Meta clusters through minimum spanning tree based clustering for performance analysis of students

T. Karthikeyan
Department of Computer Science
PSG College of Arts and Science
Coimbatore, Tamil Nadu
India

S. John Peter*
Department of Computer Science and Research Center
St Xavier's College
Palayamkottai, Tamil Nadu
India

S. Chidambaranathan
Department of MCA
St Xavier's College
Palayamkottai, Tamil Nadu
India

Abstract

Clustering is a process of discovering group of objects such that the objects of the same group are similar, and objects belonging to different groups are dissimilar. A number of clustering algorithms exist that can solve the problem of clustering, but most of them are very sensitive to their input parameters. Minimum Spanning Tree based clustering algorithm is capable of detecting clusters with irregular boundaries. In this paper we propose Minimum Spanning Tree based clustering algorithm to find Meta clusters (cluster of clusters). The algorithm uses a new cluster validation criterion based on the geometric property of data partition of the data set in order to find the proper number of clusters at each level. The algorithm works in two phases. The first phase of the algorithm create clusters with guaranteed intra-cluster similarity, where as the second phase of the algorithm create dendrogram using the clusters as objects with guaranteed inter-cluster similarity. In this paper we presented experimental result on some synthetic data sets namely student semester mark. Experimental

*E-mail: jaypeeyes@rediffmail.com (Corresponding author)

Journal of Discrete Mathematical Sciences & Cryptography
Vol. 14 (2011), No. 4, pp. 349–367

result shows that the proposed algorithm performs better than K-means algorithm. The first phase of the algorithm uses divisive approach, where as the second phase uses agglomerative approach. In this paper we used both the approaches to analyze the performance of the students.

Keywords: *Euclidean Minimum Spanning Tree, Clustering, Eccentricity, Hierarchical clustering, Dendrogram, Cluster Separation.*

1. Introduction

Clustering is a process of discovering group of objects such that the objects of the same group are similar, and objects belonging to different groups are dissimilar. Clustering, as important tool to explore the hidden structures of modern large database, has been proposed in the literature. Because of the huge variety of the problem and data distributions, different techniques, such s hierarchical, partitional, density-based and model-based approaches, have been developed and no techniques are completely satisfactory for all the cases. Minimum Spanning Tree based clustering algorithm overcomes many of the problems faced by the classical algorithms.

The problem of determining the correct number of clusters in a data set is perhaps the most difficult and ambiguous part of cluster analysis. The "true" number of clusters depends on the "level" on is viewing the data. Another problem is due to the methods that may yield the "correct" number of clusters for a "bad" classification [7]. Furthermore, it has been emphasized that mechanical methods for determining the optimal number of clusters should not ignore that the fact that the overall clustering process has an unsupervised nature and its fundamental objective is to uncover the unknown structure of a data set, not to impose one. For these reasons, one should be well aware about the explicit and implicit assumptions underlying the actual clustering procedure before the number of clusters can be reliably estimated or, otherwise the initial objective of the process may be lost. As a solution for this, Hardy [7] recommends that the determination of optimal number of clusters should be made by using several different clustering methods that together produce more information about the data. By forcing a structure to a data set, the important and surprising facts about the data will likely remain uncovered.

In some applications the number of clusters is not a problem, because it is predetermined by the context [8]. Then the goal is to obtain a