Bachelor thesis:

Robust Subspace Discovery in Principal Fitted Component–Based Clustering

Introduction:

Clustering in high-dimensional data is challenging. As the number of dimensions increases, data points tend to become more similar to each other. Only through projection into different subspaces can the data become separable. One way to identify these subspaces is by performing principal component analysis [1]. However, different data may build only cluster formulation under different feature subsets [2], [3]. Local principal component analysis is then proposed [3] to perform clustering and finding relevant feature dimensions for each cluster simultaneously.

By assuming each data point is equally important and dividing the feature dimensions into grid-like intervals, the CLIQUE algorithm counts the number of samples falling into each interval. It then successively merges dense grids in a combinatorial manner to form higher-dimensional units, continuing this process until the resulting units are no longer dense. A cluster is defined as the maximal set of connected dense intervals in high-dimensional space.

This raises the question of whether it is possible to define a mathematical loss function to automate this process using gradient-based optimization.

Assuming the data are Boolean, as in access control datasets, [1] proposes using Binomial mixture models to allow each object to be simultaneously assigned to multiple clusters. Expectation-maximization is then used to uncover the underlying cluster distribution.

Objective:

This thesis focuses on a comparative study between MAC [1] and CLIQUE, in terms of robustness to noise and generalization errors, i.e., the reconstruction error when using the cluster assignments of the nearest neighbors in the training set and their corresponding cluster centroids.

The ultimate goal is to lay the groundwork for proposing novel loss functions that address overlapping subspace identification in high-dimensional data.

Dataset:

- Synthetic dataset in [1]
- Real-world permission role dataset in [1]

Prediction outcome:

- Cluster assignment

Research questions:

Clustering in high-dimensional data is often challenging because distances between data points tend to become increasingly similar as the number of features grows. Instead of relying on axis-aligned subspace searches, as in CLIQUE—where clusters emerge through bottom-up exploration—can potentially overlapping subspaces be identified by maximizing data likelihood in a top-down manner?

Sensitivity analysis:

- The number of overlapping sources to discover

Contact

Interested? Then send your transcripts to xian@dbs.ifi.lmu.de

References:

[1] Greenacre, M., Groenen, P.J.F., Hastie, T. *et al.* Principal component analysis. *Nat Rev Methods Primers* **2**, 100 (2022). https://doi.org/10.1038/s43586-022-00184-w

[2] Data mining algorithm 2 lecture slides (https://www2.dbs.ifi.lmu.de/cms/Clustering_High-dimensional_Data.html)

[3] Charu C. Aggarwal, Joel L. Wolf, Philip S. Yu, Cecilia Procopiuc, and Jong Soo Park. 1999. Fast algorithms for projected clustering. SIGMOD Rec. 28, 2 (June 1999), 61–72. https://doi.org/10.1145/304181.304188

[4] Nandakishore Kambhatla and Todd K. Leen. "Dimension reduction by local principal component analysis." *Neural computation* 9.7 (1997): 1493-1516.