P.O.Box 8718, Beijing 100080, China	Journal of Software July 2003,14(7):1267-1274
E-mail: jos@iscas.ac.cn	ISSN 1000-9825, CODEN RUXUEW, CN 11-2560/TP
http://www.jos.org.cn	Copyright © 2003 by The Editorial Department of Journal of Software

A Mean Approximation Approach to a Class of Grid-Based Clustering Algorithms

LI Cun-Hua, SUN Zhi-Hui

LI Cun-Hua, SUN Zhi-Hui (Department of Computer Science and Engineering, Southeast University, Nanjing 210096, China)
Authors information: LI Cun-Hua was born in 1963. He is a Ph.D. candidate at the Department of Computer Science and Engineering, Southeast University. His research interests are database, information system and KDDM. SUN Zhi-Hui was born in 1941. He is a professor and doctoral supervisor at the Department of Computer Science and Engineering, Southeast University. His current research areas are database, information system and KDDM.
Corresponding author: LI Cun-Hua, Phn: 86-518-5817691, Fax: 86-518-5806171, E-mail: cli@hhit.edu.cn
Received 2002-04-23; Accepted 2002-11-04

Abstract
In recent years, the explosively growing amount of data in numerous clustering tasks has attracted considerable interest in boosting the existing clustering algorithms to large datasets. In this paper, the mean approximation approach is discussed to improve a spectrum of partition-oriented density-based algorithms. This approach filters out the data objects in the crowded grids and approximates their influence to the rest by their gravity centers. Strategies on implementation issues as well as the error bound of the mean approximation are presented. Mean approximation leads to less memory usage and simplifies computational complexity with minor lose of the clustering accuracy. Results of exhaustive experiments reveal the promising performance of this approach.

Li CH, Sun ZH. A mean approximation approach to a class of grid-based clustering algorithms. *Journal of Software*, 2003,14(7):1267~1274.
http://www.jos.org.cn/1000-9825/14/1267.htm

摘要
随着聚类分析对象数据集规模的急剧增大,改进已有的算法以获得满意的效率受到越来越多的重视.讨论了一类采用数据空间网格划分的基于密度的聚类算法的均值近似方法.该方法过滤并释放位于稠密超方格中的数据项,并利用其重心点近似计算其对周围数据元素的影响因子.给出均值近似在聚类算法中的实现策略及其误差估计.均值近似方法在有效减少内存需求、大幅度降低计算复杂度的同时对聚类精确度影响甚微.

实验结果验证了该方法能够取得令人满意的效果.

基金项目：Supported by the National Natural Science Foundation of China under Grant No.79970092 (国家自然科学基金); the Natural Science Foundation of the Education Board of Jiangsu Province of China under Grant No.02KJB520012 (江苏省教育厅自然科学基金)

References:

[1] Han J, Kamber M. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, 2000. 335~398.

[2] Berchtold S, Keim D, Kriegel HP. The X-tree: An index structure for high-dimensional data. In: Proceedings of the International Conference on Very Large Databases. Bombay, India, 1996. 28~39.

[3] Hinneburg A, Keim DA. Optimal gird-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering. In: Proceedings of the 25th International Conference on Very Large Databases. Edinburgh, Scotland, 1999. 506~517.

[4] Sheikholeslami G, Chatterjee S, Zhang A. Wave-Cluster: A multi-resolution clustering approach for very large spatial databases. In: Proceedings of the 24th International Conference on Very Large Databases. New York, 1998. 428~439.

[5] Aggrawal R, Gehrke J, Gunopulos D, Raghawan P. Automatic subspace clustering of high dimensional data for data mining applications. In: Proceedings of the ACM SIGMOD International Conference on Management of Data. Seattle, WA, 1998. 94~105.

[6] Wang W, Yang J, Muntz R. STING: A statistical information grid approach to spatial data mining. In: Proceedings of the 23rd International Conference on Very Large Databases. Athens, Greece, 1997. 186~195.

[7] Hinneburg A, Keim DA. An efficient approach to clustering in large multimedia databases with noise. In: Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD'98). New York, 1998. 58~65.

[8] Xing EP, Karp RM. CLIFF: Clustering of high dimensional microarray data via iterative feature filtering using normalized cuts. BIOINFORMATICS, 2001,1(1):1~9.

[9] Hinneburg A, Keim DA, Brandt W. Clustering 3D-structures of small amino acid chains for detecting dependences from their sequential context in proteins. In: Proceedings of the IEEE International Symposium on BioInformatics and Biomedical Engineering. Washington, DC, 2000. 43~49.

[10] Xu X, Ester M, Kriegel H, Sander J. A distribution-based clustering algorithm for mining in large spatial databases. In: Proceedings of the 14th International Conference on Data Engineering, ICDE'98. Orlando, FL, 1998. 324~331.

[11] Silverman B. Density Estimation for Statistics and Data Analysis. Chapman & Hall, 1986. 72~113.