第 28 卷 第 4 期 西南师范大学学报(自然科学版) 2003 年 8 月
Vol. 28 No. 4 Journal of Southwest China Normal University (Natural Science) Aug. 2003

文章编号:1000-5471(2003)04-0547-03

多数据流时间序列中的依赖模式发现算法研究①

王 刚[1], 吴代贤[2]

1. 重庆教育学院 计算机与现代教育技术系,重庆 400067;2. 西南师范大学 计算机与信息科学学院,重庆 400715

摘要:针对多数据流组成的时间序列中发现有用的结构模式的 MSDD 算法不能很好地对节点剪枝,以及直观地表示模式的时间关系的问题,经过研究,提出了发现多数据流时间序列结构模式的算法:时间窗口移动筛选算法(TWMA). 采用事件序列化的策略来发现多流时间序列中的依赖模式,与 MSDD 相比,在表示上更直观,发现模式的过程更灵活.

关 键 词:数据挖掘;时间序列;模式发现

中图分类号:TP311.52 文献标识码:A

文献[1]对多流时序进行了研究,提出了 MSDD 算法,发现事件的依赖关系. 由于该算法不能对中间产生的模式进行及时、有效的剪枝,使访问空间很大. 多流时序的挖掘有其自身的特点,其它单一数据流的挖掘算法[2-5]不能用在对多流时序的挖掘上. 本文正是在此基础上,提出了时间窗口移动筛选算法(TWMA). 对多数据流情况下的依赖模式发现进行了研究.

1 MSDD 算法分析

MSDD 算法要解决的问题是发现多流序列事件之间的依赖关系[1, 6-8].

该算法通过生成依赖空间来实现,得到前驱事件共$(m*n+1)^{w_p}$个,后继事件共$(m*n+1)^{w_s}$个,n 表示每条数据流中事件的个数,m 表示数据流的个数,w_p,w_s 分别表示前驱时间、后继时间. 则依赖关系树的空间最大为$(m*n+1)^{(w_p+w_s)}$.

算法的时间复杂度为 $O((m*n+1)^{w_p+w_s})$,m 为数据流的数目,n 为每个数据流中事件的个数. 该算法的不足之处在于:由于采用树结构,事件之间的时间关系表示不够直观. 该算法没有针对前驱时间段内的有效事件,后继时间段内的有效事件来生成依赖空间,而是以所有发生的事件为基础,将剪枝操作放在了子节点生成的后面,这就需要大量的运算,占用了大量的空间. 而本文提出的时间窗口移动筛选算法(TWMA)考虑到了这个问题,它先对数据流进行数据初始化,剔除明显不符合要求的事件;同时,在关系的表示上它更直观、简洁.

2 TWMA 算法

2.1 事件序列的产生

时间序列中事件的关系,表现为事件发生的相对时间. 如何将这种关系体现出来,关系到准确、快速地判断事件序列的支持数,本文提出了如下的策略.

假设窗心为‘a’,则窗口内先于‘a’发生的事件表示为集合‘{<”+ 事件序列}+‘a’’,称为前驱事件集.

窗口内后于‘a’发生的事件表示为集合‘{>”+‘a’+ 事件序列}’,称为后继事件集.

① 收稿日期:2002-12-03

作者简介:王 刚(1972-),男,四川广安人,硕士,讲师,主要从事人工智能、数据挖掘的研究.

窗口内与‘a’同时发生的的事件表示为集合‘{%”+‘a’+ 事件序列}}，称为同时发生事件集.

例如：三数据流时间序列：

a	b	c	b	d	a	e
1	2	3	1	1	2	3
A	B	X	Y	A	B	X

以事件‘c’为窗心，前驱，后继时间都为 2，则窗口中事件序列表示为：

相对时间为 0 的序列 {%c3X，%c3，%c*X，}；

后继序列 {> cB，> cbd，> c*d，> c1，> c11，> c*1，> c Y，> cYA，> c*A}；

前驱事件序列 {< bc，< abc，< a*c，< 2c，< 12c，< 1*c，< Bc，< ABc，< A*c}.

这样，集合中就没有相同的事件序列出现，也刻画了所有存在的序列. 这便于以后准确地计算序列的支持度.

2.2 算法的思想及步骤

① 初始化数据流，剔除不满足最小支持度的事件.

② 以指定流的各不同事件为窗心，建立窗口，得到事件序列的集合.

③ 将集合中事件的支持数与给定的最小支持度比较，剔除不满足要求的事件序列.

④ 输出满足条件的序列.

⑤ 重复 ② 直到所有事件都已经作为窗心.

⑥ 对窗口进行整理，得到的窗口就反映了结构模式.

2.3 算法描述

```
TWMA(S, ws, wp)
 {
S = {s1, s2, s3, s4, … }
initialiate(S); // 初始化数据流
sc = 指定流 s1 中不同的事件的集合；
for (i = 0; i < = sc. getlength() - 1; i ++) // 对每一个不同的事件
 {
for (y = 0; y < = s1. getlength(); y ++) // 以 s1 为参照
 {
if (s1. mid(y, 1) = = sc. mid(i, 1)) // 找窗心事件 sc. mid(i, 1) 的发生总次数
count ++;
}
for(k = 0; k < = count; k ++) //count 个事件窗口
 {
m = s1. find(sc. mid(i, 1), m + p); // 对每个窗口，找 sc. mid(i, 1) 窗心出现的位置
各流以位置 m 为中心，对前驱、后继时间范围内不同位置的事件序列化，结果存储在 s[k] [j + v]；//v 表示不同位置的个数.
}
计算 s[x] [j] 的支持度 sp；
if(sc. Mid(i, 1) = = 突变事件}) // 突变事件特殊处理
   sp = sp + offset; // 改变突变事件的支持度使 sp ≥ min sup port
if(sp < = min sup port)
   s[x] [j] =“ ”; // 去除不满足要求的事件
输出所有不等于空的 s[x] [j]，组合成结构模式；
}}
```

3　TWMA 与 MSDD 的比较

(1) TWMA 较之于 MSDD，可以灵活地将我们感兴趣的事件设置为窗心事件，使得发现的知识更有针对性. 这样，如果以突变事件为窗心，就可避免其因支持度小而被淘汰. 在 MSDD 中忽略了突变事件的发现，而 TWMA 克服了这个问题.

(2) 在序列模式的产生上 MSDD 算法不能及时剪枝，而 TWMA 通过预处理来克服了这个问题.

(3) 在单机环境下执行，TWMA 的运算复杂度较之于 MSDD 算法要小且随数据流的长度的增加，它的增长趋势要缓慢些.

4　结 束 语

本文在分析 MSDD 算法的基础上提出并成功地实现了更灵活、直观、全面的 TWMA 结构模式发现算法，适合数据集比较小的情况，如果数据集很大，花费的时间也会比较多的，因此，接下来的工作是研究更好的分布式、并行模型和算法[7, 8].

参考文献：

[1] Tim Oates. Searching for Structure in Multiple Streams of Data [A]. The Thirteenth International Conference on Machine Learning [C]. Italy：Barl，1996. 346 – 354.

[2] 陆玉昌. 数据挖掘与知识发现 [J] . 计算机用户，2000，5：130 – 132.

[3] 何炎祥. 时序模式的几种开采算法及比较分析 [J]. 小型微型计算机系统，2001，5(22)：120 – 123.

[4] Agrawal R，Srikant R. Mining Sequential Patterns Research [M]. California：IBM Almaden Research Center，1994. 1 – 12.

[5] Agrawal R. Parallel Mining of Association Rules [M]. California：IBM Almaden Research Center，2001. 1 – 5.

[6] 王　刚，程小平. 多流分段比较法发现多流时序的结构模式 [A]. 中国人工智能进展 [C]. 北京：北京邮电大学出版社，2001. 394 – 395.

[7] Tim Oates，Paul R，Cohen. Parallel and Distribute Search for Structure in Multivariate Time Series [A]. ICML，Machine Learning ECML – 97 [C]. Berlin，New York：Springer-Verlag，1997. 1 – 30.

[8] Matthew D，Schmill，Tim Oates. A Distribute Approach to Finding Complex Dependencies in Data [A]. University of Massach Usetts，Center Computer Science Technical Report [C]. Massachusetts：Lederle Graduate Research Center，1998. 1 – 20.

Research on The Structure Patterns of the Multiple Time Series

WANG　Gang[1]，　WU Dai-xian[2]

1. *Dept. of Computer and Modern Education Technology，College of Chongqing Education，Chongqing* 400067，*China*；

2. *School of Computer and Information Science，Southwest China Normal University，Chongqing* 400715，*China*

Abstract：It is very important to find the structure patterns from the multiple time series. A famous algorithm provided by Tim oates is the MSDD，it finds the dependency patterns by the dependency trees. The main problem is that it can't have a trim in times to the nodes，and it can 't express the time relation clearly. A TWMA(Time Window Move Algorithm) algorithm is made to solve the above problems. It is more simple，clear，concision than the other algorithm by use of the fuzzy theory and the ideas of events serial.

Key words：data mining；time series；pattern discovery

责任编辑　覃吉康