Using MTree Data Structure to Develop New Clustering Algorithms

Student: Cosmin Genoiu (genoiu@yahoo.com)
Coordinator: Conf. Dr. Ing. Cristian Mihăescu (mihaescu@software.ucv.ro)

Project Goal

Implement a new clustering algorithm based on MTree wich can bring some advantages over other classic algorithms, like the possibility of using MTree specific methods RangeQuery and KNN.

The new algorithm must be integrated into an open-source library (WEKA), and then used by an e-learning tool designed for the use of professors in order to assist them in a better monitoring of students activity from an online educational environment.

The quality metrics (sum of squared errors and time taken to build the tree) obtained using datasets for clustering have to be „pretty good, often enough".

Short Description

The main improvement brought to the MTree classic implementation is the utilisation of the Expectation - Maximization clustering algorithm to decide when and how a node will be split. Also, using this implementation there will be only two levels: one for the root and one for the leafs. The root contains the centroids of the clusters represented by the leaf nodes.

MTreeClusterer can receive 2 input parameters: the maximum number of clusters which can obtained and the seed used by the EM to split the nodes.

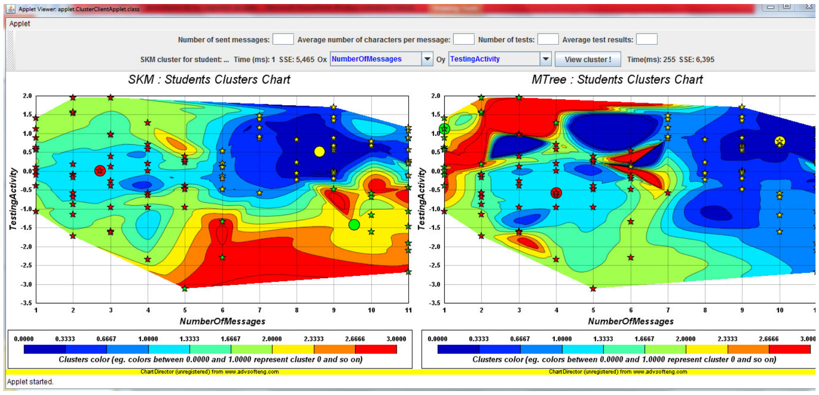
Two specific MTree methods were implemented: RangeQuery and KNN. These methods provide a faster way to obtain the nearest neighbours offering the possibility to exclude some clusters(leaf-nodes) from the search.

The Weka package containing MTreeClusterer was uploaded on sourceforge and can be downloaded from the following location:

http://sourceforge.net/projects/wekamtreeclusterer/files/MTreeClusterer.zip/download

The algorithm was integrated into an e-learning application composed of a server-side which has the role of handeling the data(connecting to a database, reading the necessary data about students from which the ARFF file is created, creating clusters (groups) of students based on students specific attributes, and a client side represented by a java applet that runs in an Internet browser, connects to the server, takes necessary data and displays the students grouped according to certain features chosen by the professor.

The tool integrates K-Means and MTreeClusterer as clustering algorithms for grouping students and facilitating the customization of parameters and number of clusters that are displayed. A data analyst can make a comparison between the two distributions and choose the one that better suits his needs. In this case, the teacher can choose one of the presented distributions to interact with the students.



Applet Viewer: applet.ClusterClientApplet.class

Applet

Number of sent messages: Average number of characters per message: Number of tests: Average test results:

SKM cluster for student: ... Time (ms): 1 SSE: 5,465 Ox NumberOfMessages Oy TestingActivity View cluster ! Time(ms): 255 SSE: 6,395

Clusters color (eg. colors between 0.0000 and 1.0000 represent cluster 0 and so on)

Applet started.

Comparation between SKM and MTree clustering results