CRPClustering: An R Package for Bayesian Nonparametric Chinese Restaurant Process Clustering with Entropy

A Vignette

Masashi Okada
Okada Algorithm Private Invention Research Laboratory, Japan
okadaalgorithm@gmail.com

CRPClustering version 1.0 2018-01-14

Abstract

Clustering is a scientific method which finds the clusters of data and many related methods are traditionally researched for long terms. Bayesian nonparametrics is statistics which can treat models having infinite parameters. Chinese restaurant process is used in order to compose Dirichlet process. The clustering which uses Chinese restaurant process does not need to decide the number of clusters in advance. This algorithm automatically adjusts it. Then, this package can calculate clusters in addition to entropy as the ambiguity of clusters.

Introduction

Clustering is a traditional method in order to find the clusters of data and many related methods are studied for several decades. The most popular method is called as K-means (Hartigan 1979). K-means is an algorithmic way in order to search the clusters of data. However its method needs to decide the number of clusters in advance. Therefore if the data is both high dimensions and a complex, deciding the accurate number of clusters is difficult and normal Bayesian methods are too. For that reason, Bayesian nonparametric methods are gradually important as computers are faster than ever. In this package, we implemented Chinese restaurant process clustering (CRP) (Pitman 1995). CRP can compose infinite dimensional parameters as Dirichlet process (Ferguson 1973). It acts like customers who sit at tables in a restaurant and has a probability to sit at a new table. As a result, Its model always automates clustering. Moreover, we added the method which calculates the entropy (Yngvason 1999) of clusters into this package. It can check the ambiguity of the result. Then, we explain the clustering model and how to use it in detail. Finally, an example is plotted on a graph.

Background

Chinese Restaurant Process

Chinese restaurant process is a metaphor looks like customers sit at a table in Chinese restaurant. All customers except for x_i have already sat at finite tables. A new customer x_i will sit at either a table which other customers have already sat at or a new table. A new customer tends to sit at a table which has the number of customers more than other tables. A probability equation is given by

$$
\begin{aligned}
&p(z_i = k | x_{1:n}, z_{1:n}^{\setminus i}, \alpha, \mu_0, \rho_0, a_0, b_0) \\
&= \begin{cases} p(x_i|\mu_k, \tau) \times \frac{n_k^{\setminus i}}{n-1+\alpha} & if \quad k \in K^+(Z_{1:n}^{\setminus i}) \\ p(x_i|\mu_k, \tau) \times \frac{\alpha}{n-1+\alpha} & \mu_k \sim N(\mu_0, (\tau\rho_0)^{-1} I) \quad if \quad k = |K^+(Z_{1:n}^{\setminus i})| + 1 \end{cases}
\end{aligned}
$$

where $n_k^{\setminus i}$ denotes the number of the customers at a table k except for i and α is a concentration parameter.

Markov Chain Monte Carlo Methods for Clustering

Markov chain Monte Carlo (MCMC) methods (Liu 1994) are algorithmic methods to sample from posterior distributions. If conditional posterior distributions are given by models, it is the best way in order to acquire parameters as posterior distributions. The algorithm for this package is given by

Many iterations continue on below

i) Sampling z_i for each i $(i = 1, 2, \cdots, n)$

$$p(z_i = k | x_{1:n}, z_{1:n}^{\setminus i}, \alpha, \mu_k, \mu_0, \tau, \rho_0) = \begin{cases} p(x_i|\mu_k, \tau) \times \frac{n_k^{\setminus i}}{n-1+\alpha} \\ p(x_i|\mu_k, \tau) \times \frac{\alpha}{n-1+\alpha} \end{cases} \qquad \mu_k \sim N(\mu_0, (\tau\rho_0)^{-1} I)$$

$$z_i \sim Multi(p(z_i = 1), p(z_i = 2), \cdots, p(z_i = \infty))$$

ii) Sampling μ_k for each k $(k = 1, 2, \cdots, \infty)$

$$\mu_k \sim p(\mu_k | x_{1:n}, z_{1:n}, \tau, \mu_0, \rho_0) = N(\mu_k | \frac{n_k}{n_k + \rho_0} \bar{x_k} + \frac{\rho_0}{n_k + \rho_0} \mu_0, (\tau(n_k + \rho_0))^{-1} I)$$

$$\bar{x_k} = \frac{1}{n_k} \sum_{i=1}^{n} \delta(z_i = k) x_i$$

First several durations of iterations which are called as "burn in" are error ranges. For that reason, "burn in" durations are abandoned.

Clusters Entropy

Entropy denotes the ambiguity of clustering. As a result of a simulation, data x_i joins in a particular table. From the total numbers n_k of the particular table k at the last iteration, a probability p_k at each cluster k is calculated. The entropy equation is given by

$$Entropy = -\sum_{k=1}^{\infty} \frac{n_k}{n} \log_2 \frac{n_k}{n}$$

Installation

CRPClustering is available through GitHub (https://github.com/jirotubuyaki/CRPClustering). If download from GitHub, you can use devtools by the commands:

```
> library(devtools)
> install_github("jirotubuyaki/CRPClustering")
```

Once the packages are installed, it needs to be made accessible to the current R session by the commands:

```
> library(CRPClustering)
```

For online help facilities or the details of a particular command (such as the function crp_gibbs) you can type:

```
> help(package="CRPClustering")
```

Methods

Method for Chinese Restaurant Process Clustering

```
> z_result <- crp_gibbs(data, mu=c(0,0), sigma=0.5, sigma_table=12,
                        alpha=1.0, ro_0=0.1, burn_in=10, iteration=100)
```

This method calculates CRP clustering.
Let's arguments be:

- data : a matrix of data for clustering. Row is each data i and column is dimensions of each data i.
- mu : a vector of center points of data. If data is 3 dimensions, a vector of 3 elements like "c(2,4,7)".
- sigma : a numeric of data variance.
- sigma_table : a numeric of table position variance.
- alpha : a numeric of a CRP concentration rate.
- ro_0 : a numeric of a CRP mu change rate.
- burn_in : an iteration integer of burn in.
- iteration : an iteration integer.

Let's return be:

- z_result : a vector denotes the number of a cluster for each data i.

Visualization Method

```
> crp_graph_2d(data, z_result)
```

This method exhibits a two dimensional graph for the method "crp_gibbs".
Let's arguments be:

- data : a matrix of data for clustering. Row is each data i and column is dimensions of each data i.
- z_result : a vector denotes the number of a cluster for each data i and it is the output of the method "crp_gibbs".

Example

Data is generated from three normal distributions and $\mu_0 = (-1, 1)$, $\mu_1 = (-1.3, -1.3)$, $\mu_2 = (1, -1)$ and $\sigma_0 = 0.3$, $\sigma_1 = 0.02$, $\sigma_2 = 0.3$. The result is plotted on a graph and each data joins in any cluster. The graph is given by below

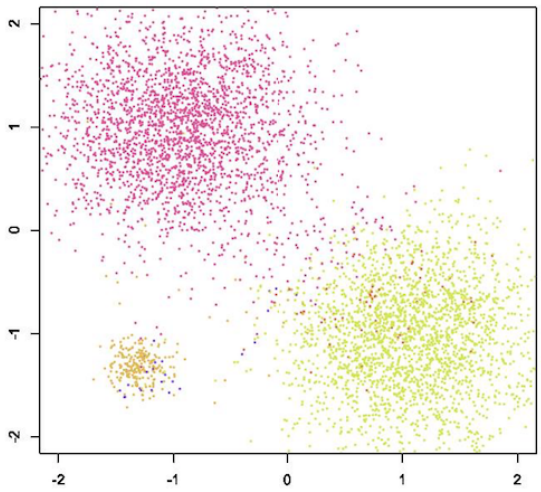

Figure 1. CRP clustering result

Conclusions

Chinese restaurant process clustering was implemented and explained how to use it. Computer resources are limited. Computer processing power is the most important problem. And several improvements are planed. Please send suggestions and report bugs to okadaalgorithm@gmail.com.

Acknowledgments

This activity would not have been possible without the support of my family and friends. To my family, thank you for much encouragement for me and inspiring me to follow my dreams. I am especially grateful to my parents, who supported me all aspects.

References

Ferguson, Thomas. 1973. "Bayesian Analysis of Some Nonparametric Problems," Annals of Statistics. 1 (2): 209–230.

Hartigan, M. A., J. A.; Wong. 1979. "Algorithm as 136: A K-Means Clustering Algorithm," Journal of the Royal Statistical Society, SeriesC. 28 (1): 100–108. JSTOR 2346830.

Liu, Jun S. 1994. "The Collapsed Gibbs Sampler in Bayesian Computations with Applications to a Gene Regulation Problem," Journal of the American Statistical Association 89 (427): 958–966.

Pitman, Jim. 1995. "Exchangeable and Partially Exchangeable Random Partitions," Probability Theory and Related Fields 102 (2): 145–158.

Yngvason, Elliott H. Lieb; Jakob. 1999. "The Physics and Mathematics of the Second Law of Thermodynamics," Physics Reports Volume:310 Issue:1 1–96.