Abstract

High dimensional data is phenomenon in real-world data mining applications. Developing effective clustering methods for high dimensional dataset is a challenging problem due to the curse of dimensionality. Usually k-means clustering algorithm is used but it results in time consuming, computationally expensive and the quality of the resulting clusters depends on the selection of initial centroid and the dimension of the data. The accuracy of the resultant value perhaps not up to the level of expectation when the dimension of the dataset is high because we cannot say that the dataset chosen are free from noisy and flawless. Hence to improve the efficiency and accuracy of mining task on high dimensional data, the data must be pre-processed by an efficient dimensionality reduction method. This paper proposes a method in which the high dimensional data is reduced through Principal Component Analysis and then bisecting k-means clustering is performed on the reduced data where there is no initialization of
the centroids.

Reference

- Chris Ding and Xiaofeng He, “K-Means Clustering via Principal Component Analysis”, In proceedings of the 21st International Conference on Machine Learning, Banff, Canada, 2004
- Sandro Saitta, Combining PCA and K-means March 26, 2007 by Filed under: PCA, k-means
- Principal component analysis From Wikipedia, the free encyclopedia
- Merz C and Murphy P, UCI Repository of Machine Learning Databases.
- Sergio M. Savaresi and Daniel L. Boley, On the performance of Bisecting K-Means and PDDP.
- A k-Means-Based Projected Clustering Algorithm, Yufen Sun, Gang Liy and Kun Xu, 2010 Third International Joint Conference on Computational Science and Optimization, DOI 10.11.09

Index Terms

Computer Science Data Mining
Reducing and Clustering high Dimensional Data through Principal Component Analysis

Key words

Keywords K-means Dimensionality Reduction Principal Component Analysis