Abstract

Clustering has been recognized as an important and valuable capability in the data mining field. Instead of finding clusters in the full feature space, subspace clustering is an emergent task which aims at detecting clusters embedded in subspaces. Most of previous works in the literature are density-based approaches, where a cluster is regarded as a high-density region in a subspace. However, the identification of dense regions in previous works lacks of considering a critical problem, called “the density divergence problem” in this thesis, which refers to the phenomenon that the region densities vary in different subspace cardinalities. Without considering this problem, previous works utilize a density threshold to discover the dense regions in all subspaces, which incurs the serious loss of clustering accuracy (either recall or precision of the resulting clusters) in different subspace cardinalities. To tackle the density divergence problem, in this thesis, we devise a novel subspace clustering model to discover the
clusters based on the relative region densities in the subspaces, where the clusters are regarded as regions whose densities are relatively high as compared to the region densities in a subspace. Based on this idea, different density thresholds are adaptively determined to discover the clusters in different subspace cardinalities. Due to the infeasibility of applying previous techniques in this novel clustering model, we also devise an innovative algorithm, referred to as DENCOS (DENsity Conscious Subspace clustering), to adopt a divide-and-conquer scheme to efficiently discover clusters satisfying different density thresholds in different subspace cardinalities. Another approach for subspace clustering in high dimensional data is proposed using Genetic Approach. The GAs work with a population of individuals representing abstract representations of feasible solutions. Each individual is assigned a fitness that is a measure of how good solution it represents. The better the solution is, the higher the fitness value it gets. The population evolves towards better solutions. The evolution starts from a population of completely random individuals and iterates in generations. In each generation, the fitness of each individual is evaluated. Individuals are stochastically selected from a current population (based on their fitness), and modified by means of operators mutation and crossover to form a new population. It is capable of optimizing the number of clusters for tasks with well formed and separated clusters. As validated by our extensive experiments on retail data set, GENETIC can discover the clusters in all subspaces with high quality, and the efficiency of GENETIC outperforms previous works using DENCOS.

Reference

Index Terms

Computer Science  Data Mining

Key words

Density Conscious  High Dimensional  Genetic

Algorithms