系统工程与电子技术 2009, 31(3) 671-676 DOI: ISSN: 1001-506X CN: 11-2422/TN

软件、算法与仿真

一种基于滑动窗口的多关系模式频度更新算法

侯伟[1], 杨炳儒[1], 吴晨生[2], 周谆[1]

1. 北京科技大学信息工程学院, 北京, 100083;
2. 北京市科学技术情报研究所, 北京, 100037

摘要： 面向多个相关数据流的挖掘算法研究尚处于起步阶段。作为多数据流挖掘算法的基础,模式频度更新算法仍然存在计数不准确、性能较低等问题,难以以此构造有效的挖掘算法。通过引入多关系挖掘概念以及目标关系定义,进而限定计数对象,提出了一种基于滑动窗口的多关系模式频度更新算法MRPFU。该算法监视各数据流窗口的更新情况,采用计数传播策略,减少了时间与空间复杂度。理论分析及实验结果证明了所提算法的有效性且具有较高性能。

关键词： 数据挖掘 数据流 滑动窗口 多关系数据挖掘 频度更新

Multi-relational pattern frequency update algorithm based on sliding window

HOU Wei[1], YANG Bing-ru[1], WU Chen-sheng[2], ZHOU Zhun[1]

1. School of Information Engineering, Univ. of Science and Technology Beijing, Beijing 100083, China;
2. Beijing Municipal Inst. of Science and Technology Information, Beijing 100037, China

Abstract: Presently,the study of mining algorithms for multiple correlated data streams is still at a primitive stage.As the basis of mining multiple data streams,the methods of updating the frequencies of patterns,are bearing problems of count deviation,low performances etc.Consequently,efficient mining algorithms are difficult to be built either.The concepts of multi-relational data mining and target relation are introduced firstly,and the count object is defined accordingly.Then an algorithm based on sliding windows for updating frequencies of multi-relational patterns is proposed,which monitors the updates of streams,adopts the strategy of count propagation,and relieves the complexity of runtime and space.The theoretical analysis and experiments prove its effectiveness and performance.

Keywords: data mining data stream sliding window multi-relational data mining frequency update

收稿日期 2008-02-01 修回日期 2008-07-05 网络版发布日期

DOI:

基金项目:

国家自然科学基金资助课题(60675030)

通讯作者:

作者简介: 侯伟(1981-),男,博士研究生,主要研究方向为数据挖掘.E-mail:dr.houwei@gmail.com

作者Email:

参考文献：

[1] Han J,Kamber M.Data mining;concepts and techniques[M].San Francisco:Morgan Kaufmann,2000.
[2] Babcock B,Babu S,Datar M,et al.Models and issues in data stream systems[C].Proc.of the 21th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems.Wisconsin:Madison,2002:1-16.
[3] Gaber M M,Zaslavsky A,Krishnaswamy S.Mining data streams:A review[J].ACM SIGMOD Record,2005,34(2):18-26.
[4] Garofalakis M N,Gehrke J,Rastogi R.Querying and mining data streams:you only get one look

[C].Proc.of the 28th Int.Conf.on Very Large Data Bases,Hong Kong,2002.
[5] Kontaki M,Papadopoulos A N,Manolopoulos Y.Effident incremental subspace clustering in data streams[C].Proc.of Datobase Engineering and Applications Symposium,Delhi,India,2006.
[6] Commnde G,Muthukrishnan S.What's new:Finding significant differences in network data streams [C].Proc.of IEEE INFOCOM,Hang Koag,2004.
[7] Chen R,Sivakumar K,Kargupta H.Distributed web mining using bayesian networks from multiple data streams[C].Proc.of ICDM',San Jose,California,2001.
[8] Xu Y,Wang K,Fu A,et al.Classification spanning correlated data streams[C].Proc.of CIKM',Arlington,Virginia,2006.
[9] Dzeroski S,Lavrac N.Relational data mining[M].Berlin:Springer,2001.
[10] Terry D,Goldberg D,Nichols D,et al.Continuous queries over append-only databases[C].proc.of ICMD,1992:321-330.
[11] Golab L,(o)zsu M T.Processing sliding window multi-joins in continuous queries over data streams [C].Proc.of the 29th VLDB Conf.,Berlin,Germany,2003.
[12] Srivastava U,Widom J.Memory-limited execution of windowed stream joins[C].Proc.of the 30th VLDB Conf.,Toronto,Canada,2004.
[13] Gedik B,Wu K,Yu P S,et al.GRUBJOIN:An adaptive multiway windowed stream join with time correlation-aware CPU load shedding[J].Transactions on Knowledge and Data Engineering,2007,19 (10):1363-1380.
[14] 钱江渡,徐宏炳,董逸生,等.基于最小生成树的数据流窗口连接优化算法[J].计算机研究与发展,2007,44 (6):1000-1007.
[15] Yin X,Han J,Yang J,et al.CrossMine:efficient classification across multiple database relations [C].Proc.of the 20th International Conference on Data Engineering,Boston,2004.
[16] Atramentov A,Leiva H,Honavar V.A multi-relational decision tree learning algorithm-implementation and experiments[C].Proc.of ILP,Szeged,Hungary,2003.
[17] Sourc forge research data[OL].http://www.nd.edu/～oss/Data/data.html,2002.

本刊中的类似文章

1. 孙宇航，孙应飞.基于网络日志的数据挖掘预处理改进方法[J]. 系统工程与电子技术, 2009,31(12): 2994-2997
2. 曾华，吴耀华，黄顺亮.非均匀类簇密度聚类的多粒度自学习算法[J]. 系统工程与电子技术, 2010,32(8): 1760-1765
3. 唐志刚, 杨炳儒, 杨珺.一种基于马尔可夫链的高维离群点挖掘算法[J]. 系统工程与电子技术, 2010,32(12): 2721-2724
4. 翟云,杨炳儒,曲武,隋海峰.基于新型集成分类器的非平衡数据分类关键问题研究[J]. 系统工程与电子技术, 2011,33(1): 196-0201
5. 石崇林, 张茂军, 吴琳, 唐宇波, 景民.基于密度的计算机兵棋推演数据快速聚类算法[J]. 系统工程与电子技术, 2011,33(11): 2428-2433
6. 赵静娴, 倪春鹏, 詹原瑞, 杜子平.一种高效的连续属性离散化算法[J]. 系统工程与电子技术, 2009,31(01): 195-199
7. 刘波, 杨路明, 邓云龙.一种自适应的XML数据流质量预测策略[J]. 系统工程与电子技术, 2009,31(6): 1485-1491
8. 程春玲, 余志虎, 张登银, 徐小龙.基于SSC-tree流聚类的入侵检测算法[J]. 系统工程与电子技术, 2012,34(3): 625-630
9. 方炜炜, 杨炳儒, 夏红科.基于SMC的隐私保护聚类模型[J]. 系统工程与电子技术, 2012,34(7): 1505-1510
10. 张毅, 张强, 王宇, 欧阳志新.抑制SAR图像相干斑的自适应红黑窗滤波算法[J]. 系统工程与电子技术, 2012,34 (8): 1576-1580
11. 赵静娴, 倪春鹏, 詹原瑞, 杜子平.一种高效的连续属性离散化算法[J]. 系统工程与电子技术, 2009,31(1): 195-199
12. 赵静娴, 倪春鹏, 詹原瑞, 杜子平.一种大规模数据库的组合优化决策树算法[J]. 系统工程与电子技术, 2009,31 (3): 583-587
13. 罗来鹏, 刘二根, 曾毅.粗糙集理论研究的矩阵方法[J]. 系统工程与电子技术, 2009,31(4): 859-862
14. 敖富江, 杜静, 颜跃进, 黄柯棣.在线挖掘数据流滑动窗口中频繁闭项集[J]. 系统工程与电子技术, 2009,31(5): 1235-1240
15. 吕宗磊, 王建东, 徐涛.基于模态代表点的聚类评价方法 [J]. 系统工程与电子技术, 2009,31(8): 1997-2002