围串标聚类 Phase 0.1

Pengfei

pengfeigaothu@gmail.com

摘要 聚类算法方案

Keywords: Clustering

1 整体目标

根据历史投标数据，对公司进行聚类，将有可能涉及围串标的公司聚成一个类。

对聚类算法的要求：

- 算法需要能够处理大数据量
- 能自动发现多个 attribute 之间的关系
- 更大数据量更多 attributes 可以提升算法性能

2 方案设计

2.1 Diagram

- 对公司进行聚类
- 在同一个投标出现过的公司才有可能聚类
- 公司之间聚类需要考虑时间维度，考虑历史投标数据

2.2 Preprocessing

- Cleaning & Imputation
- feature encoding
- 数据归一化
- 数据降维

2.3 Clustering

- gower distance / similarity measures + K-medoids / spectral clustering
- any 2 companies on a bid: using xgboost to determine if they will co-occur
- DBSCAN
- Agglomerative hierarchical clustering
- GNN / DNN embedding + K-means
- GNN clustering
- deep clustering

Key ideas:

- 扩展特征,采用 gower distance 以及 similarity distance,使用 K-prototypes 聚类方式
- 考虑使用 GNN 进行聚类: 直接聚类或者使用 K-prototypes 对 GNN embedding 聚类

2.4 Output

- 聚类结果
- 聚类结果可视化

3 Phase 0.1

3.1 Distance

Gower's distance

Gower's distance is used in statistics.
measure mixed types:

- continuous
- binary
- ordinal

Value range: 0->1.

0 is the most similar.

Definition:

Assuming p is number of features(descriptors).

$$s_{ij} = \frac{\sum_{k=1}^{p} w_k s_k}{\sum_{k=1}^{p} w_k} =< W, S > \tag{1}$$

For example, if feature k is ordinal:

$$s_k = \frac{i - j}{max(s_k) - min(s_k)} \tag{2}$$

question:

how to determine w?

Best practice is to use domain knowledge.

Historical Bis

How to encode historical bid data?

- multihot encoding with freq as values
- GNN?

3.2 Clustering

public pretrained models

Is there any public models for this task?