数据流挖掘技术及其在仿真中的应用

敖富江[1] 戚宗锋[2] 陈 彬[1] 黄柯棣[1]
(国防科技大学机电工程与自动化学院 长沙 410073)[1] (中国人民解放军63892部队 洛阳 471003)[2]

摘 要 随着仿真系统复杂程度的增加和规模的增大,仿真时间越来越长,仿真所产生的数据量越来越大,使得仿真数据具有数据流的特性,因此可以采用数据流挖掘技术处理仿真数据。综述了数据流和数据流挖掘技术的主要特点;提出了基于数据流挖掘技术的仿真应用框架;设计了通用数据流挖掘成员,以便能够快速将数据流挖掘算法集成到基于HLA体系结构的仿真系统中,并以导弹突防仿真系统为例介绍了所设计的通用数据流关联规则挖掘成员。

关键词 数据流挖掘,仿真,通用数据流挖掘成员,关联规则
中图法分类号 TP391.9 **文献标识码** A

Data Streams Mining Techniques and its Application in Simulation System

AO Fu-jiang[1] QI Zong-feng[2] CHEN Bin[1] HUANG Ke-di[1]
(College of Mechanical Engineering and Automation, National University of Defense Technology, Changsha 410073, China)[1]
(Troops 63892, Chinese People's Liberation Army, Luoyang 471003, China)[2]

Abstract With increasing of complexity and augmenting of scale of simulation system, the simulation time becomes much longer and the simulation data becomes much huger. These make the simulation data has the characters of data streams, so that the simulation data can be processed by data streams mining techniques. We summarized the characters of data streams and data streams mining techniques, and presented simulation application framework based on data streams mining techniques, and designed general data-streams-mining federate to make it easy to integrate the various data streams mining algorithms in HLA-architecture-based simulation system quickly, and introduce our general federate for mining association rule in data streams with the example of Missile-Breakthrough simulation system.

Keywords Data streams mining, Simulation, General data-streams-mining federate, Association rule

数据挖掘技术在仿真结果评估中发挥着重要作用,例如文献[1-4]中分别运用数据挖掘技术处理仿真数据。通常的做法是将仿真所产生的数据存储到永久存储器中。仿真结束后,再利用各种数据挖掘算法分析所存储的数据。但随着仿真系统复杂程度的增加和规模的增大,仿真所产生的数据可能非常巨大,存储这些数据的代价高。例如,在美国综合战区作战(synthetic theater of war, STOW)研究中,仿真48h产生了0.3TB~0.5TB的数据[5]。另外,仿真时间可能会变得越来越长,因此有必要在仿真的过程中实时从已有的数据中挖掘出用户感兴趣的模式,即形成在线评估,以便必要时对仿真系统进行调整。这要求数据挖掘算法能够实时、在线地处理连续到达的数据,并能够实时或在用户要求时快速返回当前的结果。并且由于数据量巨大,可能无法存储所有数据,要求算法只能单遍访问数据。而传统的针对静态数据集的挖掘算法通常需要保存所有数据,挖掘时多遍访问数据,因此不适合在线挖掘。

数据流挖掘技术能够较好地解决这些问题。简单地说,数据流是一种实时、有序、无限、高速到达的数据序列。仿真系统所产生的数据具有数据流的这些特点,因此可以作为数据流进行处理。例如文献[6]中将科学计算仿真数据作为数据流进行处理,并建立了数据流近似查询系统。将数据流挖掘算法嵌套到仿真系统中,能够降低存储与移动数据的开销,并使用户能够尽快了解当前仿真阶段所能提供的知识。

1 数据流概述

1.1 数据流的特点

数据流是近几年来国内外广泛研究的一个热点。它存在于多种领域,包括传感器网络、电信通话记录、气象监测与分析、股票分析、邮件过滤、网络监控与安全、Web日志分析,以及大规模科学计算的数据分析等方面。数据流具有连续性、高速性、数据量无限性、数据内容随时间改变等特征。与传统的、以关系模型存储的数据集相比,数据流具有以下特点[7]:(1)数据流中的数据元素以在线方式到达;(2)系统处理数据流时,无法控制数据元素的顺序;(3)数据流数据是无限的;(4)数据流的元素一旦被处理后,将被抛弃或者被存档——无法再容易地获取它,除非它被显式地存储在内存中。

到稿日期:2008-04-17 本文受国家自然科学基金资助项目(60573057,60704038)资助。
敖富江(1975—),男,博士生,研究方向为数据仓库、数据挖掘、复杂系统仿真等,E-mail:aofj2001@yahoo.com.cn;戚宗锋(1973—),男,高工,研究方向为雷达对抗建模与仿真;陈 彬(1981—),男,博士生,研究方向为建模方法、分布式仿真;黄柯棣(1940—),男,教授,博导,研究方向为系统仿真、控制理论与控制工程等。

对于复杂大系统仿真来说，所产生的数据具有数据流的特征。首先，它是一种连续、高速到达的数据序列；其次，对于某些仿真系统，所产生的数据量巨大，接近于无限；最后，随着时间的推进，复杂系统仿真所产生的数据的特征和数据间的关系有可能发生改变。因此，可以也有必要利用数据流挖掘技术来处理仿真数据。

1.2 数据流挖掘

相对于挖掘静态数据集，挖掘数据流更具有挑战性：第一，数据流的每个元素最多只能被检索一次，因此要求挖掘算法必须是单遍算法；第二，数据流数据是无限的，但挖掘算法所使用的内存是有限的；第三，数据流的高速性要求数据流的每一个元素应当被尽快地处理；第四，当用户请求挖掘的结果时，应当及时地响应，甚至做到在任意时刻都为用户提供对当前数据的分析结果。

图 1 给出了数据流挖掘模型。当数据流高速到达时，首先由预处理模块进行处理，结果存储于缓冲区中。预处理模块主要完成数据的变换。并且当数据流速度太快，以至于后续的挖掘引擎无法及时处理时，它还需要借助于采样等技术进行近似处理。存储器内概述数据结构通常以压缩的方式保存部分数据流数据和挖掘的结果。数据流挖掘引擎负责处理查询请求和数据的挖掘。查询请求又分为单次查询和连续查询。对于单次查询，用户提出查询请求后，只需要将当前满足查询条件的结果输出，即一次查询对应一次结果输出；对于连续查询，当用户提出查询请求后，挖掘引擎不但需要将当前满足查询条件的结果输出，而且当将来具有满足查询条件的结果时，也需要输出，即一次查询对应多次输出。结果显示模块以用户易理解的形式输出挖掘结果。

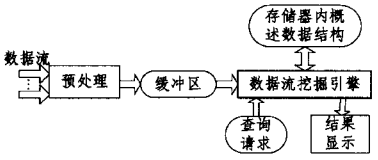

图 1 数据流挖掘模型

数据流挖掘算法通常针对特定的窗口模型。存在 3 种窗口模型：界标窗口模型(Landmark Windows)、滑动窗口模型(Sliding Windows)和衰减窗口模型(Damped Windows)。采用界标窗口模型的算法挖掘从一个特定的时间点到当前时间的所有数据，该特定时间点称为“界标”。大量的早期算法基于这种模型。当人们仅对数据流的最近信息感兴趣时，这种模型不太适合。而滑动窗口模型适用于对当前数据感兴趣的应用。在滑动窗口模型中，算法时刻维护当前时刻之前 w 个数据记录的挖掘结果。在衰减窗口模型中，每一个数据记录具有一个权重，当数据记录“衰老”时，它的权重随之降低。

另外，由于数据流数据的无限性，数据流挖掘算法通常又分为近似算法和精确算法。大多数数据流挖掘算法都是近似算法。精确算法通常仅适合于滑动窗口模型。

在挖掘内容方面，数据流挖掘算法主要集中于挖掘关联规则、分类和聚类等。挖掘关联规则的主要任务是挖掘频繁模式。传统的针对静态数据集的频繁模式挖掘算法通常需要多遍扫描数据集，因此无法被直接应用于数据流挖掘。数据流关联规则挖掘主要集中于研究单遍算法，以及能够在线更新挖掘的算法。对于数据流分类来说，训练样本通常无法一次全部获取，因此要求算法具有增量学习分类器的功能。另外，数据流的概念漂移特性对分类来说是一种极大的挑战，一成不变的分类器通常无法获得很好的分类精度。存在两种处理概念漂移的方式：一种是利用新数据更新分类器；另外一种是采用多种分类器组合的方法进行分类。同前两类算法一样，数据流聚类算法也主要集中于研究单遍算法、增量式算法，以及能够处理概念漂移问题的算法等方面。

由于窗口模型、是否近似算法、挖掘内容等方面的不同，数据流挖掘算法种类繁多。应用到仿真系统时，需要根据具体情况进行选择。

2 基于数据流挖掘的仿真应用框架

对于大型仿真系统来说，在线评估和事后评估同等重要。在线评估能够为用户提供及时的信息；事后评估能够为用户提供详细的信息。数据流挖掘算法在这两个阶段均能发挥作用。另外，某些特定的仿真成员也需要数据流挖掘算法提供决策能力，例如在双机格斗仿真中，需要利用分类算法选择下一步的动作。

因此，在仿真系统中数据流挖掘算法可以应用于以下 3 个方面：(1)以数据流挖掘成员的形式嵌入到仿真系统中(即在线评估)，搜集系统各联邦成员产生的数据，挖掘出的结果供领域专家参考，以便适当的时候调整仿真系统，或者挖掘出的结果直接输入仿真系统，调整仿真系统或联邦成员的参数，直接干涉仿真；(2)用于处理最终收集到的仿真结果数据(即事后评估)，相对于传统的数据挖掘算法，数据流挖掘算法在海量数据处理方面更具有优势；(3)直接嵌入仿真的联邦成员中，为成员提供决策能力。基于这些方面，我们提出了基于数据流挖掘的仿真应用框架，如图 2 所示。

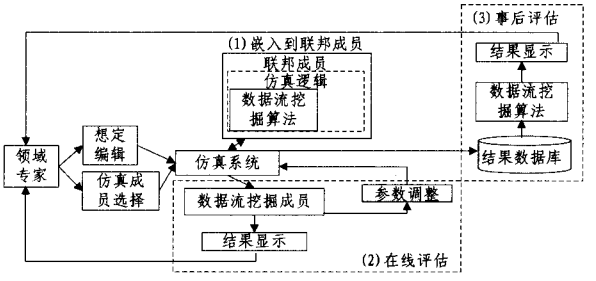

图 2 基于数据流挖掘的仿真系统应用框架

3 通用数据流挖掘成员

3.1 节介绍通用数据流挖掘成员的基本结构，3.2 节根据具体实例介绍通用数据流关联规则挖掘成员。

3.1 基本结构

根据 1.2 节的介绍，存在多种数据流挖掘算法。如何快速地将它们加入到基于 HLA 体系结构的仿真系统中，是一个需要考虑的问题。传统的 HLA 仿真成员开发方法是：根据具体的应用依次建立公布/定购关系，编写仿真逻辑，编写显示界面等代码。当公布/定购的数据发生变化，或者仿真逻辑发生变化时，只能更改源代码，并且有时需要改动大量的代码，甚至需要重新编写。

对于不同的应用，数据流挖掘成员挖掘的数据内容是不同的(公布/定购的数据不同)，并且成员所采用的数据流挖掘算法也是不同的(仿真逻辑不同)。如果采用传统的开发方

法，显然针对不同的应用，需要开发不同的成员，代码的重用性低，工作量巨大。为此，我们基于 HLA 仿真系统，开发了通用数据流挖掘成员，其架构如图 3 所示。

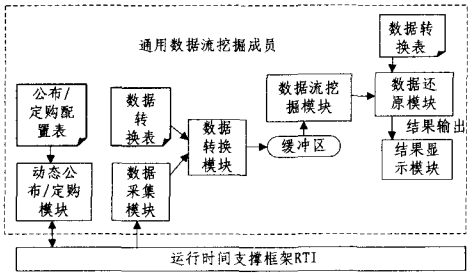

图 3　通用数据流挖掘成员

其中公布/定购配置表、数据转换表是以 XML 形式存储的配置文件，根据具体的应用进行不同的配置。动态公布/定购模块根据公布/定购配置表中描述的公布/定购对象类或交互类，确定成员的公布/定购关系（具体实现可参考文献[8]）。仿真开始后，数据采集模块根据定购关系获得仿真系统中的数据。但所获得的数据可能是枚举型、整数型或者浮点类型等。而数据流挖掘算法通常采用元组$\langle i_1, i_2, \cdots, i_m \rangle$（其中 i_k 为正整数）的形式作为输入，因此必须将数据采集模块所采集的数据进行转换。数据转换模块根据数据转换表中描述的映射规则将数据转换为数据流挖掘算法所需要的形式。而数据流挖掘模块是以动态链接库形式存在的算法，根据不同的应用可以选择不同的动态链接库。数据流挖掘模块所挖掘出的规则通常无法直接被理解，因此需要采用数据还原模块进行还原，它是数据转换模块的逆过程。不同类型的数据流挖掘算法所挖掘出的规则形式不同，因此需要采用不同的结果显示模块。结果显示模块也以动态链接库的形式存在。

3.2　通用数据流关联规则挖掘成员

关联规则挖掘寻找数据集中属性之间的有趣联系，是一类非常重要的数据挖掘应用。在仿真中经常用到的该类算法，例如文献[2]在坦克系统仿真中挖掘了天气、地形等要素与坦克命中率之间的关联规则，文献[3]在房屋设计仿真中挖掘了风速、风向与空气改变率之间的关联规则。挖掘关联规则的实质是挖掘数据中的频繁模式。频繁模式又可以细分为频繁项集、频繁闭项集、最大频繁项集、Top-K 频繁项集、基于约束的频繁项集等，相应的数据流挖掘算法已被广泛研究。每一类频繁模式挖掘算法适用于不同的范围。因此，有必要建立方便更换算法的通用关联规则挖掘成员。下面基于“导弹突防仿真系统”例子介绍我们所设计的通用数据流关联规则挖掘成员。

为了探索某型导弹的突防能力，我们建立了导弹突防仿真系统。该系统由突防导弹成员、电磁环境成员、雷达成员、干扰成员、拦截导弹成员、通用关联规则成员等组成，其成员组成如图 4 所示。

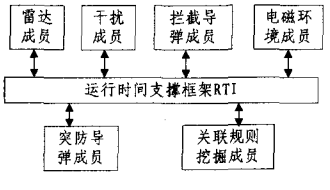

图 4　导弹突防仿真系统成员组成

通用关联规则成员用于分析哪些因素是影响导弹突防成功的主要因素，其结构如图 3 所示。首先建立条件属性集 C 为{天气、突防弹弹道、干扰方式、诱饵弹}，决策属性集 D 为{突防成功与否}。每一种属性的取值范围及其映射方式如表 1（数据转换表）所示。通用关联规则成员的“公布/定购配置表”中包含 C 和 D 中的所有属性。

表 1　数据转换表

属性	取值范围	映射
天气	晴	1
	阴	2
	雨	3
突防导弹弹道	弹道 1	4
	弹道 2	5
	弹道 3	6
干扰方式	无干扰	7
	弹载干扰	8
	远距离压制	9
	远距离多假目标	10
诱饵弹	无诱饵弹	11
	1 个诱饵弹	12
	多个诱饵弹	13
突防成功与否	成功	14
	失败	15

根据每种属性取值的不同，突防试验战情数目为 3×3×4×3=108。我们分别对每一种战情进行 100 次蒙特卡洛仿真，即共 10800 次仿真。每次仿真的条件属性和决策属性的值组成关联规则挖掘的一条事务（Transaction）数据。通用关联规则成员获得该事务后，通过数据转换模块将其转换为数据流挖掘模块所能够识别的形式。例如，某次仿真的数据为{晴，弹道 1，弹载干扰，1 个诱饵弹，成功}，通过数据转换后获得的事务为{1,4,8,12,14}。连续不断的事务组成事务数据流，例如{{1,4,8,12,14},{1,5,8,12,15},{1,5,9,12,14},{1,6,10,12,14} …}。

为了获得完全的支持度信息，并使得关联规则的数目尽可能地少，我们选择挖掘频繁闭项集。在数据流频繁闭项集挖掘方面，文献[9]中提出了一种能够在任意时刻都维护数据流滑动窗口中频繁闭项集的算法，Moment。在文献[10]中我们基于 FP-Tree 实现了具有相同功能但时空效率更优的算法 FPCFI-DS。这里的数据流挖掘模块采用 FPCFI-DS 算法。为了获得整个过程的关联规则，FPCFI-DS 算法的窗口大小设置为整个数据集，最小支持度设置为 1（即挖掘所有频繁闭项集）。

表 2　部分关联规则结果

关联规则	支持度
雨 => 成功	3168
弹道 1 => 成功	3021
多个诱饵弹 => 成功	2809
远距离多假目标 => 成功	2955
雨∩弹道 1∩无诱饵弹 => 成功	2767

所有仿真试验执行完成后，通用关联规则成员能够立即输出所有的频繁闭项集，然后通过数据还原模块获得用户所能够识别的模式。由于试验目的是为了发现哪些因素是影响导弹突防成功的主要因素，因此这里仅选取包含“成功”并且

（下转第 133 页）

念层次的深宽比变大，且结构比较平衡对称时，标记矩阵会将已确定的概念包含在概念层次中，使其传播得更远，从而确定其它更多的概念包含，对一些概念层次分支进行有效的剪枝，从而减少包含测试调用的次数，提升分类性能。

结束语 通过上面的描述，我们给出了知识库术语分类算法的相关优化技术的轮廓。一些分类之前的预处理优化技术，如上面提及的 lazy unfolding，absorption 等优化技术，可以对知识库中的复杂公理进行处理，使得知识库中的公理结构上更加简单，从而使 DL 推理(包含测试)的效率大大提高，并为后续的概念分类提供有利的支持。事实上，这两种优化技术只是保证了 tableau 算法能快速地找出概念包含 $D \sqsubseteq C$ 或者概念 $D \sqcap \neg C$ 不满足。而检测概念不包含($D \not\sqsubseteq C$)或概念满足更加困难和复杂，另外已提出一些优化技术用来解决此类问题，如 semantic branching search，dependency-directed backtracking，caching[3,4,8]等。其中部分已经在诸如 Pellet，Fact 推理机中得以实现。本文随后给出了知识库概念分类算法的完整流程，在此基础上，提出了一种动态标记矩阵的优化技术，并在理论和实验上证明了它能提升分类算法的性能。该优化技术在逻辑上是独立的，因此能很灵活地同其它优化技术融合起来。当前，知识表达在人工智能的实际应用方面扮演了很重要的角色，但领域知识经过诸如本体这样通用知识表达规范表达之后，知识层次并不完整并且需要进一步挖掘知识中隐含的未知信息[10]。本文描述的相关分类算法和优化技术正是用来解决此类问题的。知识表达同其它的智能应用的融合是一大发展趋势，未来我们会重点关注如何将领域知识表应用于信息集成、信息检索。

参 考 文 献

[1] Tsarkov D, Horrocks I. Optimised Classification for Taxonomic Knowledge Bases // Proceedings of the 2005 International Description Logic Workshop (DL 2005). Edinburgh, Scotland, UK, 2005, 147

[2] Baader F, Franconi E, Hollunder B, et al. An empirical analysis of optimization techniques for terminological representation systems // Proc. of the 3rd Int. Conf. on the Principles of Knowledge Representation and Reasoning (KR92). Cambridge, MA: Morgan-Kaufmann, 1992: 270-281

[3] Baader F, Calvanese D, McGuinness D, et al. The Description Logic Handbook: Theory, Implementation and Applications. Cambridge. UK: Cambridge University Press, 2003: 47-100, 313-355

[4] Horrocks I, Patel - Schneider F. Optimizing Description Logic Subsumption. Journal of Logic and Computation, 1999, 9(3): 267-293

[5] Horrocks I, Tobies S. Reasoning with Axioms: Theory and Practice // Proc. of the 7th Int. Conf. on Principles of Knowledge Representation and Reasoning (KR 2000). Breckenridge, Colorado, USA, 2000: 285-296

[6] Zuo M, Haarslev V. High Performance Absorption Algorithms for Terminological Reasoning // Proc. of the 2006 Int. Description Logic Workshop (DL 2006). Lake District, UK, 2006: 159-166

[7] Horrocks I, Tobies S. Optimisation of Terminological Reasoning // Proc. of the 2000 Description Logic Workshop (DL 2000). Aachen, Germany, 2000: 183-192

[8] Horrocks I. Using an Expressive Description Logic: FaCT or Fiction? // Proc. of the 6th Int. Conf. on Principles of Knowledge Representation and Reasoning (KR 98). Trento, Italy, 1998: 636-647

[9] 陆建江，张亚非，苗壮，等. 语义网原理与技术. 科学出版社，2007

[10] 邓志鸿，唐世，张铭，等. Ontology 研究综述[J]. 北京大学学报：自然科学版，2002，38(5)：730-737

(上接第 118 页)

支持度最高的 n 个(n 值根据具体情况选取)频繁闭项集作为分析的基础，并将其转换为关联规则。表 2 中列举了其中部分关联规则及它们的支持度(数据仅为举例，不代表任何真实系统)。

通过对关联规则的分析，我们得出一系列感兴趣的知识。例如，雨天时突防成功的可能性更高、采用弹道 1 突防成功的可能性更高等。

结束语 数据流挖掘技术能够为用户及时地提供数据中感兴趣的模式。而大规模复杂仿真系统所产生的数据具有数据流的特点。本文提出了基于数据流挖掘技术的仿真应用框架，设计了通用数据流挖掘成员，并以具体实例介绍了所设计的通用数据流关联规则挖掘成员。

数据流技术主要包括数据流管理技术和数据流挖掘技术。我们认为都可以将它们运用到仿真中。利用数据流管理系统管理仿真数据，以便为用户提供方便快捷的查询，可以作为进一步的研究方向。

参 考 文 献

[1] 鞠儒生，黄柯棣. 基于数据挖掘的 HLA 仿真系统测试与评估[J]. 系统工程与电子技术，2006，28(10)：1599-1602

[2] 张文明，薛青. 粗糙集方法在作战仿真数据挖掘中的应用. 系统仿真学报，2006，18(2)：179-181

[3] Morbitzer C, Strachan P, Simpson C. Application of Data Mining Techniques for Building Simulation Performance Prediction Analysis [C] // Proc. of Eighth IBPSA. Eindhoven, Nehterlands, August 2003: 11-14

[4] Liu Y, Liao W K, et al. On-Line Processing Model for Data Mining in Large Scientific Simulations [C] // Proc. of 7th Workshop on Mining Scientific and Engineering Datasets in Conjunction with SDM. Lake Buena Vista, Florida, USA, 2004: 31-38

[5] Bachinsky S, Tarbox G, Powell E. Data Collection in an HLA Environment [C] // 97S-SIW-059. 1997

[6] Abdulla G, Critchlow T. Simulation Data as Data Streams [C]. ACM SIGMOD Record, 2004, 33(1): 89-94

[7] Gaber M, Zaslavsky A, et al. Mining data streams: A Review [C]. ACM SIGMOD Record, 2005, 34(2): 18-26

[8] 陈彬，张国强，黄柯棣. 一种基于 HLA 的通用仿真数据收集方法[J]. 计算机仿真，2006，23(5)：127-130

[9] Chi Y, Wang H, Yu P S, et al. Moment: maintaining closed frequent itemsets over a stream sliding window [C] // Proc. of ICDM'04. Brighton, 2004: 59-66

[10] Ao F J, Du J, Huang K D, et al. An Efficient Algorithm for Mining Closed Frequent Itemsets in Data Streams // Proc. of CIT'08. Sydney, Australia, July 2008