Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Peer Kröger
Yifeng Lu

Knowledge Discovery in Databases II

SS 2019

Exercise 4: High Dimensional Data Clustering

Exercise 4-1 Recap: Frequent Itemset Mining

Subspace clustering algorithms usually utilize bottom-up subspace search, which is the same as the frequent itemset mining introduced in KDD1. Two bottom-up subspace search techniques: Apriori and Divide-and-Conquer (Database-projection), were discussed. Given the following data set, please find out all frequent itemsets with support not less than 2:

(a) using Apriori approach

(b) using Divide-and-Conquer approach

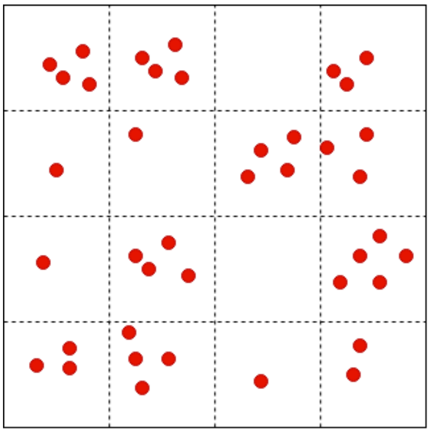
tid	a	b	c	d	e
1	1	1	0	1	1
2	1	1	0	0	0
3	0	1	1	1	0
4	1	0	1	0	1
5	0	1	0	1	1

Tabelle 1: Transaction dataset.

Exercise 4-2 Grid-based Subspace-Clustering (CLIQUE)

How many subspace clusters will be found by CLIQUE algorithm in the following dataset? ($\tau = 3, \xi = 4$)

NOTE: to keep consistency, a dense unit contains more than **or equal to** τ objects



Exercise 4-3 Density-based Subspace-Clustering (SubClu)

Show that the following statement (monotonicity of the core point property) holds:

Let D be a set of d-dimensional feature vectors, $\mathcal{A}$ the set of all attributes (dimensions/features). Further let $p \in D$ and $S \subseteq \mathcal{A}$ be a subspace (attribute subset).

Then the following holds for arbitrary $\epsilon \in \mathbb{R}^+$ and $minPts \in \mathbb{N}$:

$$\forall T \subseteq S \ : \ |\mathcal{N}_\epsilon^S(p)| \geq minPts \ \Rightarrow \ |\mathcal{N}_\epsilon^T(p)| \geq minPts$$

with $|\mathcal{N}_\epsilon^S(p)| := \{q \in D \mid L_P(\pi_S(p), \pi_S(q)) \leq \epsilon\}$.