

Knowledge Discovery in Databases II
WS 2015/2016

Übungsblatt 7: Micro-Clustering

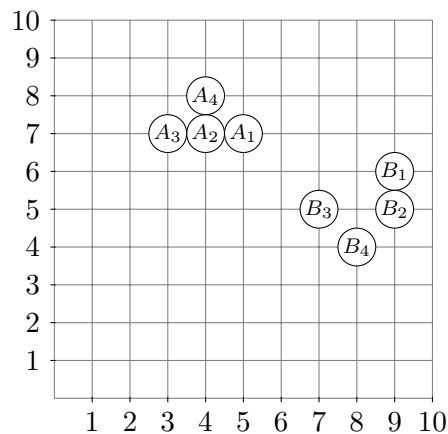
Aufgabe 7-1 **Micro Clusternig using CFs**

In this exercise, we will compress a data set by a set of cluster features having a maximum radius of *threshold*. For simplicity, it is not required to store the CFs in a CF-Tree.

- (a) Download the template `python.CF.py` and have a look at the code.
- (b) Implement a class `CF` having a constructor, a method `add` for adding a new vector, a method `unite` for joining two CFs and a method `radius` for computing the radius.
- (c) Implement a method `micro_clustering(D, threshold)` for computing a set of micro clusters from the data matrix `D` and the given threshold radius.
- (d) Where is the performance bottleneck without using a CF-Tree?

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?