Higher Order Group Prediction

ABSTRACT

Aim is to predict higher order groups given a previous group interaction history.

General Terms

Hyperedge, Hypergraph

Keywords

Hyperedge Prediction, Hypergraphs, Convex Coptimization

1. PROBLEM STATEMENT

Our input is the history of previous collaborations, $\mathbf{Y} = \{\mathbf{y_1}, \mathbf{y_2}, ..., \mathbf{y_{(T-1)}}\}$. Here, each $\mathbf{y_t} = \{v_1, v_2, ..., v_k\}$ represents a particular hyperedge (group or collaboration) that occurs at time instance t with the given vertices each represented by v_i (total number of vertices being n). Our aim is to predict the possible new hyperedge $\mathbf{y_{new}}$ that occurs at $t = T$.

2. APPROACH 1

We treat this problem as an optimization task of finding the vector $\mathbf{y_{new}}$ which is nearest to all the given previous collaboration vectors $\mathbf{y_i}$, $i \in \{1, 2, ..., (T-1)\}$ and is also sparse. We define the following regularization framework:

$$\arg\min_{\mathbf{y_{new}}} \mathbf{Dist}(\mathbf{Y}, \mathbf{y_{new}}) + \lambda \|\mathbf{y_{new}}\|_1 \quad (1)$$

where $\mathbf{Dist}(\mathbf{Y}, \mathbf{y_{new}})$ represents the distance of different kinds like euclidean distance $\|\mathbf{Y} - \mathbf{1}\mathbf{y_{new}^T}\|_2$, hamming distance etc. and $\mathbf{1}$ is a vector of all ones. The second term is lasso constraint for sparsity. The parameter λ is used for tuning the extent of regularization.

3. APPROACH 2: PREDICTING SET OF GROUPS

We treat this problem as an optimization task of finding the stack of vectors $\mathbf{Y_{new}} = \{\mathbf{y_{new}^1}, ..., \mathbf{y_{new}^P}\}$ which is nearest to all the given previous collaboration vectors $\mathbf{y_i}$, $i \in \{1, 2, ..., (T-1)\}$ and is also sparse. We define the following regularization framework:

$$\arg\min_{\mathbf{Y_{new}}} \sum_{p=1}^{P} \sum_{t=1}^{T-1} \mathbf{Dist}(\mathbf{y_t}, \mathbf{y_{new}^P}) + \lambda \sum_{p=1}^{P} \|\mathbf{y_{new}^P}\|_1 \quad (2)$$

where $(\mathbf{y_t}, \mathbf{y_{new}^P})$ represents the distance of different kinds like euclidean distance $\|\mathbf{y_t} - \mathbf{y_{new}^P}\|_2$, hamming distance etc. The second term is lasso constraint for sparsity. The parameter λ is used for tuning the extent of regularization.

4. APPROACH 3

Approach 1 is static in the sense that it does not take into account the time dimension. Moreover, **Approach 1** assumes an oversimplified and over demanding in the sense that it wants to find out new groups which are similar to the whole stack of groups observed in past. Rather, a more realistic approach should consider vectors (groups or hyperedges) that are similar to some subset of previously observed groups. Moreover, it is safe to assume that these subsets are the various communities or clusters of related hyperedges observed in past. We therefore, represent the previous collaboration vectors $\mathbf{y_{ic}}$, $i \in \{1, 2, ..., (T-1)\}$ and $c \in \{1, 2, ..., N_c\}$ (N_c being the total number of communities observed in past). We define the following regularization framework:

$$\arg\min_{\mathbf{y_{new}^P}} \sum_{t=1}^{T-1} \gamma^{-t} \mathbf{Dist}(\mathbf{y_{tc}}, \mathbf{y_{new}^P}) + \lambda \|\mathbf{y_{new}^P}\|_1 \quad (3)$$

where $\mathbf{Dist}(\mathbf{y_{tc}}, \mathbf{y_{new}^P})$ represents the distance of different kinds like euclidean distance $\|\mathbf{y_{tc}} - \mathbf{1}(\mathbf{y_{new}^P})^\mathbf{T}\|_2$, hamming distance etc, p is the index of the new vectors predicted for each community, and $\mathbf{1}$ is a vector of all ones. The second term is lasso constraint for sparsity. The parameter λ is used for tuning the extent of regularization and parameter γ penalizes more if the predicted hyperedge is not similar to hyperedges in recent past. Community c for $\mathbf{y_{tc}}$ is decided using any clustering methods like KNN.

5. ISSUES

Major issues with the above approaches in general is that $\mathbf{y_{new}}$ can be predicted same as one of the training $\mathbf{y_{tc}}$ i.e. there is no penalization for exact similarity.

- Either we penalize it directly, but how ?,

WOODSTOCK '97 El Paso, Texas USA

- Or we make constraints such that $\mathbf{Dist}(\mathbf{y_{tc}}, \mathbf{y^p_{new}}) > \delta$ where δ is the minimum diversity from the existing hyperedges in community c.
- Or we can measure distance from the cluster centers rather than the $\mathbf{y_{tc}}$ directly. More generally we can take distance from a dictionary representing $\mathbf{Y}$.

6. ADDITIONAL AUTHORS

7. REFERENCES

[1] M. Bowman, S. K. Debray, and L. L. Peterson. Reasoning about naming systems. *ACM Trans. Program. Lang. Syst.*, 15(5):795–825, November 1993.

[2] J. Braams. Babel, a multilingual style-option system for use with latex's standard document styles. *TUGboat*, 12(2):291–301, June 1991.

[3] M. Clark. Post congress tristesse. In *TeX90 Conference Proceedings*, pages 84–89. TeX Users Group, March 1991.

[4] M. Herlihy. A methodology for implementing highly concurrent data objects. *ACM Trans. Program. Lang. Syst.*, 15(5):745–770, November 1993.

[5] L. Lamport. *LaTeX User's Guide and Document Reference Manual*. Addison-Wesley Publishing Company, Reading, Massachusetts, 1986.

[6] S. Salas and E. Hille. *Calculus: One and Several Variable*. John Wiley and Sons, New York, 1978.