Abstract

Clustering analysis is the most significant step in data mining. This paper discusses the k-means clustering algorithm and various distance functions used in k-means clustering algorithm such as Euclidean distance function and Manhattan distance function. Experimental results are shown to observe the effect of Manhattan distance function and Euclidean distance function on k-means clustering algorithm. These results also show that distance functions furthermore affect the size of clusters formed by the k-means clustering algorithm.

References

- Source: collection of regression datasets by Luis Torgo (ltorgo@ncc. up. pt) at
Effect of Distance Functions on Simple K-means Clustering Algorithm

http://www.ncc.up.pt/~ltorgo/Regression/DataSets.html

- D. Randall Wilson and Tony R. Martinez 1997 "Improved Heterogeneous Distance Functions" Journal of Artificial Intelligence Research 6 (1997) 1-34 Submitted 5/96; published 1/97 © 1997 AI Access Foundation and Morgan Kaufmann Publishers. All rights reserved
- Antoni Moore 2002 "The case for approximate Distance Transforms" Presented at SIRC 2002 – The 14th Annual Colloquium of the Spatial Information Research Centre University of Otago, Dunedin, New Zealand December 3-5th 2002
- Michael Steinbach, Levent Ertöz and Vipin Kumar, "The Challenges of Clustering High Dimensional Data", Access to computing facilities was provided by AHPCRC and the Minnesota Supercomputing Institute.
- Pavel Berkhin, "Survey of Clustering Data Mining Techniques", Accrue Software, Inc. Author's Address: Pavel Berkhin, Accrue Software, 1045 Forest Knoll Dr., San Josh, CA, 95129; e-mail: pavelb@acrrue.com

Index Terms

Computer Science  Data Mining

Keywords

K-means clustering  distance functions  clustering  Euclidean distance function  Manhattan distance function