Data Clustering: 50 Years Beyond K-means

Anil K. Jain

Computer Science and Engineering
Michigan State University, USA

The practice of classifying objects according to perceived similarities is the basis for much of science. Organizing data into sensible groupings is one of the most fundamental modes of understanding and learning. As an example, a common scheme of scientific classification puts organisms in to taxonomic ranks: domain, kingdom, phylum, class, etc.). Cluster analysis is the formal study of algorithms and methods for grouping objects according to measured or perceived intrinsic characteristics. Cluster analysis does not use category labels that tag objects with prior identifiers, i.e., class labels. The absence of category information distinguishes cluster analysis (unsupervised learning) from discriminant analysis (supervised learning). The objective of cluster analysis is to simply find a convenient and valid organization of the data, not to establish rules for separating future data into categories.

The development of clustering methodology has been a truly interdisciplinary endeavor. Taxonomists, social scientists, psychologists, biologists, statisticians, engineers, computer scientists, medical researchers, and others who collect and process real data have all contributed to clustering methodology. According to JSTOR, data clustering first appeared in the title of a 1954 article dealing with anthropological data. One of the most well-known, simplest and popular clustering algorithms is K-means. It was independently discovered by Steinhaus (1955), Lloyd (1957), Ball and Hall (1965) and McQueen (1967)! A search via Google Scholar found 22,000 entries with the word clustering and 1,560 entries with the words data clustering in 2007 alone. Among all the papers presented at CVPR, ECML, ICDM, ICML, NIPS and SDM in 2006 and 2007, 150 dealt with clustering. This vast literature speaks to the importance of clustering in machine learning, data mining and pattern recognition.

A cluster is comprised of a number of similar objects grouped together. While it is easy to give a functional definition of a cluster, it is very difficult to give an operational definition of a cluster. This is because objects can be grouped into clusters with different purposes in mind. Data can reveal clusters of different shapes and sizes. Thus the crucial problem in identifying clusters in data is to specify or learn a similarity measure. In spite of thousands of clustering algorithms that have been published, a user still faces a dilemma regarding the choice of algorithm, distance metric, data normalization, number of clusters, and validation criteria. A familiarity with the application domain and clustering goals will certainly help in making an intelligent choice. This talk will provide background, discuss major challenges and key issues in designing clustering

W. Daelemans et al. (Eds.): ECML PKDD 2008, Part I, LNAI 5211, pp. 3–4, 2008.

algorithms, summarize well known clustering methods, and point out some of the emerging research directions, including semi-supervised clustering that exploits pairwise constraints, ensemble clustering that combines results of multiple clusterings, learning distance metrics from side information, and simultaneous feature selection and clustering.