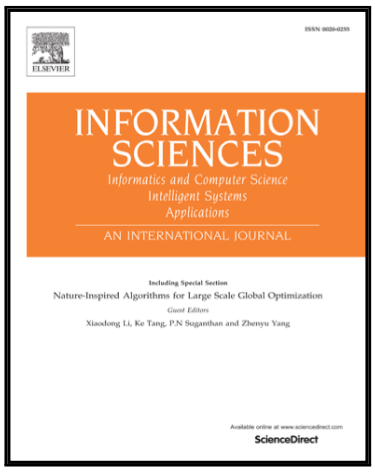

Clustering based on grid and local density with priority-based expansion for multi-density data

Shaoqun Dong, Jianjun Liu, Yuhan Liu, Lianbo Zeng, Chaoshui Xu, Tingying Zhou

PII: S0020-0255(18)30623-6
DOI: https://doi.org/10.1016/j.ins.2018.08.018
Reference: INS 13862

To appear in: *Information Sciences*

Received date: 12 September 2017
Revised date: 3 August 2018
Accepted date: 8 August 2018

Please cite this article as: Shaoqun Dong, Jianjun Liu, Yuhan Liu, Lianbo Zeng, Chaoshui Xu, Tingying Zhou, Clustering based on grid and local density with priority-based expansion for multi-density data, *Information Sciences* (2018), doi: https://doi.org/10.1016/j.ins.2018.08.018

Clustering based on grid and local density with priority-based expansion for multi-density data

Shaoqun Dong[a,b], Jianjun Liu[c,*], Yuhan Liu[c], Lianbo Zeng[a,b,*], Chaoshui Xu[d], Tingying Zhou[e]

[a]*State Key Laboratory of Petroleum Resources and Prospecting, China University of Petroleum, Beijing 102249, China*
[b]*College of Geoscience, China University of Petroleum, Beijing 102249, China*
[c]*College of Science, China University of Petroleum, Beijing 102249, China*
[d]*School of Civil, Environmental and Mining Engineering, University of Adelaide, Adelaide, 5005, Australia*
[e]*Bairong Financial Information Service Co., Ltd., Beijing 100086, China*

Abstract

Clustering based on grid and density for multi-density datasets plays a key role in data mining. In this work, a clustering method that consists of a grid ranking strategy based on local density and priority-based anchor expansion is proposed. In the proposed method, grid cells are ranked first according to local grid properties so the dataset is transformed into a ranked grid. An adjusted shifting grid is then introduced to calculate grid cell density. A cell expansion strategy that simulates the growth of bacterial colony is used to improve the completeness of each cluster. An adaptive technique is finally adopted to handle noisy cells to ensure accurate clustering. The accuracy, parameter sensitivity and computation cost of the proposed algorithm are analysed. The performance of the proposed algorithm is then compared to other clustering methods using four two-dimensional datasets, and the applicability of the proposed method to high-dimensional, large-scale dataset is discussed. Experimental results demonstrate that the proposed algorithm shows good performance in terms of accuracy, de-noising capability, robustness (parameters sensitivity) and computational efficiency. In addition, the results show that the proposed algorithm can handle effectively the problem of multi-density clustering.

Keywords: Data clustering, Grid ranking, Local density, Adjusted shifting grid, Expanding by priority, De-noising

*Corresponding author
Email addresses: liujj@cup.edu.cn (Jianjun Liu), lbzeng@sina.com (Lianbo Zeng)