

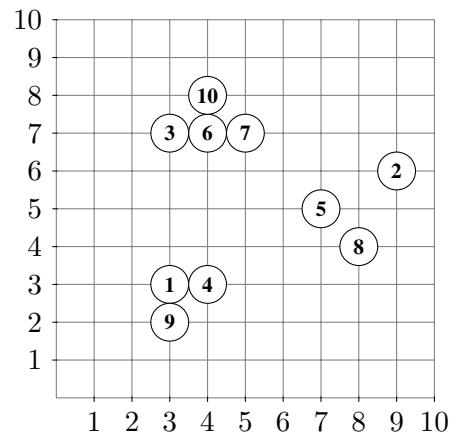
Knowledge Discovery in Databases II
 WS 2014/2015

Übungsblatt 7: Micro-Clustering

Aufgabe 7-1 BIRCH Cluster Features

Consider the following data set:

ObjID	X	Y
1	3	3
2	9	6
3	3	7
4	4	3
5	7	5
6	4	7
7	5	7
8	8	4
9	3	2
10	4	8



Calculate the BIRCH cluster features $CF = (N, \vec{LS}, SS)$

The radius R of a cluster feature CF is defined as:

$$R(CF) := \sqrt{\frac{SS}{N} - \frac{\vec{LS}^2}{N}}$$

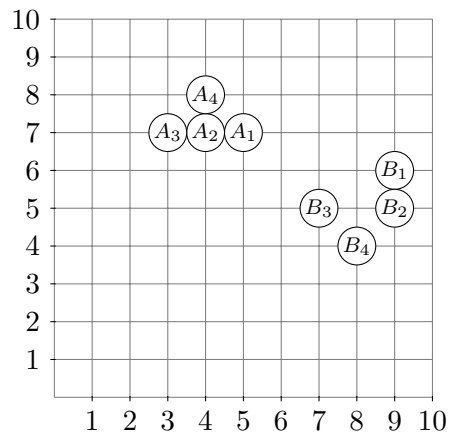
Add an element to an existing cluster feature iff adding it does not violate the threshold $R \leq T = \sqrt{2}$. If there is more than one candidate, choose the one with the smallest new radius R . Otherwise generate a new CF .

It is not necessary to construct a complete CF-tree. Use a list of leafs CF to represent the data set.

Hint: test for $R^2 \leq T^2$, not $R \leq T$.

Aufgabe 7-2 Data Bubbles

Consider the following point sets A and B :



- Represent A and B as data bubbles.
- Calculate the distance between these bubbles.
- Calculate the mean pairwise distance between the points in A and B .

How similar are the calculated distances?