Abstract

Detecting outliers in database (as unusual objects) using Clustering and Distance-based approach is a big desire. Minimum spanning tree based clustering algorithm is capable of detecting clusters with irregular boundaries. In this paper we propose a new algorithm to detect outliers based on minimum spanning tree clustering and distance-based approach. Outlier detection is an extremely important task in a wide variety of application. The algorithm partition the dataset into optimal number of clusters. Small clusters are then determined and considered as outliers. The rest of the outliers (if any) are then detected in the clusters using Distance-based method. The algorithm uses a new cluster validation criterion based on the geometric property of data partition of the dataset in order to find the proper number of clusters. The algorithm works in two phases. The first phase of the algorithm creates optimal number of clusters, where as the second phase of the algorithm detect outliers in the clusters. The key
feature of our approach is it combines the best features of Distance-based and Clustering-based outlier detection to find noise-free/error-free clusters for a given dataset without using any input parameters.

Reference


- A. Hardy, “On the number of clusters”, Computational Statistics and Data Analysis, 23,
Discovering Local Outliers using Dynamic Minimum Spanning Tree with Self-Detection of Best Number of Clusters

- E.Knorr and R.Ng, “Algorithms for Mining Distance-based Outliers in Large Data sets”, Proc.the 24th International Conference on Very Large Databases(VLDB),pp.392-403, 1998.
Discovering Local Outliers using Dynamic Minimum Spanning Tree with Self-Detection of Best Number of Clusters


Index Terms

Computer Science  Information Retrieval

Key words

Euclidean minimum spanning tree  Subtree  Clustering

Eccentricity
Cluster validity

Cluster Separation

Small Clusters

Outliers