Quotient Space Based Cluster Analysis[1]

[1,3]Ling Zhang and [2,3]Bo Zhang

[1]Artificial Intelligence Institute, Anhui University, Anhui, China
[2]Computer Science & Technology Department, Tsinghua University, Beijing, China
[3]State Key Lab of Intelligent Technology & Systems, Tsinghua University

Abstract: In the paper, the clustering is investigated under the concept of granular computing, i.e., the framework of quotient space theory. In principle, there are mainly two kinds of similarity measurement used in cluster analysis: one for measuring the similarity among objects (data, points); the other for measuring the similarity between objects and clusters (sets of objects). Therefore, there are mainly two categories of clustering corresponding to the two measurements. Furthermore, the fuzzy clustering is gained when the fuzzy similarity measurement is used. From the granular computing point of view, all these categories of clustering can be represented by a hierarchical structure in quotient spaces. From the hierarchical structures, several new characteristics of clustering can be obtained. It may provide a new way for further investigating clustering.

Keywords: Cluster analysis, granular computing, quotient space theory, hierarchical structure, fuzzy clustering

1. Introduction

In machine learning, there exist two basic problems: classification and clustering. In classification, there are many well-known theories and approaches such as SVM, neural networks, etc. In cluster analysis, there are many successful approaches as well. For example, partitioning method [1][2], density-based [3][4], k-means [5], k-nearest neighborhood [4], neural networks [6], etc. In despite of the existence of different clustering approaches, the aim of the clustering is to group the objects (or data, points) in a space into

[1] Supported by Chinese National Nature Science Foundation (Grant No. 60135010)

clusters such that objects within a cluster are similar to each other and objects in different clusters have a high degree of dissimilarity.

From granular computing viewpoint, the objects within a cluster can be regarded as an equivalence class. Then, a clustering of objects in space X corresponds to an equivalence relationship defined on the space. From the quotient space theory [7][8], it's known that it corresponds to constructing a quotient space of X. In each clustering algorithm, it's needed to define a specific metric to measure the similarity (or dissimilarity) among objects. So far various metrics have been adopted such as Euclidean distance, Manhattan distances, inner product, fuzzy membership function, etc. No matter what kind of measurement is used, in principle, there are basically two kinds: one for measuring the similarity among objects (or data, points), the other for measuring the similarity between objects and clusters (sets of objects).

From the above viewpoint, some clustering methods can be restated below.

Partitioning clustering: Given a universe X, a similarity function s(x, y) (≥0), and a threshold s. X is assumed to be partitioned into subsets $V_1, V_2, ..., V_k$ satisfying (1) $V_1, V_2, ..., V_k$ is a partition of X, (2) $\forall x, y \in X,\ if\ \ s(x, y) \geq s,$ then x and y belong to the same cluster.

k-means method: Given a set V in a distance space, and an integer k. V is grouped into k subsets $V_1, V_2, ..., V_k$. For each subset V_i, $x \in V_i \Leftrightarrow s(x, a_i) = \min\left\{s(x, a_j), j = 1, 2, ..., k\right\}$, where a_i is the center of set V_i. In the method, the similarity between object x and a cluster represented by a_i is used as well.

Density-based clustering CURD [9]: Clustering using references and density (CURD) is a revised version of CURE [10][11]. It is a bottom-up hierarchical clustering algorithm. First, taking each object (point) as a clustering center, then the most similar points are gradually grouped into clusters until k clusters are obtained. In this algorithm, the similarity function is described by two variables, so it is a multivariate clustering. Its clustering process will be stated in session 3.

Furthermore, the fuzzy clustering is discussed based on the fuzzy equivalence relation.

From the granular computing point of view, all these categories of clustering can be represented by a hierarchical structure in quo-