Big Data Dimension Reduction using PCA

Xuelong Wang

November 25, 2017

Challenge

1. Memory Barrier
   The size of data is too large to load into memeory
2. The computing time
   Time consuming if only single core used

Solution

1. Sufficient statistics
   Calcuated by scanning the data rows by rows
2. Parallel computation
   Map-Reduced method

Basic idea of PCA

Singular Value Decomposition

$$X_s = UDV^T \quad \text{,where} \quad x_{ij,s} = \frac{x_{ij} - \bar{x}}{s_j}$$

$U = (u_1, \ldots, u_r)$ is a n by r orthgonal matrix

$D = diag(d_1, \ldots, d_r)$ is a r by r diagonal matrix

$V = (v_1, \ldots, v_r)$ is a p by r orthgonal matrix

Basic idea of PCA

Principle Component and Loading

$$X_s = \underbrace{\left[d_1 u_1 \dots d_r u_r\right]}_{\text{PCs}} \underbrace{\begin{bmatrix} v_1^T \\ \vdots \\ v_r^T \end{bmatrix}}_{\text{Loading}}$$

- $Pc_j = d_j \boldsymbol{u_j} = X \boldsymbol{v_j}$ is the jth principle component
- The sample variance of Pc_j is d_j^2/n

Basic idea of PCA

Reduced matrix $X_{s,k}$

$$X_{s,k} = \sum_{j=1}^{k} d_j \boldsymbol{u_j} \boldsymbol{v_k} = U_k D_k V_k^T, \quad \text{Its Variation} \quad \sum_{j=1}^{k} d_j^2 / n.$$

Its proportion of the total variation is

$$\lambda_k = \frac{\sum_{j=1}^{k} d_j^2}{\sum_{j=1}^{r} d_j^2}$$

- If a small k such that $\lambda_k \approx 1$, we can use $U_k D_k$ in the follow up analysis

Follow-up analysis

Sufficient Statistics

Paralle Computation

Setting up

Reference

"A Brief Foray into Parallel Processing with R." 2013. https://beckmw.wordpress.com/2014/01/21/a-brief-foray-into-parallel-processing-with-r/.

"Foreach/Iterators User's Guide." 2013. https://packages.revolutionanalytics.com/doc/7.3.0/win/RevoForeachIterators_Users_Guide.pdf.

Gordon, Max. 2015. "How-to Go Parallel in R – Basics + Tips." https://www.r-bloggers.com/how-to-go-parallel-in-r-basics-tips/.

Weston, Steve. 2017. "Getting Started with DoParallel and Foreach." https://cran.r-project.org/web/packages/doParallel/vignettes/gettingstartedParallel.pdf.