Chapter 13

ASCCN: Arbitrary Shaped Clustering Method with Compatible Nucleoids

Renxia Wan
North University for Nationalities, China & Donghua University, China

Lixin Wang
Anhui Institute of Architecture and Industry, China

Xiaoke Su
Donghua University, China

ABSTRACT

Special clustering algorithms are attractive for the task of grouping an arbitrary shaped database into several proper classes. Until now, a wide variety of clustering algorithms for this task have been proposed, although the majority of these algorithms are density-based. In this chapter, the authors extend the dissimilarity measure to compatible measure and propose a new algorithm (ASCCN) based on the results. ASCCN is an unambiguous partition method that groups objects to compatible nucleoids, and merges these nucleoids into different clusters. The application of cluster grids significantly reduces the computational cost of ASCCN, and experiments show that ASCCN can efficiently and effectively group arbitrary shaped data points into meaningful clusters.

1. INTRODUCTION

In data mining field, clustering pays a very important role. Clustering is the task of categorizing a set of objects into different clusters such that objects in the same cluster are more similar to each other than objects in different clusters according to some defined criteria (Huang, 1998). It is useful in a number of tasks, for example, by partitioning objects into clusters, interesting object groups may be discovered, such as the groups of clients in a banking database having a heavy investment in real estate, and clustering of data streams also can find some important applications in tracking evolution of various phenomena in
medical, meteorological, astrophysical, seismic studies (Bhatnagar et al., 2009).

Cluster analysis has become the subject of active research in several fields such as statistics, pattern recognition, machine learning and data mining. Up to now, a wide variety of clustering algorithms has been proposed, and also received a lot of attention in the last few years (c.f., section 2). In these algorithms, discovery of arbitrary shaped clusters is often to be a real obstacle for their applications. Ester (1996) and Halkidi (2001) imply that some typical clustering algorithms such as k-means, CURE, CIARANS and so on will get some poor results if there are some nonconvex shape data sets or some ball-shaped data sets of significantly differing sizes in the database. To get the arbitrary shaped clusters, algorithms based on density are designed (DBSCAN is a typical one), but these algorithms also face challenges from the efficiency and the affectivity such as the computation time may be intolerable or parameters input is not “user-friendly”.

In this chapter, we present the new clustering algorithm ASCCN. It is a crisp partition method, and clusters objects with compatible nucleoids. The new algorithm requires only one input parameter, can discover arbitrary size and shaped clusters, is efficient even for large data sets especially data with high dimension. The rest of this chapter is organized as follows: In section 2, we survey related work. In section 3, we define the compatible relation. The new algorithm ASCCN is presented in section 4. The experimental results are reported to illustrate the new algorithm in section 5. Finally, we draw our conclusions in section 6.

2. RELATED WORK

There are many clustering algorithms proposed, these algorithms may be classified into partitioning, hierarchical, density and grid based methods (Han et al., 2001). Partitioning methods determine a partition of the points into clusters, such that the points in a cluster are more similar to each other than to points in different clusters. They start with some arbitrary initial clusters and iteratively reallocate points to clusters until a stopping criterion is met. They tend to find clusters with hyperspherical shapes. Examples of partitioning algorithms include k-means (MacQueen, 1967), k-prototypes (Huang, 1998), PAM (Kaufma et al., 1990), EM (Bradley et al., 1998), MaxEntEDA (Tan et al., 2005), and MeSH Graph (Zhang, H. et al., 2008). Hierarchical clustering methods can be either agglomerative or divisive, the agglomerative method starts with each point as a separate cluster, and successively performs merging until a stopping criterion is met, and the divisive method begins with all points in a single cluster and performs splitting until a stopping criterion is met. The result of a hierarchical clustering method is a tree of clusters called a dendogram. Examples of hierarchical clustering methods include BIRCH (Zhang et al., 1996), CURE (Guha et al., 1998), MeSH Ontology (Zhang, J. et al., 2008) and SM/DynGSC (Song et al., 2009). Density-based clustering methods try to find clusters based on the density of points in regions. Dense regions that are reachable from each other are merged to formed clusters. Density-based clustering methods excel at finding clusters of arbitrary shapes. Examples of density-based clustering methods include DBSCAN (Ester et al., 1996) and DENCLUE (Hinneburg et al., 1998). Grid-based clustering methods quantize the clustering space into a finite number of cells and then perform the required operations on the quantized space. Cells containing more than a certain number of points are considered to be dense. Contiguous dense cells are connected to form clusters. Examples of grid-based clustering methods include STING (Wang et al., 1997), PNMBG (Wan et al., 2009) and CLIQUE (Agrawa et al., 1998). Furthermore, some fuzzy methods are also introduced into the clustering task (Cannon et al., 1986; Hore et al., 2007; Kwok et al., 2002; Pal et al., 1995).