A Fast Approach to Clustering Datasets using DBSCAN and Pruning Algorithms

Abstract

Among the various clustering algorithms, DBSCAN is an effective clustering algorithm used in many applications. It has various advantages like no a priori assumption needed about the number of clusters, can find arbitrarily shaped clusters and can perform well even in the presence of outliers. However, the performance is seriously affected when the dataset size becomes large. Moreover, the selection of the two input parameters, Eps and MinPts, has a great impact on the clustering performance. To solve these two problems, this paper modifies the traditional DBSCAN algorithm in two manners. The first method uses K-dimensional tree instead of the traditional R-tree algorithm while the second method includes a locally sensitive hash procedure to speed up the process of clustering and increase the efficiency of clustering. The algorithms use a k-distance graph method to automatically calculate Eps and MinPts. Experimental results show that both the algorithms are efficient in terms of scalability and speeds up the clustering process in an efficient manner.

References

A Fast Approach to Clustering Datasets using DBSCAN and Pruning Algorithms


extraction methods on classification performance in surface inspection systems, Special Issue
Experiments in parallel clustering with DBSCAN, Euro-Par 2001 Parallel Processing,
with sampling techniques. International Journal of Business Intelligence and Data Mining, Vol.
5, No. 1, Pp. 94–111.
algorithms for large data sets, Proceedings of the 33rd Hawaii International Conference on

Index Terms

Computer Science

Artificial Intelligence

Keywords
DBSCAN  Speed Optimization  Nearest Neighbour Search  KD-Tree  Locally
Sensitive Hashing