ABSTRACT

Clustering analysis is not only major tools to uncover the underlying structure of a given data set, but also the promising ways to reveal the local input-output relations of a complex system. The clustering analysis is needed in a large variety of engineering and scientific problems, such as pattern recognition and classification, machine learning, computer vision, and system modeling and more. The goal of clustering analysis is to split data into a plausible number of subgroups such that the distance between objects within a subgroup is smaller than the distance between objects belonging to different subgroups. Unfortunately, without any prior knowledge the proper number of clusters is difficult to estimate. We use several useful information, such as distance matrix, covariance matrices derived by grid clustering algorithm, entropy, density and more are combined together in our research. In our research, we propose ellipse density entropy clustering algorithm because it only uses samples without any prior knowledge to reduce sensitivities about parameters and can automatically determine the number of clusters. It can avoid the situation which user set wrong parameters and wrong clustering results. In this thesis, we use gaussian distribution for (1) the separated clusters, (2) the overlapping clusters and (3) the containing noise clusters respectively and demonstrate it is feasible. We compare with Fuzzy C-means (FCM). The experiment results show that our scheme is reliable and stable.

Keywords : Data Mining, Knowledge Discovery, Clustering analysis, Entropy, Ellipse Density Entropy Clustering Algorithm


