Clustering in Different Perspectives

Boris Mirkin

School of Computer Science and Information Systems, Birkbeck, University of London

Abstract. This talk is an attempt at structuring and systematising the development of clustering as a discipline. As the attendants to this conference are well aware, clustering is devoted to finding and describing cohesive or similar or homogeneous parts in data, the clusters. Common clustering structures are: partition, hierarchy, and a single cluster. Among distinguishabe goals of clustering are data structuring, data describing, data generalizing, associating aspects, and information visualising.

At least four different frameworks can be distingushed in the literature on clustering as a data analysis discipline: probabilistic statistics, machine learning, data mining, and knowledge discovery. In probabilistic statistics, clustering is a method to fit a prespecified probabilistic model of the data generating mechanism. In machine learning, clustering is a tool for prediction. In data mining, clustering is a tool for finding patterns and regularities within the data. In knowledge discovery, clustering is a tool for updating, correcting and extending the existing knowledge; in this regard, clustering is but empirical classification.

Different frameworks may lead to different views on the same issues such as that of the optimal number of clusters. This is a very much important question in the probabilistic statistics perspective and it is of minor importance in the knowledge discovery perspective. Yet new experimental results on this issue will be presented.

The talk will focus then on the data mining and knowledge discovery perspectives. Specifically, the data recovery paradigm (in data mining) and cluster based conceptual descriptions and production rules (in knowledge discovery) will be outlined following a recent book by tha author (Mirkin 2005). Two real world applications will be used for illustration.

References

B. Mirkin (2005) *Clustering for Data Mining: A Data Recovery Approach*. Chapman and Hall/CRC.

Keywords