# Data Clustering: User's Dilemma

Anil K. Jain

Department of Computer Science and Engineering
Michigan State University (USA)
http://www.cse.msu.edu/~jain/

**Abstract.** Data clustering is a long standing research problem in pattern recognition, computer vision, machine learning, and data mining with applications in a number of diverse disciplines. The goal is to partition a set of n d-dimensional points into k clusters, where k may or may not be known. Most clustering techniques require the definition of a similarity measure between patterns, which is not easy to specify in the absence of any prior knowledge about cluster shapes. While a large number of clustering algorithms exist, there is no optimal algorithm. Each clustering algorithm imposes a specific structure on the data and has its own approach for estimating the number of clusters. No single algorithm can adequately handle various cluster shapes and structures that are encountered in practice. Instead of spending our effort in devising yet another clustering algorithm, there is a need to build upon the existing published techniques. In this talk we will address the following problems: (i) clustering via evidence accumulation, (ii) simultaneous clustering and dimensionality reduction, (iii) clustering under pair-wise constraints, and (iv) clustering with relevance feedback. Experimental results show that these approaches are promising in identifying arbitrary shaped clusters in multidimensional data.

P. Perner (Ed.): MLDM 2007, LNAI 4571, p. 1, 2007.