of any partition of the data can be computed and it is closely linked to the fitted model. This allows for a model based comparison of any pair of partitions.

2 Sampling importance resampling

Under the assumptions we introduced above, following the arguments of [5], we can write the conditional posterior distribution of s_i given $x_1,\dots,x_i$, as

$$p(s_i = j|\mathbf{s}_{<i},\theta^*,\mathbf{x}^{(j)}_{<i},x_i) = \begin{cases} \frac{n_j}{\alpha+i-1} p(x_i|\theta^*_j,\mathbf{s}_{<i},\mathbf{x}^{(j)}_{<i}) & j\in\{k\} \\ \frac{\alpha}{\alpha+i-1} p(x_i|\theta^*_{k+1}) & j=k+1, \end{cases}$$

where $\mathbf{x}^{(j)}_{<i} = \{x_{i'} : i'<i, s_{i'}=j\}$, $j=1,\dots,k$, and $\mathbf{x}^{(k+1)}_{<i}=\emptyset$, since $\forall i'<i$, $s_{i'}\in\{k\}$.

We can marginalise the conditional posterior of s_i with respect to θ^*, obtaining

$$p(s_i = j|\mathbf{s}_{<i},\mathbf{x}^{(j)}_{<i},x_i) = \begin{cases} \frac{n_j}{\alpha+i-1} p(x_i|s_i=j,\mathbf{s}_{<i},\mathbf{x}^{(j)}_{<i}) & j\in\{k\} \\ \frac{\alpha}{\alpha+i-1} p(x_i|s_i=k+1,\mathbf{s}_{<i},\mathbf{x}_{<i}) & j=k+1, \end{cases}$$

where

$$p(x_i|s_i=j,\mathbf{s}_{<i},\mathbf{x}_{<i}) = \\ \int_\Theta p(x_i|\theta,s_i=j,\mathbf{s}_{<i},\mathbf{x}^{(j)}_{<i})p(\theta|s_i=j,\mathbf{s}_{<i},\mathbf{x}^{(j)}_{<i})d\theta \tag{5}$$

and

$$p(x_i|s_i=k+1,\mathbf{s}_{<i},\mathbf{x}_{<i}) = \int_\Theta p(x_i|\theta)dG_0(\theta). \tag{6}$$

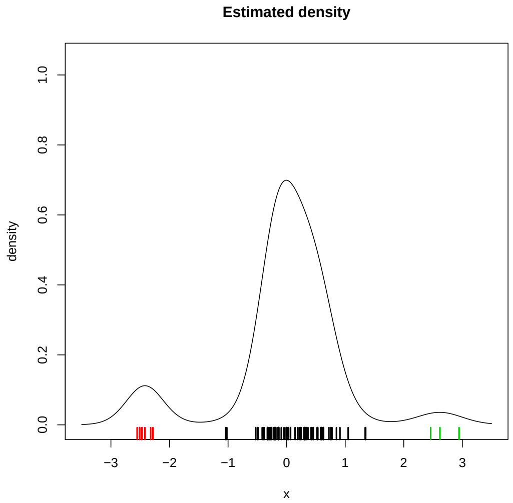
Notice that when G_0 is a conjugate prior for (1), the computation of (5) and (6) is often straightforward.

The following importance sampler has been introduced in [5].

SIR algorithm. For $i=1,\dots,n$, repeat steps (A) and (B)

(A) Compute

Fig. 1 Estimated posterior density function and clustering



$$g(x_i|\mathbf{s}_{<i},\mathbf{x}_{<i}) \propto \sum_{j=1}^{k+1} \frac{n_j}{\alpha+i-1} p(x_i|s_i=j,\mathbf{s}_{<i},\mathbf{x}_{<i}^{(j)}),$$

with $n_{k+1}=\alpha$.

(B) Generate s_i from the multinomial distribution with

$$p(s_i=j|\mathbf{s}_{<i},\mathbf{x}_{<i}^{(j)},x_i) \propto \frac{n_j}{\alpha+i-1} p(x_i|s_i=j,\mathbf{s}_{<i},\mathbf{x}_{<i}^{(j)}).$$

Taking R independent replicas of this algorithm we obtain $s_i^{(r)}$, $i=1,\dots,n$, $r=1,\dots,R$, and $\theta_j^* \sim p(\theta|\mathbf{x}^{(j)})$, with $\mathbf{x}^{(j)}=\{x_i : i\in\{n\}, s_i=j\}$, and compute the importance weights

$$w_r \propto \prod_{i=1}^{n} g(x_i|\mathbf{s}_{<i},\mathbf{x}_{<i})$$

such that $\sum_{r=1}^{R} w_r = 1$. Should the variance of the importance weights be too small, the efficiency of the sampler could be improved by resampling as follows [2]. Compute $N_{\text{eff}} = (\sum_{r=1}^{R} w_r^2)^{(-1)}$. If $N_{\text{eff}<\frac{R}{2}}$, draw R particles from the current particle set with probabilities equal to their weights, replace the old particle with the new ones and assign them constant weights $w_r=\frac{1}{R}$.

3 Pairwise similarities and community detection

Intuitively, we can state that two individuals, i and j, are similar if x_i and x_j are generated by the same mixture component, i.e. if $s_i=s_j$. Label switching prevents

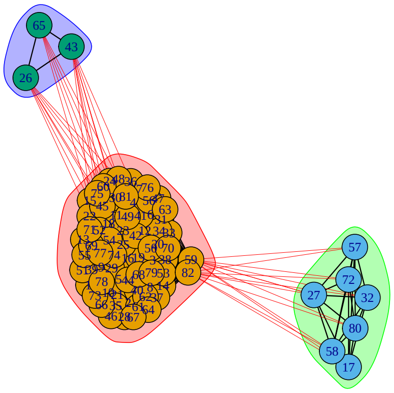


Fig. 2 The graph induced by the posterior similarity and the clusters detected by the map equation algorithm

us from identifying mixture components, but not from assessing similarities among individuals. In fact, the algorithm introduced in the previous section may help us in estimating pairwise similarities between individuals. The posterior probability that x_i and x_j are generated by the same component, i.e. the posterior probability of the event $\left\{s_i = s_j\right\}$, can be estimated as

$$\hat{p}_{ij} = \sum_{r=1}^{R} w_r I\left(s_i^{(r)}, s_j^{(r)}\right),$$

where $I(x,y) = 1$ if $x = y$ and $I(x,y) = 0$ otherwise. We can then define a similarity matrix S with ij-th element $s_{ij} = \hat{p}_{ij}$.

The matrix S can be used to build the weighted undirected graph $G = (V,E)$, where each node in the set V represents an individual in the sample, i.e. $V = \{n\}$, and the set E contains all the edges in G. Furthermore, the weight of the generic edge (i,j) is given by $w_{ij} = s_{ij}$ if $i \neq j$, and $w_{ij} = 0$ otherwise. We can then define a random walk $\mathscr{X}$ on G, with state space V. Let $d_i = \sum_{j=1}^{n} w_{ij}$, $i = 1,\ldots,n$ and $D = \operatorname{diag}(d_1,\ldots,d_n)$. We define the transition matrix of $\mathscr{X}$ as $P = D^{-1}W$. If G is connected, $\mathscr{X}$ has π as invariant distribution, with $\pi_i = \frac{d_i}{\sum_{i,j} w_{ij}}$ [4]. The random walk we have just defined represents an artificial stochastic flow such that the probability of moving from i to j is proportional to w_{ij}, i.e. to the similarity between i and j. Such a dynamics induces some high density subsets of V, i.e. subsets where the random walker spends a long time before moving to other clusters, separated by low weight edges. In such a context, community detection algorithms attempt to identify an optimal partition of V. We shall refer, in particular, to the so called map equation [10, 11] that attempts to find a partition of V such that the length of the code describing the behaviour of $\mathscr{X}$ is minimised. Let M be a partition of V. The map equation computes the entropy $L(M)$, which is strictly related to P and π. The optimal partition minimises $L(M)$. As stated in [10], “the map equation calculates the minimum description length of a random walk on the network for a two-level code that separates the important structures from the insignificant details based on the partition M”.

As an example, figures 1 and 2 show the results of an application to the well known galaxy data set [9].

References

1. Antoniak, C.E.: Mixtures of Dirichlet processes with applications to Bayesian non parametric problems. Annals of Statistics. **2**, 1152–1174 (1974).
2. Cappé, O., Moulines, E., and T., Rydén: Inference in Hidden Markov Models. Springer, New York (2005)
3. Frühwirth-Schnatter, S.: Finite Mixture and Markov Switching Models. Springer: Berlin (2006)
4. Lovász, L.: Random walks on graphs: a survey. In: Combinatorics, Paul Erdös is eighty, pp. 353–397. János Bolyai Math. Soc., Budapest (1993)
5. MacEachern, S.N., Clyde, M., and Liu, J.S: Sequential importance sampling for nonparametric Bayes models: The next generation. The Canadian Journal of Statistics, **27**, 251–267 (1999)
6. McLachlan, G., and Peel, D.: Finite Mixture Models. Wiley, New York (2000)
7. Medvedovic, M., and Sivaganesan, S.: Bayesian infinite mixture model based clustering of gene expression profiles. Bioinformatics, **18**, 1194–1206 (2002)
8. Medvedovic, M., Yeung, K.Y., and Bumgarber, R.E.: Bayesian mixture model based clustering of replicated microarray data. Bioinformatics, **20**, 1222–1232 (2004)
9. Roeder, K.: Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. JASA **85** 617–624
10. Rosvall, M. and Bergstrom C. T.: Maps of random walks on complex networks reveal community structure. PNAS. **105**, 1118–1123 (2008) doi: www.pnas.org/cgi/doi/10.1073/pnas.0706851105
11. Rosvall, M., Axelsson, D., and Bergstrom C.T.: The map equation. Eur. Phys. J. Special Topics 178, 1323 (2009) doi: 10.1140/epjst/e2010-01179-1
12. Stephens, M.: Dealing with label switching in mixture models. J. R. Statistic. Soc. B. **62**, 795–809 (2000)