Libras 属性个数特别多的数据集上效率提升较少,仍需进一步加以改进.

References:

[1] Xu R, Wunsch D. Survey of clustering algorithms. IEEE Trans. on Neural Networks, 2005,16(3):645−678. [doi: 10.1109/TNN.2005.845141]

[2] Sun J, Liu J, Zhao LY. Clustering algorithms research. Ruan Jian Xue Bao/Journal of Software, 2008,19(1):48−61 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/48.htm [doi: 10.3724/SP.J.1001.2008.00048]

[3] Frey BJ, Dueck D. Clustering by passing messages between data points. Science, 2007,315(5814):972−976. [doi: 10.1126/science.1136800]

[4] Frey BJ, Dueck D. Response to comment on "clustering by passing messages between data point". Science, 2008,319(5864):726. [doi: 10.1126/science.1151268]

[5] Rodriguez A, Laio A. Clustering by fast search and find of density peaks. Science, 2014,344(6191):1492−1496. [doi: 10.1126/science.1242072]

[6] Arora P, Deepali D, Varshney S. Analysis of *K*-means and *K*-medoids algorithm for big data. Procedia Computer Science, 2016,78:507−512. [doi: 10.1016/j.procs.2016.02.095]

[7] Peker M. A decision support system to improve medical diagnosis using a combination of *k*-medoids clustering based attribute weighting and SVM. Journal of Medical Systems, 2016,40:116. [doi: 10.1007/s10916-016-0477-6]

[8] Kaufman L, Rousseeuw PJ. Finding Groups in Data: An Introduction to Cluster Analysis. New York: John Wiley and Sons, 1990.

[9] Ng RT, Han JW. Clarans: A method for clustering objects for spatial data mining. IEEE Trans. on Knowledge and Data Engineering, 2002,14(5):1003−1016. [doi: 10.1109/TKDE.2002.1033770]

[10] Barioni MCN, Razente HL, Traina AJM, Jr CT. Accelerating *K*-medoids-based algorithms through metric access methods. The Journal of Systems and Software, 2008,81:343−355. [doi: 10.1016/j.jss.2007.06.019]

[11] Park HS, Jun CH. A simple and fast algorithm for *K*-medoids clustering. Expert Systems with Applications, 2009,36:3336−3341. [doi: 10.1016/j.eswa.2008.01.039]

[12] Zadegan SMR, Mirzaie M, Sadoughi F. Randed *K*-medoids: A fast and accurate rank-based partitioning algorithm for clustering large datasets. Knowledge-Based Systems, 2013,39:133−143. [doi: 10.1016/j.knosys.2012.10.012]

[13] Kashef R, Kamel MS. Efficient bisecting *K*-medoids and its application in gene expression analysis. In: Proc. of the 5th Int'l Conf. on Image Analysis and Recognition. Varzim, 2008. 423−434. [doi: 10.1007/978-3-540-69812-8_42]

[14] Lai PS, Fu HC. Variance enhanced *K*-medoids clustering. Expert Systems with Applications, 2011,38:764−775. [doi: 10.1016/j.eswa.2010.07.030]

[15] Jiang YB, Zhang JM. Parallel *K*-medoids clustering algorithm based on hadoop. In: Proc. of the 5th IEEE Int'l Conf. on Software Engineering and Service Science (ICSESS). 2014. 649−652. [doi: 10.1109/ICSESS.2014.6933652]

[16] Yue J, Mao SJ, Li M, Zou XS. An efficient PAM spatial clustering algorithm based on MapReduce. In: Proc. of the 22nd Int'l Conf. on Geoinformatics. 2014. 1−6. [doi: 10.1109/GEOINFORMATICS.2014.6950803]

[17] Han LS, Xiang LS, Liu XY, Luan J. The *K*-medoids algorithm with initial centers optimized based on a *P* system. Journal of Information and Computational Science, 2014,11(6):1765−1773. [doi: 10.12733/jics20103217]

[18] Han LS, Xiang LS, Liu XY. *P* system based on the MapReduce for the most value problem. Journal of Information and Computational Science, 2014,11(13):4697−4706. [doi: 10.12733/jics20104502]

[19] Zhang Q, Couloigner I. A new and efficient *K*-medoids algorithm for spatial clustering. In: Proc. of the Int'l Conf. on Computational Science and Its Applications. LNCS 3482, Singapore: Springer-Verlag, 2005. 207−224. [doi: 10.1007/11424857_20]

[20] Chu SC, Roddick JF, Ran JS. An efficient *K*-medoids-based algorithm using previous medoid index, triangular inequality elimination criteria, and partial distance search. In: Proc. of the 4th Int'l Conf. on Data Warehousing and Knowledge Discovery, Vol.2454. Aixen-Provence, 2002. 63−72. [doi: 10.1007/3-540-46145-0_7]

[21] Chu SC, Roddick JF, Chen TY, Pan JS. Efficient search approaches for *K*-medoids-based algorithms. In: Proc. of the Int'l Conf. on Computers, Communications, Control and Power Engineering. 2002. 712a−715a. [doi: 10.1109/TENCON.2002.1181751]

[22] Chiang CS, Chu SC, Roddick JF, Pan JS. New search strategies and new derived inequality for efficient *K*-medoids-based algorithm. Chinese Journal of Electronics, 2007,16(1):82−87.

[23] Chu SC, Roddick JF, Pan JS. Improved search strategies and extensions to *K*-medoids-based clustering algorithms. Int'l Journal of Business Intelligence and Data Mining, 2008,3(2):212−231. [doi: 10.1504/IJBIDM.2008.020520]

[24] Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, 2012.

[25] Lichman M. UCI Machine Learning Repository. Irvine: University of California, School of Information and Computer Science, 2013. http://archive.ics.uci.edu/ml

附中文参考文献:

[2] 孙吉贵,刘杰,赵连宇.聚类算法综述.软件学报,2008,19(1):48−61. http://www.jos.org.cn/1000-9825/19/48.htm [doi: 10.3724/SP.J.1001.2008.00048]

附录

定理 3.2 的简要证明.

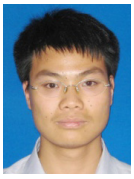
需要分 4 种情形.

1) $P\in C_i,O_i\in O\cap O'$,若 $dist(P,O_i)<\min_j\{dist(P,O'_j)\mid O'_j\in O'-O\}$,此时说明在所有的新交换中心点中,点 P 与新中心点的距离均大于 $dist(P,Q_i)$,因此,$P\in C_i$;

2) $P\in C_i,O_i\in O\cap O'$,若 $dist(P,O_i)<\min_j\{dist(P,O'_j)\mid O'_j\in O'-O\}$ 不成立,此时说明在所有的新交换中心点中,存在新中心点,使得点 P 与该新中心点的距离小于 $dist(P,Q_i)$,因此,P 需要重新分配到该新中心点(最小距离)所代表的簇中,即,满足 $dist(P,O'_l)=\min_j\{dist(P,O'_j)\mid O'_j\in O'-O\}$;

3) 当 $P\in C_i,O_i\notin O'$,说明点 P 所在的簇 i 的中心点已经更换,若 $dist(P,O_i)>\min_j\{dist(P,O'_j)\mid O'_j\in O'-O\}$,则说明存在新中心点,使得点 P 到该新中心点的距离要小于原始簇的距离 $dist(P,Q_i)$,此时需要 P 重新分配到该新中心点(最小距离)所代表的簇中,即,满足 $dist(P,O'_l)=\min_j\{dist(P,O'_j)\mid O'_j\in O'-O\}$;

4) 当 $P\in C_i,O_i\notin O'$,说明点 P 所在的簇 i 的中心点已经更换,若 $dist(P,O_i)>\min_j\{dist(P,O'_j)\mid O'_j\in O'-O\}$ 不成立,则说明所有新更换的中心点到 P 的距离均大于原始簇的距离 $dist(P,Q_i)$,此时需要比较所有中心点,以确定 P 所属的簇,即,满足 $dist(P,O'_l)=\min_j\{dist(P,O'_j)\mid O'_j\in O'\}$.



余冬华(1988－),男,江西赣州人,博士生,主要研究领域为机器学习,生物信息学.

任世军(1962－),男,博士,教授,主要研究领域为聚类分析,图论.

郭茂祖(1966－),男,博士,教授,博士生导师,CCF 专业会员,主要研究领域为机器学习,生物信息学.

刘晓燕(1963－),女,博士,副研究员,主要研究领域为生物信息学,数据挖掘.

刘扬(1976－),男,博士,副教授,CCF 专业会员,主要研究领域为机器学习,图像处理,计算机视觉.

刘国军(1979－),男,博士,讲师,CCF 专业会员,主要研究领域为机器学习,计算机视觉,模式识别.