Abstract

Cluster analysis has been an object of research since the 1980s for finding the natural groups of objects in the data so that similar objects stay within the same clusters while different objects stay in different clusters. It is undoubted that cluster analysis is important for a wide range of scientific and industrial processes such as data mining, computer vision, signal processing, and census research. *k*-means-like algorithms are the most used unsupervised machine learning techniques for handling such problems of cluster analysis. In the context of a *k*-means-like algorithm, a proper structure for representing the clusters and an appropriate distance measure for measuring the distance between objects and clusters must be accordingly defined. In general, a *k*-means-like algorithm seeks for the optimal clusters that can minimize the total distance from all objects to their nearest clusters, which makes the cluster representation and distance measure become very important to achieve such clustering target. With different types of data, the formulas of cluster representations must differ accordingly. For instance, *centroid* is specified for numerical data while *mode* and *representative* are specified for categorical data. For the data with both numerical and categorical attributes, the *prototype* structure can be effectively applied by hybridizing the *centroid* and *representative*.

On one hand, a basis *k*-means algorithm is a local optimization technique that can easily return a locally optimal solution. To achieve a better or the global solution, the clustering algorithm must seek the solutions several times with different initial states. For this reason, a "good" initial state is very important to achieve the global optimum. In this research, we first aim to propose a new scheme to use a dimension reduction technique so-called Locality-Sensitivity Hashing (LSH) to predict the "good" initial state of the cluster so that the global optimal can be potentially obtained. The empirical experiment using real and synthetic datasets showed that our proposed method LSH-k-representatives and LSH-*k*-prototypes not only can outperform other related works in terms of clustering accuracy but also have the best consistency for clustering categorical and mixed data, respectively. However, the proposed LSH-based cluster prediction requires extra processes in order to create the LSH hash table, which makes the proposed method not ready for handling big data yet.

On the other hand, the complexity of a *k*-means-like algorithm varies linearly with the volume of data while the volume of data is exploded day by day. Thus, it is essential to reduce the complexity of *k*-means-like algorithms so that they become capable of processing big and real data. Dimension reduction and data sample are the most used techniques that can approximate the clustering procedures. However, these approaches change the nature of the data instead of changing the algorithm to make it more appropriate. This dissertation also fills such shortcoming by proposing a new heuristic approach for approximately reducing the complexity of a typical *k*-means-like algorithm. In detail, the proposed method can avoid the potential unnecessary distance computations from objects to cluster representations in each iteration. Consequently, after applying our proposed method into LSH*k*-representatives, the incorporated algorithm can process up to 2 to 32 times faster than its own original version with comparable clustering accuracy.

Moreover, we also extend a kernel-based representation of so-called LSH*k*-prototypes to make it capable of fuzzy clustering of categorical data. The LSH-based cluster prediction technique is then extended to estimate the fuzzy clusters of categorical data. Eventually, the proposed fuzzy clustering algorithm so-called LSHF*k*-centers can outrun other state-of-the-art fuzzy clustering approaches.

Keywords: Cluster analysis, fuzzy cluster analysis, categorical data, mixed data, LSH, cluster initialization, big data.