Discovering Clusters of Arbitrary Shapes

Rachsuda Jiamthapthaksin, Jiyeon Choo, Christian Giusti1, Dan Jiang, Christoph F. Eick, and Ricardo Vilalta
Department of Computer Science, University of Houston
Houston, Texas 77204-3010, U.S.A
{rachsuda, jchoo, djiang, ceick, vilalta}@cs.uh.edu

Clustering is an unsupervised learning technique that is used for exploratory data analysis, for summary generation, and as a preprocessing step for other data mining tasks. Our on-going research applies clustering in scientific discovery, such as identifying pollution hotspots in environmental data, discovering co-location patterns concerning chemicals on the Martian surface and in the Texas ground water, and for region discovery in general. Due to the characteristics and diverse nature of the data used, clusters may be of arbitrary shapes and can be nested within one another. Examples of such shapes are chain-like patterns that represent active and inactive volcanoes, as depicted in Figure 1 (a). Generally traditional clustering algorithms, such as k-means and k-medoids, fail to detect non-spherical shapes, as demonstrated in Figure 1 (b). Our research aims to find solutions for this challenging problem. We focus on developing novel techniques that discover clusters of arbitrary shapes efficiently and effectively. Additionally, the following other criteria are taken into consideration: speed and scalability, comprehensive parameters tuning, robustness to noise and outliers, and the capability to detect clusters at different levels of granularity.

Two approaches are currently explored in order to find clusters of arbitrary shapes. The first approach is a generic clustering framework that approximates arbitrary shape clusters using unions of small convex polygons [CJCCGE2007], as depicted in Figure 2. One of benefits of this framework its flexibility by allowing for three plug-in components: 1) the formation of large number of sub-clusters, 2) the construction of neighboring relation of sub-clusters (Fig. 3), and 3) the fitness function that measures cluster quality. In this framework, fitness functions play an important role in capturing arbitrary shapes of clusters—unfortunately, the use of fitness functions in clustering algorithm has not been studied much by past research. We investigate two approaches for acquiring fitness functions: 1) using shape...

---

1 Department of Mathematics and Computer Science, University of Udine
Via delle Scienze, 33100, Udine, Italy. {giusti@dimi.uniud.it}
signatures [SS06] as summarized information to capture shape characteristic of clusters, and 2) learning fitness functions based on the domain expert’s or other feedback.

Figure 2: An illustration of MOSAIC’s Clustering Approach for the Complex9 Dataset

Figure 3: An illustration of the neighboring relations between sub-clusters

The second approach explores the use of density estimation techniques for discovering arbitrary shape clusters [Jiang2006, JEC2007]. These techniques rely on influence functions in which a point's influence on another point decreases as the distance between the two points increases. Based on influence functions, density functions are constructed, as depicted in Fig. 4, and clusters are formed by using hill climbing algorithms that associates objects in the dataset with maxima of the so-defined function: objects that are associated with the same maximum belong to the same cluster.

The ability to discover clusters with arbitrary shapes is important in many real applications, 3 of which are highlighted in this paragraph. Finding non-spherical regions is of critical importance for hotspot and co-location discovery in spatial datasets [EVJW2006, EDPS2007]; otherwise, arbitrary shape, regional patterns will remain undetected in spatial data mining. Another application is the use of clustering as a preprocessing step for classification. In [VAE2003], a decomposition of classes into subclasses in order to simplify the approximation of class distributions is proposed. It is clear that such decompositions do not always have spherical shapes,
and therefore the capability to discover arbitrary shape clusters is important for class decomposition. Third, in the image processing field an arbitrary shape clustering technique is important to obtain a spatial clustering of features extracted from an image. In [GPP2006] a method for discovering objects in images using spatial clustering of features is presented. The method extracts features from a picture through a weak segmentation procedure, and uses the displacement of features over the picture to subdivide the picture in meaningful parts, relying on clustering algorithms.

References


