Chapter 9

Clustering Techniques: A Review on Some Clustering Algorithms

Harendra Kumar
Gurukula Kangri Vishwavidyalaya, India

ABSTRACT

Clustering is a process of grouping a set of data points in such a way that data points in the same group (called cluster) are more similar to each other than to data points lying in other groups (clusters). Clustering is a main task of exploratory data mining, and it has been widely used in many areas such as pattern recognition, image analysis, machine learning, bioinformatics, information retrieval, and so on. Clusters are always identified by similarity measures. These similarity measures include intensity, distance, and connectivity. Based on the applications of the data, different similarity measures may be chosen. The purpose of this chapter is to produce an overview of much (certainly not all) of clustering algorithms. The chapter covers valuable surveys, the types of clusters, and methods used for constructing the clusters.

INTRODUCTION

Clustering is the task of dividing the set of data points (populations) into a number of groups (clusters) such that data points in the same groups are more similar to other in the same group than those data points in other groups. Or clustering is a process of organizing data points into groups whose members are similar in some way. A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. Clustering can be considered the most important unsupervised learning problem as it finds a structure in a collection of unlabeled data. Such grouping is pervasive in the way human
process information, and one of the motivations for using clustering algorithms is to provide automated tools to help in constructing categories or taxonomies. These formed clusters are used as the basis for the data analysis (or data processing techniques). The formed clusters include dense areas of the data space, groups with small distances, cluster data members, intervals or particular statistical parameter. Therefore, a data clustering can be formulated as a multi-objective optimization problem. The selection of appropriate clustering algorithm and parameter settings depend on the types of data set and intended used for the results.

Almost all clustering problem is NP hard for almost all clustering objective functions. Distance-based and density-based are two of many categories of clustering algorithms. The distance-based clusters are formed by adding points that minimize intra-cluster distances and maximize inter-cluster distances. The radius and diameter of clusters can be used as intra-cluster characteristics. Density based clustering help searches for signals of unknown shape. In a density-based clustering, a cluster is connected dense component which can grow in any direction to increase the density. Density based algorithm looks for neighbours of those data points that have at least a given number of neighbouring points within a given distance on the plane and forms clusters of data-points that can be related through their common neighbours.

Clustering algorithms work very differently so it is very difficult to conclude which algorithm is the best without examining the formed data clusters. Besides choosing the right clustering algorithm, choosing the right features also plays a critical role in clustering. Moreover, there are no universally accepted and effective criteria for selecting the clustering schemes and valid features. Although, validation criteria can provide some insights on the quality of clustering solutions but even how to choose the appropriate criterion is still a problem requiring more research to do. Han and Kamber (2001) have given a very good introduction to contemporary data mining clustering techniques in their text book. Genther et al. (1994) presented a modified fuzzy clustering algorithm for parametric defuzzification in fuzzy rule base systems. Some recent defuzzification methods have been discussed by Kumar (2017). The general discussion about hierarchical clustering is available in most of the clustering books. Zahn (1971) has discuss about divisive hierarchical clustering techniques that uses the minimum spanning tree of graph. Leung et al. (2000) have derived an interesting hierarchical clustering technique which is based on scale-space theory by using a blurring process, in which each datum is regarded as a light point in an image and a cluster is represented as a blob. Morzy et al. (1999) introduce a hierarchical algorithm that uses sequential patterns as the basic element found in the database to efficiently generate data clusters and defined a co-occurrence measure, as the standard of fusion of smaller clusters. Selim and Ismail (1984) gave rigorous proof of the finite convergence of the $K$-means type algorithms.
Cuckoo Search Optimized Reduction and Fuzzy Logic Classifier for Heart Disease and Diabetes Prediction
www.igi-global.com/article/cuckoo-search-optimized-reduction-and-fuzzy-logic-classifier-for-heart-disease-and-diabetes-prediction/179319?camid=4v1a