Abstract

With the ever-increasing large volumes of data, efficient, reliable, robust algorithms are needed to discover hidden patterns in the data. Clustering is an unsupervised machine learning technique that aims to identify groups of observations with similar characteristics. However, clustering high-dimensional data is overwhelming due to the challenge that only a subset of the observed variables is informative to the clustering process. Many variables selection techniques have been utilized in conjunction with clustering to determine the best feature subspace. However, the majority of techniques neglect the challenge of addressing highly correlated data. Moreover, they are often computationally intensive and require a long period of time to run. In this dissertation, the author first provides a comprehensive literature review of the most popular R packages available for clustering. Although some studies were conducted for the same purpose, their scope was limited to only nine clustering packages in R. The comprehensive literature review also discusses the main characteristics of these R packages, addresses their limitations, highlights the similarities and differences between them, provides a guide for the R users on the most appropriate method to use, and identifies the current gaps in the literature as future directions. Toward the second part of this dissertation, the author introduces novel, robust algorithms for clustering high-dimensional data. The introduced algorithms efficiently select the significant variables in conjunction with the clustering process. Rigorous testing is performed to assess the performance of the introduced algorithms in terms of synthetic datasets, benchmark datasets, and real-world application scenarios. The experimental results show that the novel algorithms significantly outperform the state-of-the-art R packages for clustering.