P.O.Box 8718, Beijing 100080, China	Journal of Software, April 2007,18(4):905-918
E-mail: jos@iscas.ac.cn	ISSN 1000-9825, CODEN RUXUEW, CN 11-2560/TP
http://www.jos.org.cn	Copyright © 2007 by *Journal of Software*

基于滑动窗口的进化数据流聚类

常建龙, 曹 锋, 周傲英

常建龙, 曹 锋, 周傲英
(复旦大学 计算机科学与工程系,上海 200433)
作者简介: 常建龙(1972－),男,博士生,主要研究领域为数据流,数据挖掘.曹锋(1977－),男,上海崇明人,博士生,主要研究领域为数据流,数据挖掘.周傲英(1965－),男,博士,教授,博士生导师,CCF高级会员,主要研究领域为数据流,数据挖掘,XML数据管理,P2P计算.
联系人: 周傲英 Phn: +86-21-65643503, E-mail: ayzhou@fudan.edu.cn, http://www.fudan.edu.cn

Received 2005-12-22; Accepted 2006-06-09

Abstract
To address the sliding window based clustering, two types of exponential histogram of cluster features, false positive and false negative, are introduced in this paper. With these structures, a clustering algorithm based on sliding windows is proposed. The algorithm can precisely obtain the distribution of recent records with limited memory, thus it can produce the clustering result over sliding windows. Furthermore, it can be extended to deal with the clustering problem over N-n window (an extended model of the sliding window). The evolving data streams in the experiments include KDD-CUP'99 and KDD-CUP'98 real data sets and synthetic data sets with changing Gaussian distribution. Theoretical analysis and comprehensive experimental results demonstrate that the proposed method is of high quality, little memory and fast processing rate.

Chang JL, Cao F, Zhou AY. Clustering evolving data streams over sliding windows. *Journal of Software*, 2007,18(4):905-918.
DOI: 10.1360/jos180905
http://www.jos.org.cn/1000-9825/18/905.htm

摘要
提出了纳伪(false positive)和拒真(false negative)两种聚类特征指数直方图分别来支持纳伪误差和拒真误差窗口的聚类分析;然后,提出一种基于滑动窗口的数据流聚类方法.该方法在占用窗口大小的次线性内存空间前提下,及时保存最近数据记录的分布状况,从而实现对滑动窗口内的数据进行聚类.此外,它还可被扩展用于N-n窗口(滑动窗口的扩展模型)的数据聚类.实验采用KDD-CUP'99和KDD-CUP'98真实数据集以及变换高斯分布的人工数据集构造进化数据流.理论分析和实验结果表明,该方法具有良好的聚类质量、较小的内存开销和快速的数据处理能力.

基金项目: Supported by the National Natural Science Foundation of China under Grant Nos.60496325, 60496327 (国家自然科学基金)

References:

[1] Aggarwal CC, Han J, Wang J, Yu PS. A framework for clustering evolving data streams. In: Freytag JC, Lockemann PC, Abiteboul S, Carey MJ, Selinger PG, Heuer A, eds. Proc. of the Int'l Conf. on Very Large Data Bases. Berlin: Morgan Kaufmann Publishers, 2003. 81-92

[2] Chalaghan LO, Mishra N, Meyerson A, Guha S. Streaming data algorithms for high-quality clustering. In: Proc. of the 18th Int'l Conf. on Data Engineering. San Jose, 2002. 685-694. http://doi.ieeecomputersociety.org/10.1109/ICDE.2002.994785

[3] Domingos P, Hulten C. Mining high-speed data streams. In: Proc. of the KDD. 2000. http://citeseer.ist.psu.edu/domingos00mining.html

[4] Guha S, Meyerson A, Mishra N, Motwani R, Callaghan LO. Clustering data streams: Theory and practice. IEEE Trans. on Knowledge and Data Engineering, 2003,3(15):515-528.

[5] Guha S, Mishra N, Motwani R, Callaghan LO. Clustering data stream. In: Proc. of the 41st Annual Symp. on Foundations of Computer Science. Redondo Beach: IEEE Computer Society, 2000. 359-366.

[6] Nam H, Won S. Statistical grid-based clustering over data streams. SIGMOD Record, 2004,33(1):32-37.

[7] Ordonez C. Clustering binary data streams with k-means. In: Zaki MJ, Aggarwal CC, eds. Proc. of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD). San Diego, 2003. 12-19.

[8] Zhou A, Cai Z, Wei L, Qian W. M-Kernel merging: Towards density estimation over data streams. In: Proc. of the 8th Int'l Conf. on Database Systems for Advanced Applications (DASFAA). Kyoto, 2003. 285-292.

[9] Aggarwal CC, Han J, Wang J, Yu PS. A framework for projected clustering of high dimensional data streams. In: Nascimento MA, -zsu MT, Kossmann D, Miller RJ, Blakeley JA, Schiefer KB, eds. Proc. of the VLDB. Toronto: Morgan Kaufmann Publishers, 2004. 852-863.

[10] Babcock B, Datar M, Motwani R, Callaghan LO. Maintaining variance and k-medians over data stream windows. In: Proc. of the 22nd ACM SIGACT-SIGMOD-SIGART Symp. Principles of Database Systems. San Diego: ACM Press, 2003. 234-243.

[11] Dai B, Huang J, Yeh M, Chen M. Clustering on demand for multiple data streams. In: Proc. of the ICDM. Brighton, 2004. 367-370.

[12] Nasraoui O, Cardona C, Rojas C, Gonzlez F. TECNO-STREAMS: Tracking evolving clusters in noisy data streams with a scalable immune system learning model. In: Proc. of the 3rd IEEE Int'l Conf. on Data Mining (ICDM). Melbourne, 2003. 235-242.

[13] Yang J. Dynamic clustering of evolving streams with a single pass. In: Dayal U, Ramamritham K, Vijayaraman TM, eds. Proc. of the ICDE. Bangalore, 2003. 695-697.
[14] Han J, Kamber M. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, 2001.

[15] Charikar M, Callaghan LO, Panigrahy R. Better streaming algorithms for clustering problems. In: Proc. of the 35th Annual ACM Symp. on Theory of Computing. San Diego, 2003. 30-39.

[16] Beringer J, Hullermeier E. Online Clustering of Parallel Data Streams. Data & Knowledge Engineering. 2005. http://wwwiti.cs.uni-magdeburg.de/iti_dke/publications/DKE_draft.pdf

[17] Cao F, Ester M, Qian W, Zhou A. Density-Based clustering over an evolving data stream with noise. In: Proc. of the SIAM Conf. on Data Mining (SDM). 2006.

[18] Zhu WH, Yin J, Xie YH. Arbitrary shape cluster algorithm for clustering data stream. Journal of Software, 2006,17(3):379-387 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/17/379.htm

[19] Babcock B, Babu S, Datar M, Motwani R, Widom J. Models and issues in data stream systems. In: Popa L, ed. Proc. of the 21st ACM Symp. on Principles of Databases Systems (PODS). Madison, 2002. 1-16.

[20] Zhang T, Ramakrishnan R, Livny M. Birch: An efficient data clustering method for very large databases. In: Jagadish HV, Mumick IS, eds. Proc. of the SIGMOD. Montreal: ACM Press, 1996. 103-114.

[21] Datar M, Gionis A, Indyk P, Motwani R. Maintaining stream statistics over sliding windows. In: Proc. of the 13th Annual ACM-SIAM Symp. on Discrete Algorithms (SODA). San Francisco, 2002. 635-644. http://www-cs.stanford.edu/~datar/papers/sicomp_streams.pdf

附中文参考文献:
[18] 朱蔚恒,印鉴,谢益煌.基于数据流的任意形状聚类算法.软件学报,2006,17(3):379-387. http://www.jos.org.cn/1000-9825/17/379.htm