**Ludwig-Maximilians-Universität München**
**Institut für Informatik**
PD Dr. Matthias Schubert
Markus Mauder

## Knowledge Discovery in Databases II

WS 2014/2015

## Übungsblatt 5: Clustering

### Aufgabe 5-1 $\delta$-Biclustering

(a) Formulate the $\delta$-biclustering algorithm in pseudocode.

(b) Given the following data matrix:

$$A := \begin{pmatrix} 4 & 4 & 4 \\ 4 & 0 & 5 \\ 1 & 2 & 3 \end{pmatrix}$$

Execute the $\delta$-biclustering algorithm to find the first bicluster. Use $\delta = 0.5$.

### Aufgabe 5-2 Clustering algorithms and the "curse of dimensionality"

The following effects of the "curse of dimensionality" are most relevant for clustering in high-dimensional vector spaces:

(a) **Complexity of pattern search.** More attributes are equivalent to more variables in an optimization problem. Exhaustive search is becoming harder or impossible with crowing dimensionality.

(b) **Irrelevance of distance differences.** Concepts like "proximity", "distance", or "neighborhood" are becoming less meaningful with growing dimensionality.

(c) **Irrelevant attributes.** A growing number of attributes are considered spuriously and are thus likely irrelevant for certain patterns. Relevance varies between different subsets of the considered data objects.

This problem is independent from the problem of irrelevance of distance differences (above), but it too is characterized by questionale full-dimensional object distances.

(d) **Correlated attributes.** With a growing number of attributes, the likelihood of random correlations between attributes grows. Like the problem of irrelevant attributes, this problem too can be observed in low-dimensional data.

Correlation between attributes change the shape of a data set and its contained patterns considerably. In addition, the task of interest is not only to discover hidden patterns despite these correlations, but also to discover these correlations themselves.

Consider the clustering algorithms you know from the lectures: CLIQUE, SUBCLU, PROCLUS, PreDeCon, Cheng&Church, p-cluster, ORCLUS, 4C und CASH. Which of these problems are considered, which are neglected, and which are not solved satisfactorily?

| | CLIQUE | SUBCLU | PROCLUS | PreDeCon | Cheng&Church | p-cluster | ORCLUS | 4C | CASH |
|---|---|---|---|---|---|---|---|---|---|
| Problem (a) | | | | | | | | | |
| Problem (b) | | | | | | | | | |
| Problem (c) | | | | | | | | | |
| Problem (d) | | | | | | | | | |