Accepted Manuscript

Mechanisms to improve clustering uncertain data with *UK*means

Chuan-Ming Liu, Zhendong Niu, Kuan-Teng Liao

PII: S0169-023X(16)30328-7

DOI: 10.1016/j.datak.2018.05.004

Reference: DATAK 1653

To appear in: *Data & Knowledge Engineering*

Received Date: 28 November 2016

Revised Date: 9 April 2018

Accepted Date: 21 May 2018

Please cite this article as: C.-M. Liu, Z. Niu, K.-T. Liao, Mechanisms to improve clustering uncertain data with *UK*means, *Data & Knowledge Engineering* (2018), doi: 10.1016/j.datak.2018.05.004.

Mechanisms to Improve Clustering Uncertain Data with UKmeans

Chuan-Ming Liu[a,*], Zhendong Niu[b,*], Kuan-Teng Liao[a]

[a] *Computer Science and Information Engineering, National Taipei University of Technology, Taipei, Taiwan*
[b] *School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China*

Abstract

Uncertain data in Kmeans clustering, namely UKmeans, have been discussed in decade years. UKmeans clustering, however, has some difficulties of time performance and effectiveness because of the uncertainty of objects. In this study, we propose some modified UKmeans clustering mechanisms to improve the time performance and effectiveness, and to enable the clustering to be more complete. The main issues include (1) reducing the consideration of time performance in clustering, (2) increasing the effectiveness of clustering, and (3) considering the determination of the number of clusters. In time performance, we use simplified object expressions to reduce the time spent in comparing similarities. Regarding the effectiveness of clustering, we propose compounded factors including the distance, the overlapping of clusters and objects, and the cluster density as the clustering standard to determine similarity. In addition, to increase the effectiveness of clustering, we also propose the concept of a cluster boundary, which affects the belongingness of an object by the overlapping factor. Finally, we use the evaluating approach of the number of uncertain clusters to determine the appropriate the number of clusters. In the experiment, clustering results generated using strategies commonly used in processing uncertain data clustering in UKmeans clusters are compared. Our proposed model shows more favorable

*Corresponding author
Email addresses: cmliu@csie.ntut.edu.tw (Chuan-Ming Liu), zniu@bit.edu.cn (Zhendong Niu), t8599008@csie.ntut.edu.tw (Kuan-Teng Liao)