Name:	Sungjae Jung
Affiliation:	Dept. of Biomedical Informatics, Ajou University School of Medicine, Suwon, South Korea
Email:	sungjae.2425@gmail.com
Presentation type (s):	**Poster**

A regression analysis method for distributed large-scale data analysis

Sungjae Jung, MS[1], Rae Woong Park, MD PhD[1,2]
[1]Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, South Korea; [2]Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, South Korea

Abstract

Traditional regression methods are designed to analyze data from a single source, and estimates coefficients according to the maximum likelihood estimation method via the Newton-Raphson method. The use of distributed data in regression analysis has been explored in previous studies; however, these methods are not suitable for analyzing large amounts of multi-source data. To enable large-volume data processing, we applied the concept of data chunks. A single regression model can be derived from separated multi-source datasets rather than integrated data. This method can be applied to distributed research networks configured with a common data model.

Introduction

Traditional regression methods such as logistic regression, Poisson regression and Cox Proportional Hazard model are designed to analyze data from a single source. Applying existing statistical methods to a distributed research network (DRN) creates the opportunity to conduct research using data from multiple organizations. The use of distributed data has been explored in previous studies[1, 2]; however, these methods are not suitable for analyzing large amounts of multi-source data. The purpose of this study is to propose a logistic regression method that can analyze large amounts of data from multiple data sources at a distributed environment.

Methods

A regression coefficient is estimated according to the maximum likelihood estimation (MLE) method via the Newton-Raphson method[3, 4]. The likelihood function such as log likelihood function[1] and Breslow's partial likelihood function[2] is depends on a regression method. The first and second derivatives of likelihood function are necessary to estimate coefficients. Newton-Raphson iteration can be updated as follow:

$$\beta^{new} = \beta^{old} - \left(\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^T}\right)^{-1}\frac{\partial l(\beta)}{\partial\beta}$$

To enable large-volume data processing, we applied the concept of data chunks (Figure 1), used for large-scale data processing in R, in each Newton-Raphson iteration (Figure 2). This method can be extended to a distributed research network (Figure 3).

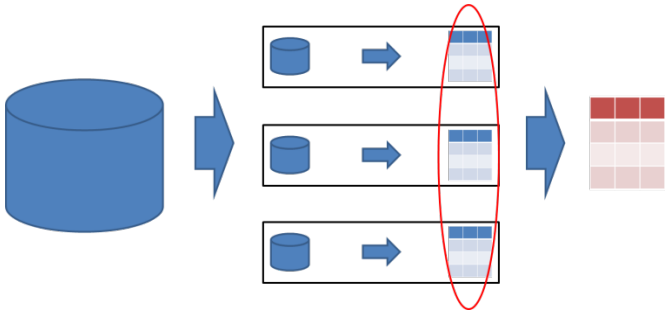

Figure 1. The concept of data chunks. Split huge sized data into smaller sized data chunks. Summed matrix elements are same as the result of using the original data.

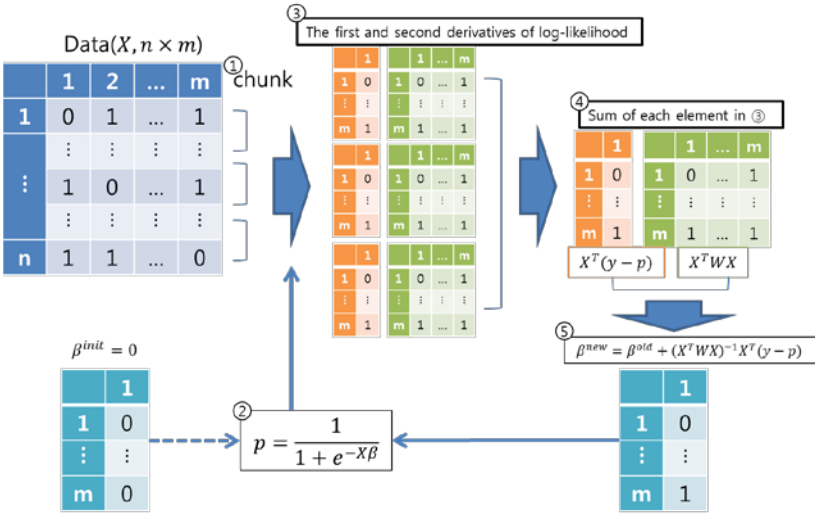

Figure 2. Process to estimate logistic coefficient with data chunk. ①Split the data into data chunks. ②Calculate probabilities with logit for each data chunk. ③Calculate the first and second derivatives of log-likelihood function for each data chunk. ④Sum the matrix elements of ③. ⑤Calculate new coefficients. ⑥Iterate ② to ⑤ until variation of estimated coefficients converged to specific precision.

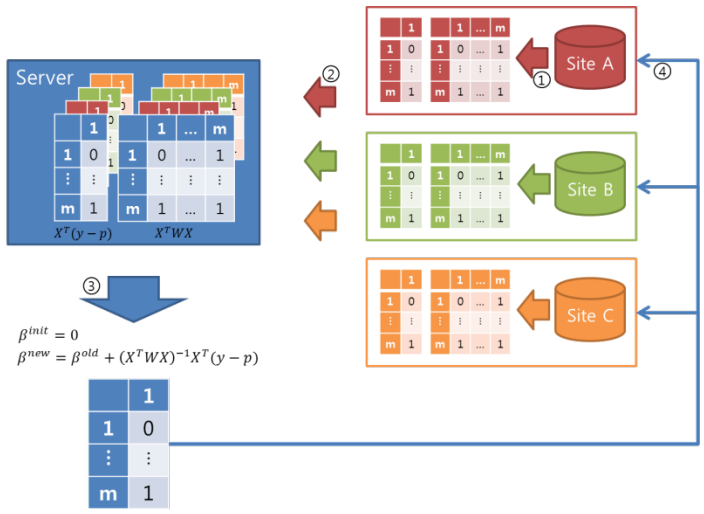

Figure 3. Process to estimate logistic coefficient in a distributed research network. ①Calculate the first and second derivatives of log-likelihood function as follow method of Figure 2. ②Send the matrices to the server. ③Calculate new coefficients. ④Send the estimated coefficients to each client. ⑤Iterate ① to ④ until variation of estimated coefficients converged to specific precision.

Conclusion

We propose a method enabling to derive a ~~A~~ single regression model ~~can be derived~~ from separated multi-source and large volume datasets in distributed environment without merging the source data~~rather than integrated data~~. This method can be applied to DRNs configured with a common data model. We expect that this method will ~~to~~ facilitates ~~-~~collaborative research requiring logistic regression for multiple data sources when the data ~~can not~~cannot be take out from the data partners.~~at home and abroad.~~

References

1. Wu Y, Jiang X, Kim J, Ohno-Machado L. Grid Binary LOgistic REgression (GLORE): building shared models without sharing data. J Am Med Inform Assoc. 2012;19(5):758-64.
2. Lu CL, Wang S, Ji Z, Wu Y, Xiong L, Jiang X, et al. WebDISCO: a web service for distributed cox model learning without patient-level data sharing. J Am Med Inform Assoc. 2015;22(6):1212-9.
3. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning : data mining, inference, and prediction. 2nd ed. New York: Springer; 2009. xxii, 745 p. p.
4. Hosmer DW, Lemeshow S, Cook ED. Applied logistic regression. 2nd ed. New York: Wiley; 2000. xii, 375 p. p.