

基于类 Apriori 原则的区域周期模式挖掘算法

李鹏岳¹, 何 杰^{2,3}, 谭志国³

(1 陆军军医大学(第三军医大学) 生物医学工程系 医学图像学教研室, 重庆 400038;

2 国防科技大学 计算机学院, 湖南 长沙 410073; 3 武警警官学院 信息通信系, 四川 成都 610213)

摘 要: 时间序列在各领域广泛存在, 针对目前已有的周期模式难以描述仅在时间序列部分区域存在的周期行为的现状, 提出了区域周期模式的概念. 在对此区域周期模式进行详细形式化描述的基础上, 进一步提出了一种基于类 Apriori 原则的区域周期模式挖掘算法. 该算法能在保证挖掘结果完整性的同时, 较大程度地提高挖掘效率. 通过在网络安全公开数据集上的实验表明, 该算法能快速精确地挖掘出时间序列中所有的区域周期模式.

关键词: 时间序列; 区域周期模式; Apriori 原则; 模式挖掘

中图分类号: TP301

文献标识码: A

文章编号: 1000-7180(2018)09-0085-05

A Regional Periodic Pattern Mining Algorithm Based on Apriori-like Principle

LI Peng-yue¹, HE Jie^{2,3}, TAN Zhi-guo³

(1 Department of Medical Image, School of Biomedical Engineering, Army Medical University
(Third Military Medical University), Chongqing 400038, China;

2 College of Computer, National University of Defense Technology, Changsha 410073, China;

3 Department of Information and Communication, Officers College of PAP, Chengdu 610213, China)

Abstract: Time series exists widely in various fields. In view of the fact that the existing periodic models are difficult to describe periodic behaviors only activate in partial regions of time series, a novel concept called regional periodic pattern is proposed. After formalizing this regional periodic pattern, a regional periodic pattern mining algorithm based on apriori-like principle is further proposed, which can greatly improve the efficiency of mining under the premise of ensuring the integrity of the mining results. Experiments on a public dataset of network security show that the proposed algorithm can quickly and accurately mine all regional periodic patterns in time series.

Key words: time series; regional periodic patterns; apriori principle; pattern mining

1 引言

对时间序列分析的一个重要方面是其周期模式的挖掘, 挖掘时间序列上的周期性行为在诸多领域都具有重要的研究价值.

目前已有的周期模式主要包括全周期模式和部分周期模式^[1-3]. 而有时在某一个时间序列在整体趋势上可能并没有明显的周期性, 但在时间序列的某一部分上隐含着短暂的周期性行为. 由于出现的概

率并不固定以及延续时间不能估计, 因此在基于整体的模式挖掘中这种周期行为往往会被遗漏.

针对这种只在部分区域出现的周期性, 本文提出了一种新型的周期模式, 称之为区域周期模式. 区域周期模式的研究有利于挖掘时间序列中隐藏的周期性行为, 从而为揭示时间序列特征提供了更好的理论依据. 文献[4]提出的密集型周期模式虽然也定义在部分时间序列上, 但其只是区域周期模式的一种个例.

由于传统用于全周期模式或者部分周期模式的

挖掘算法^[2,3,5],不能直接移植用来进行区域周期模式挖掘. 本文首先对提出的区域周期模式进行了形式化的描述和分析,然后提出了一种基于类 Apriori 原则的区域周期模式挖掘算法,并验证了算法的可行性和有效性.

2 区域周期模式的形式化描述

2.1 基础概念

假定 S 是一个等时间采样的事件时间序列, $S = e_1, e_2, \dots, e_n, e_i \in L (i = 1, 2, \dots, n)$. 其中 n 是时间序列的长度, L 是事件集合, 指代所有事件类型对应的字符集. 定义 p 为 S 的周期, 则 S 可以划分为至多 $C = \lfloor n/p \rfloor$ 个等长周期区间(留下部分被丢弃). 这些周期区间被称为周期段, 可以表示为: $E_i = [e_{ip+1}, \dots, e_{i(p+p)}], i \in [0, C)$.

模式 s 为定义在 $(2^L - \{\emptyset\}) \cup \{*\}$ 上的一个非空序列 $s = e_1, \dots, e_p$. 其中 $*$ 代表通配符, 即可以匹配中的任何字符. 定义模式 s 的阶数 L_len 为 s 中非 $*$ 字符的个数, 一个模式如果拥有 j 个非 $*$ 字符, 则称其为 j 阶模式. 举例来讲, 模式 $s = a * c * a$ 中含有三个非 $*$ 事件, 则称 s 为三阶模式. 同时, 如果一个模式 $s' = e'_1, \dots, e'_p$ 与 s 具备相同的长度, 且 $e'_i = e_i$ 或者 $e'_i = *, i \in [1, p]$, 则称 s' 为 s 的子模式, 反之, s 为 s' 的祖先模式, 如果两者的阶数只相差 1, 则称 s 为 s' 的父模式.

定义一个时间序列 S 中关于模式 s 的支持度 supp 和置信度 conf 为:

$$\text{supp}(s, S) = \sum_{i=0}^{C-1} \text{match}(s, E_i) \quad (1)$$

式中, $\text{match}(s, E_i) = \begin{cases} 1, & \text{如果 } E_i \text{ 匹配 } s \\ 0, & \text{其他情况} \end{cases}$

$$\text{conf}(s, S) = \frac{\text{supp}(s, S)}{C} \quad (2)$$

公式(1)中周期段 E_i 匹配模式 s 定义为当且仅当, 对于每一个位置 $j \in [1, p], s$ 中 e_j 等于 $*$ 或者与 E_i 中位置 j 的字符相同. 同时, 如果一个模式 s' 是 s 的子模式, 可推理知, E_i 匹配模式 s' .

2.2 区域周期模式的定义

定义 1: 全周期模式和部分周期模式

给定长度为 n 的事件时间序列 S, L 为 S 的事件集合. 指定 min_conf 为最小置信度, min_supp 为最小支持度, p 为周期. 则有, 如果一个模式 $s = e_1, \dots, e_i, \dots, e_p, (i \in [1, p], e_i \in L)$ 满足: $\text{conf}(s, S) \geq \text{min_conf}, \text{supp}(s, S) \geq \text{min_supp}$. 则称 s 为 S 的一个全周期模式. 如果一个模式 $s = e_1, \dots, e_i, \dots, e_p,$

$(i \in [1, p], e_i \in L \cup \{*\})$ 满足: $\text{conf}(s, S) \geq \text{min_conf}, \text{supp}(s, S) \geq \text{min_supp}$, 则称 s 为 S 的一个部分周期模式.

定义 2: 区域周期模式

给定长度为 n 的事件时间序列 S, L 为 S 的事件集合. 指定 min_conf 为最小置信度, min_supp 为最小支持度, p 为周期, 则有 $C = \lfloor n/p \rfloor$, 周期段 $E_i = [e_{ip+1}, \dots, e_{i(p+p)}], i \in [0, C)$. 如果 Sub 为 S 的一个子序列且 $\text{Sub} = [E_i, E_{i+1}, \dots, E_j], (i \geq 0, j \leq p)$, 称 Sub 为 S 的一个区域, 称 Sub 所含周期段数量 $(j - i + 1)$ 为 Sub 的区域长度, 记为 l . 若存在一个模式 $s = e_1, \dots, e_i, \dots, e_p, (i \in [1, p], e_i \in L \cup \{*\})$ 满足: $\text{conf}(s, S) \geq \text{min_conf}, \text{supp}(s, S) \geq \text{min_supp}$, 则基于定义 1, s 为 Sub 的一个部分周期模式. 同时, 称 s 为 S 的一个区域周期模式.

举例说明, 假定一个事件时间序列 $S = \text{abcabcabcabcabdcddadaaaacbcbbccac}$ 且周期 $p = 3, \text{min_conf} = 0.6, \text{min_supp} = 4$. 如果一个模式 $s = ab*$, 有 $C = \lfloor 37/3 \rfloor = 12, \text{supp}(s, S) = 5$, 则 $\text{conf}(s, S) = 5/12 < \text{min_conf}$, 故 s 不是 S 的部分周期模式. 若定义 S 的一个区域 $\text{Sub}_S = \text{abcabcabcabcabd}$, 则有: $\text{conf}(s, \text{Sub}_S) = 5/5 = 1 > \text{min_conf}, \text{supp}(s, \text{Sub}_S) = 5 > \text{min_supp}$. 根据定义 1 和定义 2, 模式 $s = ab*$ 是序列 Sub_S 的一个部分周期模式, 是序列 S 的一个区域周期模式.

由定义 1 可知, 全周期模式和部分周期模式的区别仅在于全周期模式中的字符不能为 $*$, 即全周期模式的事件必须全部来自事件集合, 模式的阶数等于周期 p . 而部分周期模式放松了这一要求, 不必完全匹配, 可以包含阶数从 1 到 p 的所有满足条件的模式. 此外, 相对于全周期模式和部分周期模式需要在所有周期段中满足最小支持度和最小置信度这两项条件, 区域周期模式只需要在区域周期段(即总序列的部分子集)中满足以上条件即可.

3 基于类 Apriori 原则的区域周期模式挖掘算法

基于频繁模式和周期模式的相似性, 频繁模式挖掘算法中常用的部分基本原则往往可以运用于周期模式挖掘算法中. Apriori 原则^[6]即是其中之一: “一个 k 项集是频繁的, 当且仅当它的所有子项集是频繁的”. Han 等人^[3]将此原则移植到了周期模式挖掘中: “每一个周期为 p 的(全/部分)周期模式的

子模式也是一个周期为 p 的频繁模式^①”。这就使得我们可以通过 $i-1 (i \in [2, p))$ 阶周期模式结合 1 阶周期模式来推导所有可能的 i 阶周期模式. 在区域周期模式挖掘中, Apriori 原则同样适用, 可以描述为: “所有周期为 p 的区域周期模式的所有子模式同样是周期为 p 的区域周期模式”. 可基于区域周期模式的定义证明如下:

证明: 假设一个模式 s 是时间序列 S 的一个区域周期模式, 则存在 S 的某一个区域 Sub, 使得: $\text{supp}(s, \text{Sub}) \geq \min_supp$ 且 $\text{conf}(s, \text{Sub}) \geq \min_conf$. 假定 s' 是模式 s 的一个子模式, 不失一般性, 根据子模式的含义可知: $\text{supp}(s', \text{Sub}) \geq \text{supp}(s, \text{Sub}) \geq \min_supp$ 且 $\text{conf}(s', \text{Sub}) \geq \text{conf}(s, \text{Sub}) \geq \min_conf$ 由此可知, s' 同样 Sub 的部分周期模式, 即为 S 的区域周期模式. 得证.

3.1 获取一阶区域周期模式集合 F_1

在全(部分)周期模式中, 由于每一个模式的活跃期都定义在时间序列的完整时间段上, 所以当最小置信度 $\min_conf$ 大于 0.5 时^②, 每一个周期段位置上的一阶周期模式都只有 1 个. 但在区域周期模式中, 由于一阶区域周期模式可以活跃在时间序列的多个区域内, 因此满足条件的一阶区域周期模式不仅可以有多个, 而且可以在多个区域范围上. 对于每一个一阶区域周期模式, 本文都只记录其满足条件的最大区域. 求解一阶区域周期模式过程的算法如表 1 所示. 主要过程为: 首先初始化一个数据结构 F_1 , 用来记录所有一阶区域周期模式及其最大区域和置信度. 然后对于每一个周期段上的位置 j 和每一个 L 中的字符 l (对应这里的一阶模式), 检查所有可能的区域并且判断在每一个区域上 l 是否满足最小置信度和最小支持度条件. 同时, 本文提出了三种剪枝策略用来降低计算代价.

第一, 本文将连续的周期段序号作为一个分组来处理, 如算法第 8 行. 这样的处理将有效降低遍历的数目. 举例来讲, 如在位置 j 上, 字符 a 活跃在 $E_1, E_2, E_4, E_5, E_6, E_7, E_{10}$ 等 7 个周期段上. 如果按周期段序号来遍历区域, 则共有 $\{E_1\}, \{E_1, E_2\}, \{E_1, E_2, E_4\} \cdots \{E_7, E_{10}\}, \{E_{10}\}$ 共 $7 \times (7+1)/2 = 28$ 个区域, 如果将连续的周期段序号作为一个分组, 得新组 $E_{1_new} = \{E_1, E_2\}, E_{2_new} = \{E_4, E_5, E_6,$

$E_7\}, E_{3_new} = \{E_{10}\}$, 则只需遍历 $3 \times (3+1)/2 = 6$ 个区域. 之所以能够这么做, 是因为在本算法中对于每一个模式都计算其能够活跃的最大区域, 对于任意模式, 连续的活跃周期段都会增加其支持度和置信度, 因此连续的活跃周期段是不可分隔的. 这里的活跃周期段是指所有在当前位置出现当前字符的周期段.

表 1 一阶区域周期模式求解算法

输入	事件时间序列 $S = e_1, e_2, \dots, e_n$, 周期 p , 最小支持度 m , 最小置信度 c
输出	存放一阶区域周期模式数据结构的集合 F_1
1	$C = \lfloor n/p \rfloor$
2	划分 S 为 C 个周期段, 得到 $SE = E_1, E_2, \dots, E_C, E_i = [e_{ip+1}, \dots, e_{ip+p}], i \in [0, C)$
3	For j in $[1 : p]$
4	$J = [E_{0j}, E_{1j}, \dots, E_{(C-1)j}]$ // 周期段位置为 j 的所有事件集合
5	$LJ = \text{distinct}(J)$ // J 中独立的事件类型
6	For l in LJ
7	$TP = J(l)$ // TP 记录 J 中出现字符 l 的时间点
8	$TS \ TP$ // 转换 TP 到多个时间段 (把连续的时间点作为一个时间段)
9	For start in $[1 : TS]$
10	For end in $[TS : \text{start}]$
11	$R = TS[\text{start} : \text{end}]$
12	If $\text{supp}(l, R) \geq m$ 且 $\text{conf}(l, R) \geq c$
13	If $R \notin F_{Lj}.R$
14	$F_{Lj} \leftarrow \{R, \text{conf}(l, R)\}$
15	Goto line 9 for next start
16	End
17	End
18	$F_j \leftarrow F_{Lj}$
19	End
20	$F_1 \leftarrow F_j$
21	End
22	Output F_1

第二, 当遍历区域时, 设置结束边界从远周期段到近周期段, 如算法第 10 行. 如前文中的 E_{1_new} 作为开始边界时, 本文首先判断 E_{3_new} 作为结束边界时当前模式是否满足约束条件. 如果满足, 则记录当前区域和模式 (算法第 14 行), 并且直接跳转到下一开始边界 (算法第 15 行), 而不再继续判断更近的 E_{2_new} 作为结束边界. 这同样是因为对于每一个模式, 算法都试图获取其满足条件的最大区域. 如果在某一个区域内满足条件, 则不必判断其区域内的子区域. 这样的策略可以进一步有效降低所遍历区域的个数.

第三, 在循环过程中, 当发现一个满足条件的模式时, 在记录该模式和区域之前, 需判断当前区域是

① 这里称为频繁模式是因为在全周期模式中周期模式的子模式不能称为周期模式, 但在部分周期模式中, 其子模式同样是部分周期模式, 只是阶数较低. 下文统称为阶数不同的周期模式.

② 在周期模式的应用中, 通常使用大于 0.5 的值作为最小置信度。

否已经包含于已记录的某一个区域,如果被包含,则不记录当前区域,直接跳转到下一开始边界(算法第 13 行). 原理同上.

基于以上三个原则,本算法能够快速且完整地发现所有一阶区域周期模式及其最大活跃区域. 图 1 给出了某一时间序列在周期为 13 时,所有一阶区域周期模式及其最大活跃范围的示意.

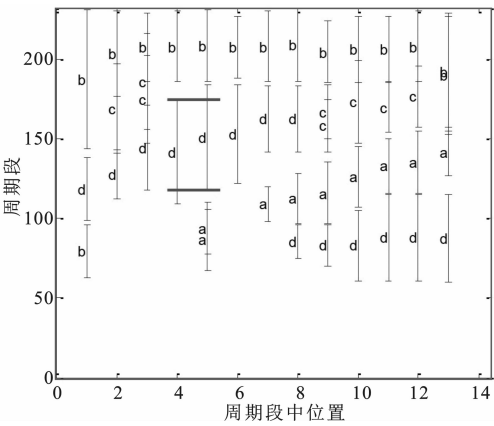


图 1 一阶区域周期模式分布

3.2 推导高阶区域周期模式

在获取 F_1 之后,就可以基于 Apriori 原则,通过结合 F_1 与 $i-1(i \in [2, p])$ 阶周期模式来推导可能的 i 阶周期模式,直到某阶模式为空. 同时,需要注意的是,由于每一个周期段位置上的一阶周期模式可能有多个,并且分布在多个不同区域. 如果简单将其与其他模式结合来推导高阶模式,可能产生巨大的组合数和运算代价. 在最差情况下, L 中的每一个字符在每一个位置上都可能是一个周期模式,并且对于同一个位置上的同一个模式,还可能有多多个区域. 如图 1, 周期段位置 1 上面对应的一阶周期模式($b * * * * * * * * * * * *$)就有两个活跃区域. 所以不能简单的计算所有的模式组合. 这里本文提出了交叉组合概念,即在组合可能的高阶模式时,只有 $i-1$ 阶周期模式与 F_1 中的模式在活动区域上交叉时,才进行组合操作,并且组合后的大区域范围也只定义在交叉区域上. 如图 1 中,当模式($* * * d * * * * * * * * * *$)与($* * * * d * * * * * * * * *$)试图组合出二阶模式($* * * dd * * * * * * * * *$)时,首先判断两者活动区域是否有交叉,若交叉,则进一步获取新的检测范围,即图 1 上两条水平线之间的周期段范围. 之后,在此交叉范围内遍历可能的区域,进而判断是否有当前组合模式所满足条件的区域. 其中遍历可能子区域时,同样采用如步骤一所示的三种剪枝方法.

通过以上两个步骤,本算法就可以较快的遍历出所有可能的区域周期模式及对应的最大活跃区域.

4 实验评估

实验的数据集来自某网络安全实验室中的一个 P2P 僵尸网络流量数据集 ISOT^[7]. 为了挖掘其中的 P2P 僵尸网络所呈现的区域周期性行为,本实验仅关注该数据集中报文大小在 60 字节到 120 字节之间的本地请求报文,并且忽略其中的 DNS 类型或者 PING 类型报文. 本文定义该数据集中的事件为某本地 IP 地址每秒发送上述特定类型报文的数量.

为将报文数量的具体数值转换为不同事件,本文规定转换方法为:若给定本地 IP 地址在数据集整个时间段中一秒内发送以上特定报文数量的最大值为 q ,则记该 IP 地址一秒内发送此类报文数量 $x \in [0, q/5]$ 的事件为 a ,一秒内发送此类报文数量 $x \in (q/5, 2q/5]$ 的事件为 b ,一秒内发送此类报文数量 $x \in (2q/5, 3q/5]$ 的事件为 c ,一秒内发送此类报文数量 $x \in (3q/5, 4q/5]$ 的事件为 d ,一秒内发送此类报文数量 $x \in (4q/5, 5q/5]$ 的事件为 e . 依据此方法,则可为每一个本地 IP 地址提取出一个用字符集 $\{a, b, c, d, e\}$ 表示的事件时间序列.

本文随机选取了该数据集的 4 个本地 IP 地址,并对它们的区域周期行为进行了分析. 为确定各 IP 地址对应时间序列的周期,本文采用自相关函数分析来为其推荐周期取值^[8],结果如表 2 所示.

表 2 各个 IP 地址的推荐周期

IP 地址	推荐周期
172. 16. 0. 11	10, 4
172. 16. 0. 12	13
172. 16. 2. 2	3
172. 16. 2. 11	11

为对比分析本文算法的性能和效率,本文还实现了一种基于迭代的区域周期模式挖掘算法. 该算法通过迭代遍历时间序列上所有可能的区域,并利用 Hit 算法^[3]求解这些区域上的部分周期模式(所得即为整个时间序列的区域周期模式),最终得出区域周期模式集合.

基于表 2 的推荐周期,本文分别利用基于类 Apriori 原则的区域周期模式挖掘算法(记为算法一)和基于迭代的区域周期模式挖掘算法(记为算法二),来对表中 4 个 IP 地址对应的事件时间序列进行了区域周期模式挖掘运算. 本实验中,设置 $\text{min_conf}=0.65$, $\text{min_supp}=20$. 实验环境位于同一台

PC 主机,基于 MATLAB 7.2 完成.实验结果如表 3 所示.

由于算法二遍历了所有可能的区域,所以本文认为其结果是完整可信的.由表 3 可知,算法一能够得到与算法二一致的结果,说明其结果集也是完整的,也说明本文采用类 Apriori 原则进行区域周期模式挖掘是合理的.

表 3 区域周期模式挖掘结果

IP 地址	模式阶数	模式数量	
		算法一	算法二
172.16.0.11 (1)	1	2	2
	172.16.0.11 (2)	1	1
172.16.0.12	1	46	46
	2	84	84
	3	84	84
	4	43	43
	5	8	8
	1	83	83
172.16.2.2	2	48	48
	3	6	6
	1	29	29
	2	48	48
172.16.2.11	3	47	47
	4	30	30
	5	12	12
	6	2	2

本次实验中 5 次执行区域模式挖掘算法的时间开销如图 2 所示.其中 172.16.0.11(1)表示该 IP 地址的时间序列周期取 10 时的运算时间,172.16.0.11(2)为该 IP 地址的时间序列周期取 4 时的运算时间.由图可知,同算法二相比,算法一在保证得到完整解的前提下,大大减少了运算时间开销,这也验证了本文算法的高效性.

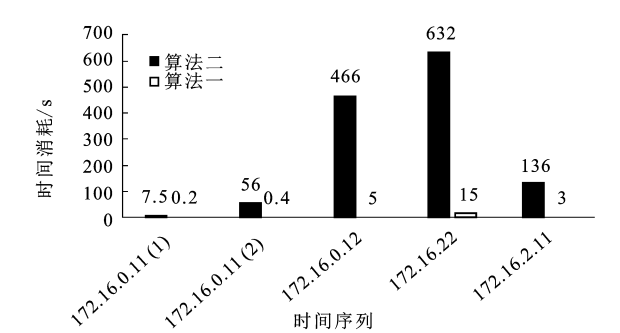


图 2 算法运行时间开销

5 结束语

本文提出了一种区别于全周期模式和部分周期模式的区域周期模式,为进一步挖掘时间序列的周

期性行为提供了一种新型的分析工具.在对区域周期模式进行详细形式化描述的基础上,本文进一步提出了一种基于类 Apriori 原则的区域周期模式挖掘算法,可以在保证得到完整解的前提下,较大幅度地提高算法挖掘效率.最后利用网络安全领域的公开数据集测试和评估了本文算法的性能,验证了算法的可行性和高效性.

参考文献:

[1] Qiao Y, Yang Y, He J, et al. Detecting parasite P2P botnet in eMule-like networks through quasi-periodicity recognition [M]. Berlin Heidelberg, Springer, 2012:127-139.

[2] Zhang D, Lee K, Lee I, et al. Hierarchical trajectory clustering for spatio-temporal periodic pattern mining [J]. Expert Systems with Applications, 2018: 92.

[3] Han J, Yin Y, Dong G. Efficient mining of partial periodic patterns in time series database[C]// International Conference on Data Engineering, 1999. Los Angeles, USA, IEEE Xplore, 1999:106-115.

[4] Sheng C, Hsu W, Lee M L. Mining dense periodic patterns in time series data[J]. IEEE Transactions on Knowledge & Data Engineering, 2003, 15 (3): 613-628.

[5] Xylogiannopoulos K F, Karampelas P, Alhadj R. Repeated patterns detection in big data using classification and parallelism on LERP Reduced Suffix Arrays[J]. Applied Intelligence, 2016, 45(3):1-31.

[6] Agrawal R, Srikant R. Fast algorithms for mining association rules[C]// Very Large Data Bases Conference, Santiago de Chile,1994 , 23 (3) :487-499.

[7] Saad S, Traore I, Ghorbani A, et al. Detecting P2P botnets through network behavior analysis and machine learning[C]// Ninth International Conference on Privacy, Security and Trust. Montreal, Quebec, Canada IEEE, 2011:174-180.

[8] 张玲,陈硕,张旻. 基于自相关获取周期的时间序列模式挖掘算法[J]. 信息工程大学学报, 2015, 16 (2): 209-213.

作者简介:

李鹏岳 男,(1988-),博士研究生.研究方向为计算机应用等. E-mail:llyue0511@163.com.

何杰 男,(1984-),博士,讲师.研究方向为网络安全.

谭志国 男,(1978-),博士,讲师.研究方向为机器视觉.