Discovering Arbitrarily Shaped Clusters in High-Dimensional Numerical Data

Abstract

Clustering is an unsupervised machine learning process, which groups similar data points in a cluster based on the intrinsic structure of the unlabeled dataset. Clustering has applications in geospatial analysis, social-network analysis, community detection, image processing, bioinformatics, anomaly detection, robotics and navigation, customer segmentation, product categorization, and many other areas. Datasets of these areas are usually unlabeled, high-dimensional, large-sized, arbitrarily shaped, and may contain noise. Unsupervised categorization of these datasets through clustering for exploratory data analysis is very important and has many real-life applications. Traditional clustering methods like *k*-Means, *k*-Medians, *k*-Medoids, and GMM fail to cluster such datasets, since they work well for low-dimensional and small to medium-sized datasets with spherical clusters. Seminal algorithms like single-linkage hierarchical clustering (SLHC), DBSCAN, CLIQUE, and Spectral clustering (SC) cluster arbitrarily shaped datasets. However, these methods require user defined input parameters, which are very difficult to estimate. To overcome these limitations, recently the STICA algorithm is reported, which is a parameter-less algorithm and automatically generates optimal number of arbitrarily shaped clusters. However, SLHC, DBSCAN, CLIQUE, SC, and STICA fail to preserve intrinsic behavior of high-dimensional datasets. Most of the real-world datasets with high ambient dimensions contain very low intrinsic dimensional manifolds. Euclidean metric fails to preserve intrinsic manifold geometry and topology leading to failure of these methods. As the STICA algorithm is simple and parameter-less, there is a promising scope of improving the STICA algorithm for high-dimensional data. One way of combating Euclidean (L_2-norm) failure, Intrinsic Fractional L_p-quasi-norm (IFLp) with $0 < p < 1$ may be used. No such work has been reported in the literature yet. However, IFLp will require estimation of k and p parameters from the intrinsic characteristics of the dataset. Another way of combating Euclidean failure may be using shared nearest neighbor (SNN) count for neighborhood estimation. This was used for limited attempt on lower dimensional dataset. The SNN computation involves determining the kth nearest neighbors for all data points, where estimation of k is difficult. To overcome the Euclidean failure in high dimensional datasets, in this research, IFLp-quasi-norm or SNN or hybrid of these two will be incorporated for spanning tree (ST) constructions using BFS-tree as done in STICA algorithm. To make this intended algorithm parameter-less, the values of p, k, and edge-length threshold for BFS-tree construction will be estimated based on the intrinsic characteristics of the dataset within the algorithm. Depending on several approaches, several ST construction mechanisms will be developed. Development of theoretical and computational basis, C language implementation, testing, and experimental validation using synthetic and real-world datasets with performance evaluation will be undertaken in this research project.