

Exploration of heterogeneous treatment effects under distributed storage

Abstract

Exploration of heterogeneous treatment effects under distributed storage

Keywords: subgroup analysis, two-step algorithm, distributed storage, ADMM

1. Introduction

With different industry life cycle stages, the pharmaceutical industry has also produced different medical theories. At present, based on the individualized characteristics of different patients, individualized medical theory will be an inevitable trend of the medical industry. Complex analytical tools are needed in personalized treatment strategies. One of the key statistical challenges is to correctly identify subgroups from heterogeneous populations. To address this issue, a popular approach is to use a mixture model analysis Everitt & Hand (1981), treating data as coming from different subgroups, each with its own parameter values. Farewell (1982) analysis the survival data with long-term survivors by mixture model. All along, the mixture model is continuous improvement. Muthén & Shedden (1999) discussed the analysis of an extended finite mixture model where the latent classes corresponding to the mixture components for one set of observed variables influence a second set of observed variables. Pauler & Laird (2000) introduced a general finite mixture of nonlinear hierarchical models that allows estimates of component membership probabilities and random effect distributions for longitudinal data arising from multiple subpopulations. Rasmussen (2000) presented the infinite gaussian mixture model. Maugis et al. (2009) discussed variable selection for clustering with gaussian mixture models. Shen & He (2015) proposed a structured logistic-normal mixture model for

the purpose of identifying a subgroup that has an enhanced treatment effect as well as the variables that are predictive of the subgroup membership. The mixture model-based approach needs to specify an underlying distribution and the number of mixture components in the population which is often difficult to do in practice. To solve this problem, Ma & Huang (2017) proposed a concave pairwise fusion approach to subgroup analysis. They developed an alternating direction method of multipliers algorithm with concave penalties. With the rapid development of data storage and communication, the medical industry has also developed a new direction. The data exchange between hospitals has improved the accuracy of diagnosis and treatment. So far, statistical research on subgroup analysis has remained at the stand-alone level. There is few approach to do the subgroup analysis over the distributed stored data.

For distributed storage data, we can abstract the computing structure into two parts, one is the master and the other is the workers. All data interactions only occur between the master and the worker and no data communication between workers. The assumption of such a computing environment fits in with the operational logic of the actual distributed computing platform, and on the other hand, the data security of each node can also be guaranteed. more specifically, all sample data only exists in each worker node. There is no sample data in the master, which only do the map-reduce operation, and some necessary calculations after reduce data from all workers.

In this paper, We propose a two-step approach that allows us to minimize the influence of the limitations of data space separation and to perform subgroup analysis from an overall perspective. Let y_{im} be the response variable for the i^{th} subject in m^{th} node. $X_{im} = (x_{im1}, \dots, x_{imp})$ presents a set of covariates. We consider subgroup analysis of the heterogeneity driven by unknown or unobserved latent factors. Hence, we consider

$$y_{im} = \mu_{im} + x_{im}^T \beta + \epsilon_{im}, i = 1, \dots, n_m; m = 1, \dots, M, \quad (1)$$

where μ_{im} is unknown subject-specific intercepts, $\beta = (\beta_1, \dots, \beta_p)^T$ is the vector of unknown coefficients for x_{im} , and ϵ_{im} is the error term independent of x_{im}

50 with $E(\epsilon_{im}) = 0$ and $Var(\epsilon_{im}) = \sigma^2$. Model 1 can be divided into two parts, combined with the actual situation of biomedicine, x_{im} represents some patient-related essential variables, such as age, gender, etc. μ_{im} represents the factors contributing to the heterogeneity, such as different treatments. Then μ_{im} can be written as $\mu_{im} = \mu + z_{im}^T \theta$, where z_{im} represents the different treatment effects. 55 It is worth noting that we are talking about personalized medicine, which means different patients have different effects on the same treatment $\mu_{im} = \mu + z_{im}^T \theta_{im}$. Thus, model 1 becomes

$$y_{im} = \mu + z_{im}^T \theta_{im} + x_{im}^T \beta + \epsilon_{im}, i = 1, \dots, n_m; m = 1, \dots, M. \quad (2)$$

Throughout this paper, we focus on model2 by considering that even the samples of one subgroup are distributed stored, our estimation method can still 60 correctly identify their subgroups. Several authors have studied the problem of exploring homogeneity effects of covariates over a single machine. Ma & Huang (2017) proposed a concave pairwise fusion penalized least squares approach for this purpose and derive an alternating direction method of multipliers algorithm Boyd et al. (2011) for implementing the following approach.

$$Q_n(\mu, \beta; \lambda) = \frac{1}{2} \sum_{i=1}^n (y_i - \mu_i - X_i^T \beta)^2 + \sum_{1 \leq i < j \leq n} P(|\mu_i - \mu_j|, \lambda), \quad (3)$$

65 where $P(\cdot, \lambda)$ is a concave penalty function with a tuning parameter $\lambda \geq 0$. However, when we introduce the node information, the objective function 3 is very difficult to solve. The objective function becomes as follow.

$$Q_n(\mu, \beta; \lambda) = \frac{1}{2} \sum_{m=1}^M \sum_{i=1}^{n_m} (y_{im} - \mu_{im} - X_{im}^T \beta)^2 + \sum_{1 \leq i < j \leq \sum_m n_m} P(|\mu_{im} - \mu_{jm}|, \lambda). \quad (4)$$

Compared with function 3, function 4 brings a lot of problems that make the original estimation method malfunction. One problem is that The operation 70 on matrix X becomes unrealizable. Another problem is $|\mu_i - \mu_j|, 1 \leq i < j \leq \sum_m n_m$ needs a lot of data interaction, almost impossible in actual operation. Therefore, we proposed a two-step algorithm based on ADMM algorithm

and stratified sampling. Through the operation of map-reduce and two-step iteration, our algorithm can accurately identity the subgroups of data in each
75 physical node with low computational cost.

The rest of this paper is organized as follows. In Section 2 we describe the two-step approach in detail. In Section 3 we solve the two-step approach by ADMM algorithm. In Section 4, we do some numerical simulation in the spark cluster mode, compared with the calculation results when the data is stored in
80 a single machine. In Section 5, the real data is distributed at different physical nodes, mimicking the hospital data interoperability.

2. Two-step Algorithm over distributed stored data

3. Computation

4. Simulation studies

85 5. Real data example

References

- Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J. et al. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3, 1–122.
- 90 Everitt, B., & Hand, D. (1981). Finite mixture distributions chapman and hall. *New York*, .
- Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, (pp. 1041–1046).
- Ma, S., & Huang, J. (2017). A concave pairwise fusion approach to subgroup
95 analysis. *Journal of the American Statistical Association*, 112, 410–423.
- Maugis, C., Celeux, G., & Martin-Magniette, M.-L. (2009). Variable selection for clustering with gaussian mixture models. *Biometrics*, 65, 701–709.

- Muthén, B., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the em algorithm. *Biometrics*, 55, 463–469.
- 100 Pauler, D. K., & Laird, N. M. (2000). A mixture model for longitudinal data with application to assessment of noncompliance. *Biometrics*, 56, 464–472.
- Rasmussen, C. E. (2000). The infinite gaussian mixture model. In *Advances in neural information processing systems* (pp. 554–560).
- Shen, J., & He, X. (2015). Inference for subgroup analysis with a structured
105 logistic-normal mixture model. *Journal of the American Statistical Association*, 110, 303–312.