Random Partition Models for Microclustering Tasks

Brenda Betancourt,[1] Giacomo Zanella,[2] and Rebecca C. Steorts[3]

University of Florida,[1], Bocconi University,[2] and Duke University[3]

Summary

- Many models assume the number of data points in each cluster grows linearly with the total number of data points.
- Examples include infinitely exchangeable clustering processes.
- **Entity resolution**: Integration of multiple data sources removing duplicated information → clustering approach.
- The size of each cluster remains small even for large databases.
- Requires models that yield clusters whose sizes grow sublinearly with the size of the data set.

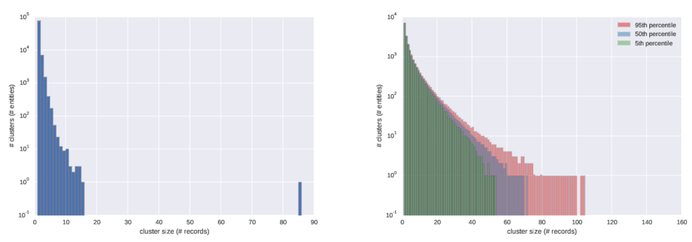

FIGURE 1: Two illustrations of the microclustering problem: (*Left*) a histogram of the number of records per entity (i.e., cluster sizes) for a database of 100,000 campaign finance donations from 2011–2012. *Right*: A histogram of the cluster sizes generated from a Chinese restaurant process simulation with concentration parameter 0.1.

Microclustering property

Traditional Clustering

- A number of popular clustering applications assume priors on partitions such as the Dirichlet process and Pitman-Yor process distributions.
- *Any* infinitely exchangeable partition distribution on the positive integers has a Kingman paintbox representation.
- Having a Kingman paintbox representation implies that the number of data points in each cluster grows (a.s.) linearly with the total number of data points N.
- By contrast, in entity resolution (and other problems), we expect the size of the clusters to be small even for large datasets.

Microclustering

A sequence of random partitions $(\Pi_n)_{n=1}^{\infty}$ satisfies the *microclustering property* if $M_n/n \to 0$ in probability as $n \to \infty$, where the size M_n of the largest cluster of Π_n.

- Equivalently, $M_n \,/\, n \to 0$ in probability as $n \to \infty$.
- Microclustering requires sacrificing either (i) exchangeability or (ii) self-consistent marginalization of the partitions.
- We sacrifice (ii) and assume that $(\mathcal{C}_1, \mathcal{C}_2, \dots)$ is an exchangeable sequence of clusters with finite sizes.

Exchangeable Sequences of Clusters

We propose a flexible and tractable prior distribution for a random partition that is appropriate for microclustering tasks such as entity resolution.

- K: number of clusters (random).
- $\Pi_n = \{C_1, \dots, C_K\}$ is a partition of $[n]$
- $S_k,\ k = 1, \dots K$: size of the cluster C_k.
- $z_i,\ n = 1, \dots n$: cluster assignment of the nth data point.

A random partition $\Pi_n \sim ESC_{[n]}(P_{\boldsymbol{\mu}})$ can be generated as follows.

1. Conditional on E_n, sample $\boldsymbol{\mu} \sim P_{\boldsymbol{\mu}}$ and $S_1, S_2, \dots | \boldsymbol{\mu} \overset{iid}{\sim} \boldsymbol{\mu}$.
2. Define K as the unique positive integer such that $\sum_{j=1}^{K} S_j = n$.
3. Define the cluster allocation variables $(z_1, \dots, z_n)$ as a uniformly at random permutation of the vector

$$(\underbrace{1, \dots, 1}_{S_1 \text{ times}}, \underbrace{2, \dots, 2}_{S_2 \text{ times}}, \dots\dots, \underbrace{K, \dots, K}_{S_K \text{ times}})$$

.

where $\boldsymbol{\mu} = (\mu_s)_{s=1}^{\infty}$, μ_s is the probability of a cluster of size s, $P_{\boldsymbol{\mu}}$ is the distribution of $\boldsymbol{\mu}$ and

$$E_n = \left\{ \text{there exists } k \in \mathcal{N} \text{ such that } \sum_{j=1}^{k} S_j = n \right\}.$$

Conditional on the event E_n, K is defined as the unique positive integer such that $\sum_{j=1}^{K} S_j = n$.

Properties of ESC Model

Prediction Rule

Let $\mathbf{z} = (z_1, ..., z_n)$ be the cluster allocation variables of $\Pi_n \sim ESC_{[n]}(P_{\boldsymbol{\mu}})$. For any $i = 1, \dots, n$, the conditional distribution of z_i given $\mathbf{z}_{-i} = \mathbf{z} \backslash z_i$ and $\boldsymbol{\mu}$ is

$$\mathcal{P}(z_i = j | \mathbf{z}_{-i}, \boldsymbol{\mu}) \propto \begin{cases} (s_j + 1)\frac{\mu_{(s_j+1)}}{\mu_{s_j}} & \text{if } j = 1, \dots, k_{-i}, \\ (k_{-i} + 1)\mu_1 & \text{if } j = k_{-i} + 1\,, \end{cases}$$

where k_{-i} is the number of clusters in $\mathbf{z}_{-i}$.

Number of clusters

Assume $\sum_{s=1} s\mu_s < \infty$ and let K_n be the number of clusters of Π_n. As $n \to \infty$ it holds

$$\frac{K_n}{n} \xrightarrow{p} \left(\sum_{s=1}^{\infty} s\mu_s \right)^{-1},$$

where $\xrightarrow{p}$ denotes convergence in probability.
This implies that, if the mean of $\boldsymbol{\mu}$ is finite (i.e. $\sum_{s=1}^{\infty} s\mu_s < \infty$) the number of clusters of Π_n grows linearly with the number of data points n.

Proportion of clusters of given size

Assume $\sum_{s=1} s\mu_s < \infty$.

(a) Let $M_{s,n}$ be the number of clusters of size s in Π_n. As $n \to \infty$

$$\frac{M_{s,n}}{n} \xrightarrow{p} \frac{\mu_s}{\sum_{\ell=1}^{\infty} \ell \mu_\ell}.$$

(b) The size S_j of a cluster chosen uniformly at random from the clusters of Π_n converges in distribution to $\boldsymbol{\mu}$ as $n \to \infty$.

This implies that, asymptotically, the distribution of the size of a randomly chosen cluster from Π_n coincides with $\boldsymbol{\mu}$

Model Specification

ESC-NB Model

We model $\boldsymbol{\mu}$ as a Negative Binomial distribution truncated on $\{1, 2, \dots\}$, $\boldsymbol{\mu} = NegBin(r, p)$, meaning that, for all $s = 1, 2, \dots$; μ_s is the following deterministic function of r and p

$$\mu_s(r, p) = \gamma \frac{\Gamma(s + r)p^s}{\Gamma(r)s!},$$

where $\gamma = \frac{(1-p)^r}{1-(1-p)^r}$.

ESC-D Model

A more flexible choice is to assume $\boldsymbol{\mu} \sim Dir(\alpha, \boldsymbol{\mu}^{(0)})$, where $\boldsymbol{\mu}^{(0)} = (\mu_s^{(0)})_{s=1}^{\infty}$ is a sequence of non-negative numbers satisfying $\sum_{s=1}^{\infty} \mu_s^{(0)} = 1$. We then impose a parametric form on $\boldsymbol{\mu}^{(0)}$ in an analogous way to the ESC-NB model.

- $r > 0$ and $p \in (0, 1)$ are given prior distributions, $r \sim Gamma(\eta_r, s_r)$ and $p \sim Beta(u_p, v_p)$.
- The hyperparameters η_r, s_r, u_p and v_p are chosen to reflect the prior expectations on the distribution of the cluster sizes S_j.

Entity Resolution Model

The observed data $\boldsymbol{x} = (x_1, \dots, x_n)$ consist of n records and each record x_i contains L fields $(x_{i\ell})_{\ell=1}^{L}$. We assume that fields within clusters are independent and each of them depends only on the following field specific parameters:

- a distortion probability $\beta_\ell \in (0, 1)$ that reflects the proneness to errors of each field, and
- a density vector $\boldsymbol{\theta}_\ell = (\theta_{\ell d})_{d=1}^{D_\ell} \in [0, 1]^{D_\ell}$ that characterizes the distribution of categories per field, where $\sum_{d=1}^{D_\ell} \theta_{\ell d} = 1$.

Let $y_{j\ell}$ represent the true ℓ-th feature of the entity associated to cluster $C_j \in \Pi_n$ for $j = 1, \dots, K$, and $\zeta(\Pi_n, i)$ be a function that maps record i to its latent cluster assignment z_i according to Π_n.
The microclustering model for entity resolution is as follows:

$$x_{i\ell} | y_{j\ell}, z_i, \boldsymbol{\theta}_\ell, \beta_\ell \overset{iid}{\sim} \beta_\ell \boldsymbol{\theta}_\ell + (1 - \beta_\ell)\delta_{y_{z_i\ell}} \quad (1)$$

$$y_{j\ell} \overset{iid}{\sim} \boldsymbol{\theta}_\ell \quad (2)$$

$$z_i = \zeta(\Pi_n, i) \quad (3)$$

$$\Pi_n \sim ESC_{[n]}(\boldsymbol{\mu}) \quad (4)$$

where δ_y is the Dirac-delta function at y, and $\boldsymbol{\theta}_\ell$ is fixed and assumed to be the empirical distribution of the data.

Chaperones Algorithm

If we let $c_i \in \Pi_n$ denote the cluster containing element i, then each iteration consists of:

1. Randomly choose two *chaperones*, $i, j \in \{1, \dots, n\}$ from a distribution $P(i, j \,|\, x_1, \dots, x_n)$ where the probability of i and j given $x_1, \dots, x_n$ is greater than zero for all $i \neq j$. This distribution must be independent of the current state of the Markov chain Π_n; however, crucially, it may depend on the observed data points $x_1, \dots, x_n$.

2. Reassign each element $k \in c_i \cup c_j$ by sampling from $P(\Pi_n \,|\, n, \Pi_n \backslash k, c_i \cup c_j, x_1, \dots, x_n)$.

Step 2 is almost identical to the restricted Gibbs moves found in existing split-merge algorithms (Jain and Neal, 2004), except that the chaperones i and j can also change clusters, provided they do not abandon any of their children.

Application

The Data

- Social Diagnosis Survey (SDS): panel study of objective and subjective quality of life in Poland focused on individuals aged 16 and above.
- We take a subsample of $n = 10,000$ records from the survey waves in the years 2011, 2013, and 2015.
- The true number of clusters in the datasets is $K = 5,505$ with 2,505 singleton clusters, 1,505 clusters of size two, and 1,495 clusters of size three.
- We use the fields of sex, date of birth (day, month and year), province of residence, and education level.

Results

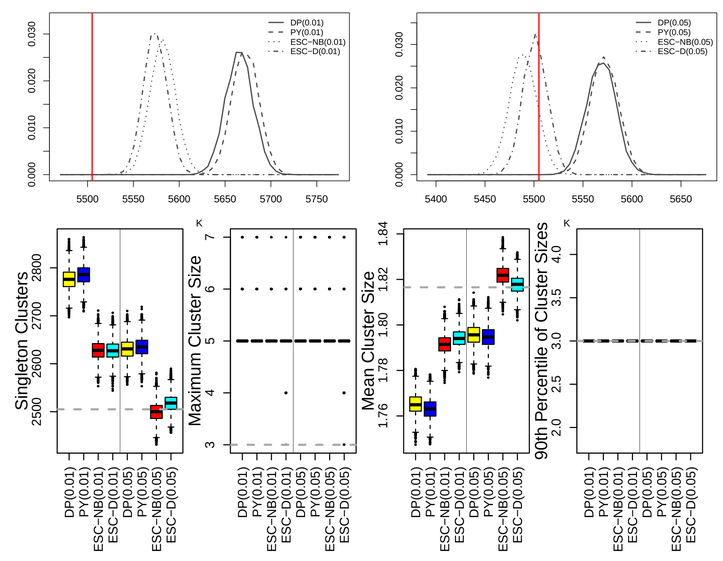

FIGURE 2: Top: Posterior distributions of the number of clusters for DP, PY, ESC-NB and ESC-D models truncating the prior distribution of the distortion probabilities at $c = 0.01$ (top left) and $c = 0.05$ (top right) for SDS subsample dataset of $n = 10,000$ records. Bottom: Side-by-side boxplots of the posterior distributions of the number of singleton clusters, max, mean and 90th percentile of cluster sizes for all models and truncation values. The red vertical lines (top) and gray dashed horizontal lines (bottom) represent the true statistic values. The ESC-NB and ESC-D models with truncation value of $c = 0.05$, ESC-NB(0.05) and ESC-D(0.05), capture the true partition statistics and show the best performance.

Discussion

- We overcome major limitations in the existing literature for microclustering models – a lack of interpretability, identifiability, and full characterization of the model asymptotic properties.
- The resulting framework offers great flexibility in terms of the prior distribution of cluster sizes, is computationally tractable and suitable for general microclustering tasks.
- The theoretical results allow to devise simple and efficient Markov chain Monte Carlo algorithms to perform statistical inference.
- Extensions to include string features in the entity resolution model.

Acknowledgements: This work was supported in part by NSF Big Data Privacy, NSF Career and the Foerster-Bernstein Postdoctoral Fellowship.