4.3.2 GCStream 算法在不同数据流速度下的聚类质量

能够快速聚类数据流,是数据流聚类算法的一个重要特性.因此,本文在 KDDCUP99 数据流上以不同的数据流速度(1*k*/s,2*k*/s,7*k*/s)验证本文算法聚类质量.聚类结果如图 10 所示.首先,本文算法能够在这 3 种速度下处理完数据流,说明 GCStream 算法有能力处理速度较快的数据流.然后,分析聚类质量评价指标结果可以得出,随着数据流速度的上升,CMM 指标值有所下降,但是下降幅度并不大;Purity 指标值下降幅度比 CMM 值略大,但仍保持在较高的水平.说明 GCStream 算法在聚类高速数据流时依然可以保存较高的聚类质量.

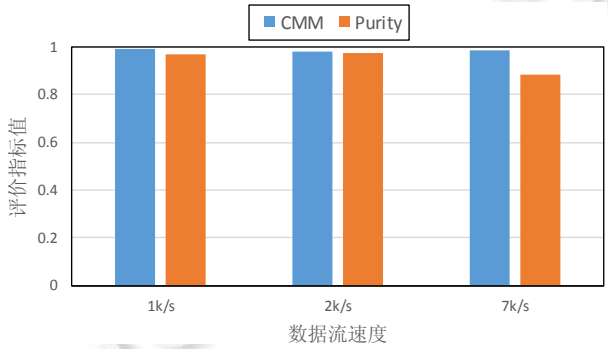


Fig.10 Cluster quality comparison under different stream rate

图 10 不同数据流速度下的聚类质量对比

4.3.3 GCStream 算法在不同网格边长下的聚类质量

本节实验主要测试不同网格边长对聚类质量的影响.以 KDDCUP99 为测试数据流,我们分别设置网格边长 *len*=40,100,120,160,其中,*len*=100 为本文整理数据集时发现的 KDDCUP99 数据集中簇之间的最小距离.聚类结果如图 11 所示.从图 11 可以看出,当网格边长大于 100 时,聚类结果的 CMM 值和 Purity 值随着网格边长的增加均有明显的下降.当网格边长小于 100 时,聚类质量总体相对稳定.实验结果说明:本文实验设置的网格边长 *len*=100 是比较准确的,并且 GCStream 算法聚类质量随着网格边长的增加而有所下降.

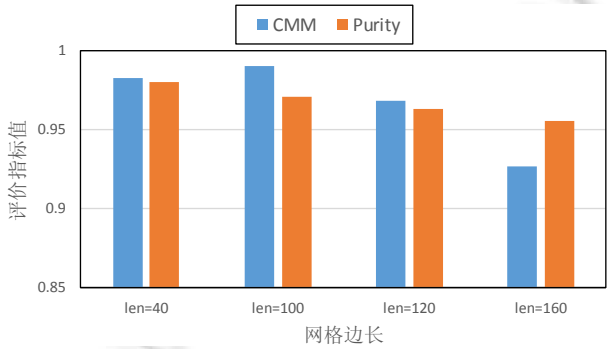


Fig.11 Cluster quality comparison under different grid sides

图 11 不同网格边长下的聚类质量对比

4.3.4 GCStream 算法捕捉簇的演变能力

数据聚类算法的一个重要特性是能够适应和捕捉簇的演变.为了验证 GCStream 算法的这两个特性,本文在人工数据集 MTD 上对 GCStream 算法进行了评估.在这个测试中,本文设置数据流到达速度为 1000*pt*/s,整个 MTD 数据流在 116s 内处理完.该数据集的分布如图 12 所示.图 12(a)~图 12(c)分别显示了 MTD 数据集中簇的生成顺序.其中,簇 1 和簇 2 中的数据是交叉分布的,在同一时刻,既有簇 1 中的数据到达也有簇 2 中的数据到达,

所以簇 1 和簇 2 能够同时生成.图 13 中显示了 GCStream 算法处理下的 MTD 数据分布.图 13(a)~图 13(d)分别显示了在 t=5,t=54,t=84,t=116 时刻生成的聚类结果.图中深颜色的区域代表当前时刻的生成的簇,浅蓝色的区域代表即将消失掉的簇.可以看出,GCStream 能够发现 4 个不同形状的簇并且不受噪声影响.图 14 显示了 MTD 数据流中簇的演变时刻.不同颜色的线条表示不同的簇,线条的长度表示簇存在的时间段.可以看到,簇 1 和簇 2 在初始时刻产生,在 55 时刻消失;簇 3 在 54 时刻产生,在 85 时刻消失;簇 4 在 84 时刻产生.除此之外,本文测得 GCStream 算法在人工数据集 MTD 的上的 Purity 均值为 0.983,CMM 均值为 1.说明 GCStream 算法具有较高的聚类质量.

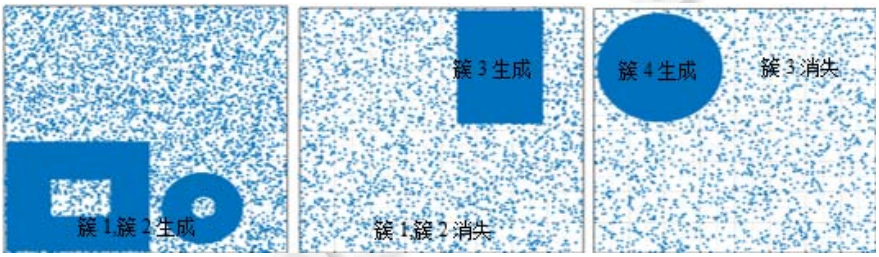
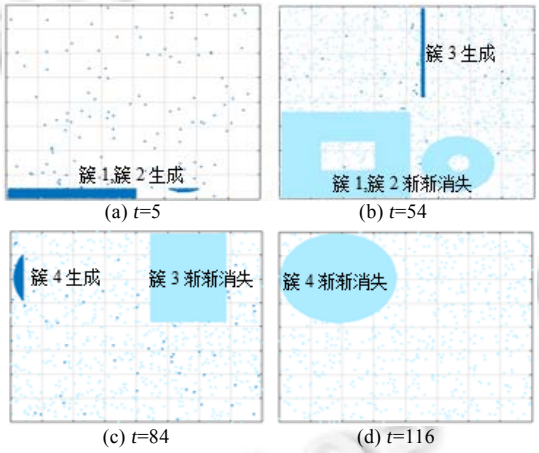


Fig.12 MTD data distribution

图 12 MTD 数据分布



(a) t=5 (b) t=54

(c) t=84 (d) t=116

Fig.13 Data distribution of MTD data set changes with time

图 13 MTD 数据集的数据分布随时间的变化

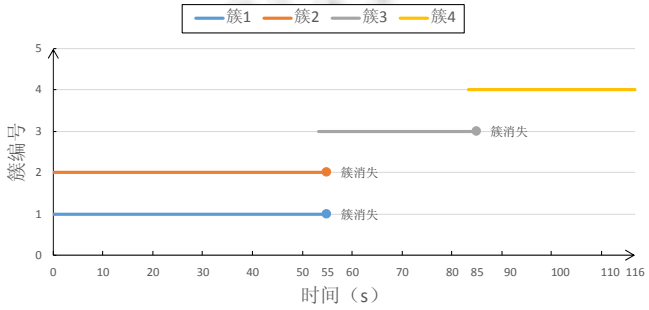


Fig.14 Evolution of clusters in MTD datasets

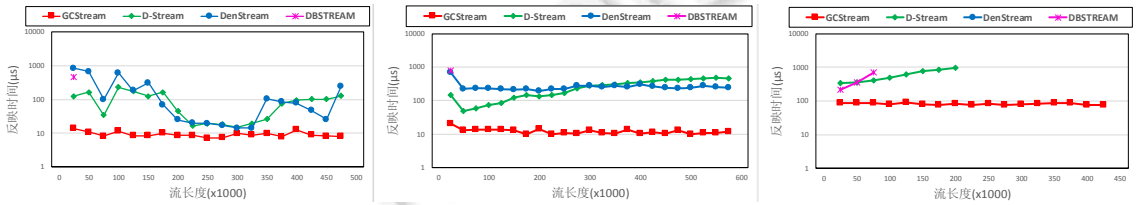
图 14 MTD 数据集中簇的演变

4.4 聚类效率评价

实时更新聚类结果对于数据流聚类算法至关重要.本文分别在多个数据集和不同维度上对各算法的效率进行了对比.

4.4.1 GCStream 算法在不同数据集上的效率

本节在 3 个 UCI 数据流上测试了 GCStream 与对比算法的聚类效率.设置数据流到达速率为 1000*pt*/s,并且每隔 25s 显示一次聚类结果.如果各算法能够在 25s 内处理完这段时间内到达的数据,则证明该算法能够正常运行;否则,说明该算法的效率不足以处理 1000*pt*/s 的数据流.图 15 显示了 25s 间隔内不同算法的响应时间对比.其中,DBSTREAM 算法在 3 个数据流上只在开始时正常运行,随后便运行失败;DenStream 和 D-Stream 算法在 PAMAP2 数据流上运行失败;而本文的 GCStream 算法能以 1000*pt*/s 的速度正常处理 3 个数据流并且所需时间最少,这说明 GCStream 算法效率比对比算法高.



(a) KDDCUP99 (b) CoverType (c) PAMAP2

Fig.15 Response time comparison on multiple datasets

图 15 多数据集上反映时间对比

4.4.2 网格边长与数据维度对 GCStream 算法效率的影响

本文在 MOAD 数据流上测试 GCStream 与对比算法在不同维度和不同网格边长上的聚类效率.图 16(a)、图 16(b)分别显示了网格边长 *len*=6 和 *len*=12.4 时,各算法在不同维度上平均效率.在不同大小的网格边长上比较可看出:随着网格边长的增加,GCStream,D-Stream 以及 DenStream 算法效率都有所提升.在不同数据维度上的算法效率比较可以看到:在数据维度小于 100 维时,GCStream 算法的效率是最高的;当数据维度大于 100 维时,GCStream 算法的效率也是比较高的,基本处于各算法效率的第 2 位.

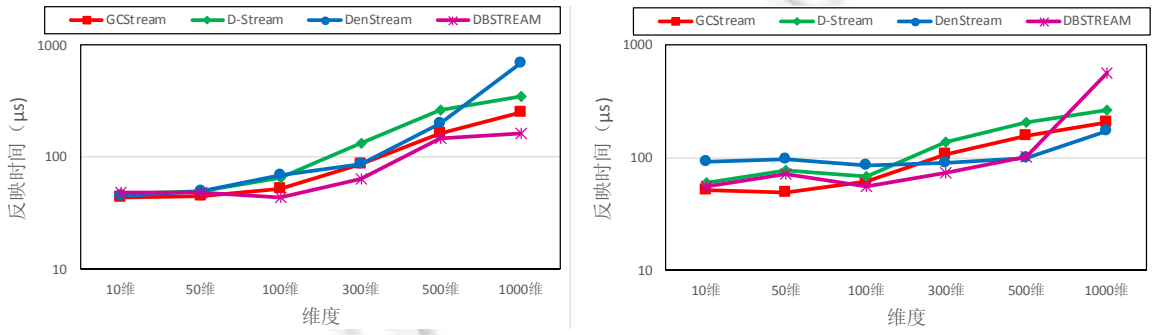


Fig.16 Response time comparisons in multiple dimensions and different grid lengths

图 16 不同网格边长和多维度上反映时间对比

5 结束语

本文针对现有数据流聚类算法在实时处理高速、大量的数据流时聚类效率和精度不高的问题,提出了一种基于网格耦合和核心网格的数据流聚类算法 GCStream.首先,通过网格耦合实现了对数据流更精确的汇总,提

高算法聚类质量;其次,本文根据数据流中局部权重较高的网格相比于局部权重较低的网格更可能为簇中心的特点引入了核心网格,然后以核心网格为簇中心生成簇,并且根据核心网格集合的变化来捕捉簇的演变;最后,通过真实数据集上进行实验,对比了本文所提方法与其他方法的聚类效果和聚类效率.实验结果表明,本文所提算法的聚类效果和聚类效率都优于对比方法.

由于本文算法的实验都是在网格边长相等的基础上进行的,没有考虑不同维度上的数据分布差异.所以本文的未来研究工作将着重研究根据不同维度上的数据分布采用不同的网格边长来使网格划分更精确,以进一步提高聚类质量.

References:

[1] Isaksson C, Dunham MH, Hahsler M. SOStream: Self organizing density-based clustering over data stream. In: Proc. of the Machine Learning and Data Mining in Pattern Recognition. Berlin, Heidelberg: Springer-Verlag, 2012. 264−278. [doi: 10.1007/978-3-642-31537-4_21]

[2] Silva JA, Faria ER, Barros RC, *et al*. Data stream clustering: A survey. ACM Computing Surveys, 2013,46(1):1−31.

[3] Zhang X, Furtlehner C, Germain-Renaud C, Sebag M. Data stream clustering with affinity propagation. IEEE Trans. on Knowledge and Data Engineering, 2014,26(7):1644−1656. [doi: 10.1109/TKDE.2013.146]

[4] Gong SF, Zhang YF, Yu G. Clustering stream data by exploring the evolution of density mountain. Proc. of the VLDB Endowment, 2017,11(4):393−405. [doi: 10.1145/3164135.3164136]

[5] Gama J, Žliobaitė I, Bifet A, Pechenizkiy M, Bouchachia A. A survey on concept drift adaptation. ACM Computing Surveys, 2014, 46(4):1−37. [doi: 10.1145/2523813]

[6] Masud M, Gao J, Khan L, Han J, Thuraisingham BM. Classification and novel class detection in concept-drifting data streams under time constraints. IEEE Trans. on Knowledge and Data Engineering, 2011,23(6):859−874. [doi: 10.1109/TKDE.2010.61]

[7] Aggarwal CC, Han J, Wang J, Yu PS. A framework for clustering evolving data streams. In: Proc. of the 29th Very Large Data Bases (VLDB) Conf. Berlin: VLDB Endowment. 2003. 81−92. [doi: 10.1016/B978-012722442-8/50016-1]

[8] Chen Y, Tu L. Density-based clustering for real-time stream data. In: Proc. of the 13th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2007. 133−142. [doi: 10.1145/1281192.1281210]

[9] Amini A, Saboohi H, Herawan T, Wah TY. MuDi-stream: A multi density clustering algorithm for evolving data stream. Journal of Network and Computer Applications, 2016,59(1):370−385. [doi: 10.1016/j.jnca.2014.11.007]

[10] Tu L, Chen Y. Stream data clustering based on grid density and attraction. ACM Trans. on Knowledge Discovery from Data, 2009, 3(3):1−27. [doi: 10.1145/1552303.1552305]

[11] Wan L, Ng WK, Dang XH, Yu PS, Zhang K. Density-based clustering of data streams at multiple resolutions. ACM Trans. on Knowledge Discovery from Data, 2009,3(3):1−28. [doi: 10.1145/1552303.1552307]

[12] Hahsler M, Bolaños M. Clustering data streams based on shared density between micro-clusters. IEEE Trans. on Knowledge and Data Engineering, 2016,28(6):1449−1461. [doi: 10.1109/TKDE.2016.2522412]

[13] Nguyen HL, Woon YK, Ng WK. A survey on data stream clustering and classification. Knowledge & Information Systems, 2015, 45(3):535−569. [doi: 10.1007/s10115-014-0808-1]

[14] O'callaghan L, Mishra N, Meyerson A, Guha S, Motwani R. Streaming-data algorithms for high-quality clustering. In: Proc. of the ICDE. 2002. 685−694. [doi: 10.1109/ICDE.2002.994785]

[15] Aggarwal CC, Han J, Wang J, Yu PS. A framework for projected clustering of high dimensional data streams. Proc. of the VLDB Endowment, 2004. 852−863. [doi: 10.1016/B978-012088469-8.50075-9]

[16] Cao F, Estert M, Qian W, Zhou A. Density-based clustering over an evolving data stream with noise. In: Proc. of the Siam Int'l Conf. on Data Mining. Bethesda, 2006. 328−339. [doi: 10.1137/1.9781611972764.29]

[17] Stolfo J, Fan W, Lee W, Prodromidis A, Chan PK. Cost-based modeling and evaluation for data mining with application to fraud and intrusion detection. In: Proc. of the Results from the JAM Project by Salvatore. 2000. 1−15.

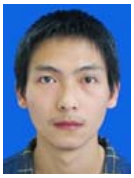
[18] Reiss A, Stricker D. Introducing a new benchmarked dataset for activity monitoring. In: Proc. of the Int'l Symp. on Wearable Computers. IEEE Computer Society, 2012. 108−109. [doi: 10.1109/ISWC.2012.13]

[19] Reiss A, Stricker D. Creating and benchmarking a new dataset for physical activity monitoring. In: Proc. of the Workshop on Affect & Behaviour Related Assistance. 2012. 1−8. [doi: 10.1145/2413097.2413148]

[20] Bifet A, Holmes G, Kirkby R, Pfahringer B. MOA: Massive online analysis. Journal of Machine Learning Research, 2010,11(2): 1601−1604.

[21] Kranen P, Kremer H, Jansen T, Seidl T, Bifet A, Holmes G, Pfahringer B. Clustering performance on evolving data streams: Assessing algorithms and evaluation measures within MOA. In: Proc. of the Int'l Conf. on Data Mining Workshops. 2010. 1400−1403. [doi: 10.1109/ICDMW.2010.17]

[22] Kremer H, Kranen P, Jansen T, Seidl T, Bifet A, Holmes G, Pfahringer B. An effective evaluation measure for clustering on evolving data streams. In: Proc. of the SIGKDD. San Diego, 2011. 868−876. [doi: 10.1145/2020408.2020555]



张东月(1993－),男,河北衡水人,硕士,主要研究领域为数据挖掘.

吴湘云(1964－),男,讲师,主要研究领域为微分方程,概率论与数理统计,数据分析.

周丽华(1968－),女,博士,教授,博士生导师,CCF 专业会员,主要研究领域为数据挖掘,社交网络分析.

赵丽红(1974－),女,讲师,主要研究领域为数据挖掘.