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

A Vignette

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

CRPClustering version 1.0 2018-01-12

Abstract

Clustering is a scientific method which finds the clusters of data. And lots of methods are traditionally researched for long terms. Bayesian Nonparametric is a 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 a number of clusters in advance. This algorithm automatically adjusts it. And this package can calculate clusters in addition to entropy as ambiguity of clusters.

Introduction

Clustering is a traditional method in order to find clusters of data. And lots of methods are invented for several decades. The most popular method is called as K-mean(Hartigan 1979). K-mean is an algorithmic way in order to search clusters of data. But its method needs to decide a number of clusters in advance. So if the data is both high dimensions and a complex, deciding accurate numbers of clusters is difficult. And normal bayesian methods are too. For that reason, Bayesian Nonparametric methods are gradually important as computers are faster. In this package, we implement Chinese Restaurant Process Clustering (CRP)(Pitman 1995). CRP can compose infinite dimensional parameters as Dirichlet Process(Ferguson 1973). It acts like a customers who sit at tables in a restaurant and has a probability to sit at a new table. As result, Its model always automates Clustering. And we add the method which calculates entropy(Yngvason 1999) of clusters into this package. It can check ambiguity of the result. Then we explain the Clustering model and how to use it in detail. finally, an example is explained.

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 a number of customers more than other tables. A probability equation is given by

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

$n_k^{\setminus i}$ expresses the number of the customers at a table k. And α is a concentration parameter.

Markov Chain Monte Carlo Methods for CRP Clustering

Markov chain Monte Carlo methods (MCMC) are algorithmic methods(Liu 1994) 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

Lots of 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 u_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 burn_in are error ranges. For that reason, burn_in durations are abandoned.

Cluster Entropy

Entropy expresses ambiguity of Clustering. As the result of 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

$$H(x) = -\sum_{k=1}^{\infty} \frac{n_k}{n} \log_2 \frac{n_k}{n}$$

Installation

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")
```

Method

Calculating Method for CRP 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 args be

- data : a matrix of data for Clustering. Rows are data i and cols are dimensions of data.
- mu : numbers of center points of data. If data is 3 dimensions, a vector of 3 elements like "c(2,4,7)".
- sigma : a number of data variance.
- sigma_table : a number of table position variance.
- alpha : a number of a CRP concentrate rate.
- ro_0 : a number of a CRP mu change rate.
- burn_in : iteration numbers of burn in.
- iteration : iteration numbers.

Let's return be

- z_result : a vector expresses cluster numbers 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 args be

- data : a matrix of data for Clustering. Rows are data i and cols are dimensions of data.
- z_result : a vector expresses cluster number for each data i. And 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 as a graph and each data joins in any cluster. The graph is given by below.

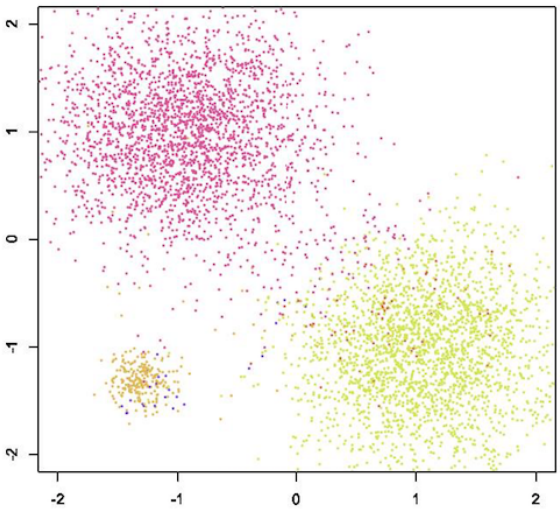

Figure 1. CRP Clustering Result

Conclusions

Chinese Restaurant Process Clustering was implemented and explained how to use it. Computers resources is 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 lots of 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.