MACHINE LEARNING FOR DATA STREAMS

# CLUSTERING

1. **Efficient and Effective Clustering Methods for Spatial Data Mining (1994):**
   * A cluster is represented by its **medoid**, or the most, centrally located data point in the cluster.
   * The clustering process is formalized as **searching a graph in which each node is a K-partition represented by a set of *K* medoids**, and two nodes are neighbors if the only differ by one medoid.
   * Starts with a randomly selected node. For the current node, it checks at most the *maxneighbor* number of neighbors randomly, and if a better neighbor is found, it moves to the neighbor and continues; otherwise it records the current node as a *local mínimum*, and restarts with a new randomly selected node to search for another local mínimum. It stops after the *numlocal* number of the so-called *local mínima* have been found, and returns the best of these.
   * Suffers from the same drawbacks as the above IO method wrt. efficiency.
   * It may not find a real local minimum due to the searching trimming controlled by *maxneighbor*.
2. **BIRCH: An Efficient Data Clustering Method for Very Large Databases (1996):**
   * Balanced Iterative Reducing and Clustering using Hierarchies.
   * Incrementaly and dynamically clusters incoming multidimensional metric data points to try to produce the best quality clustering with the available resources (i.e., available memory and time constraints).
   * Can typically find a good clustering with a **single scan of the data**, and improve the quality further with a few additional scans.
   * First clustering algorithm proposed in the database area to **handle “noise”** (data points that are not part of the underlying pattern) effectively.
   * Considers **metric** atributes, as in most of the Statistics literature (atributes whose values satisfy the requirements of Euclidean space, i.e., self identity and triangular inequality.
   * Offers opportunities for **parallelism**, and for **interactive** or **dynamic performance tuning** based on knowledge about the dataset, gained over the course of the execution.
   * Comparison with CLARANS (1)