06 Fast mining of relationships of large-scale data

- An efficient algorithm for L1-graph construction -

Abstract

As a result of a deployment of ICT, large-sized data is available these days. Since graphical structures can effectively capture sparse relationships among data, they are actively studied recently. In this study, we developed an efficient algorithm to construct L1-graph from large size of data. We reduce the computation cost by analyzing conditions with which nodes have an edge between them. Since our algorithm significantly improves the computation speed comparing to existing approaches, it can efficiently reveal relationships behind data. This indicates that we can effectively analyze large-sized data which is not considered to apply graph mining approaches. We can effectively perform recommendation and prediction based on big data. In the near future, we can obtain world-wide data as by using IoT. We believe our algorithm is useful in analyzing such data.

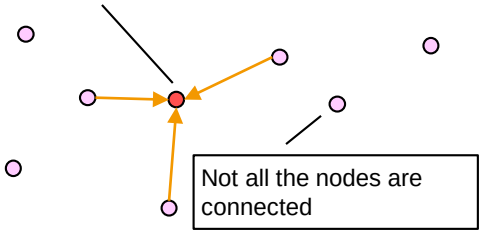
Background: L_1-grpah

- Effectively represent sparse relationships among data by using L_1 regularization

Compute edge weights by minimizing the objective function

$$\min_{w_p} \frac{1}{2M}\|x_p - w_p X\|_2^2 + \lambda\|w_p\|_1$$

w_p: Edge weight
M: #dimensions
x_p: data point
X: data matrix
λ: parametesr



Not all the nodes are connected

Problem: high computation cost

- Compute coefficients of the regression (lasso) for all the node

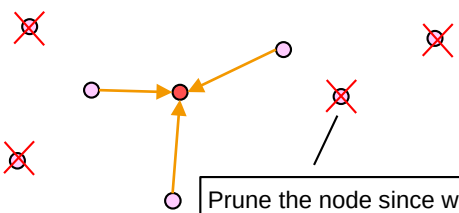
$$w_p[u] \leftarrow \begin{cases} z_p[u|w_p] - \lambda & z_p[u|w_p] > 0 \text{ かつ } |z_p[u|w_p]| > \lambda \\ z_p[u|w_p] + \lambda & z_p[u|w_p] < 0 \text{ かつ } |z_p[u|w_p]| > \lambda \\ 0 & |z_p[u|w_p]| \le \lambda \end{cases}$$

in which $z_p[u|w_p] = \frac{1}{M}\sum_{i=1}^{M} x_u[i](x_p[i] - \hat{x}_{p[i]u})$

Need parameter $z_p[u|w_p]$ in computing edge weight $w_p[u]$

Approach: node pruning

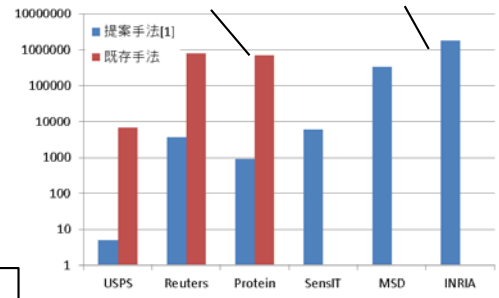
- Compute lower bound $\underline{z}_p[u|w_p]$ and upper bound $\overline{z}_p[u|w_p]$ of parameter $z_p[u|w_p]$ by using SVD



Prune the node since we know $|z_p[u|w_p]| \le \lambda$ from the bounds

Up to 1300 times faster than previous approach

We can handle large-size data



Processing time for each dataset[s]

Reference

[1] Y. Fujiwara, Y. Ida, J. Arai, M. Nishimura, S. Iwamura, "Fast Algorithm for the Lasso based L_1-Graph Construction," *PVLDB*, 10(3), pp. 229-240, 2016

Contact

Yasuhiro Fujiwara Distributed Computing Technology Project, NTT SIC
Email : fujiwara.yasuhiro(at)lab.ntt.co.jp