DOI:10.13203/j.whugis20150616　　文章编号:1671-8860(2018)05-0766-07

顾及背景知识的多事件序列关联规则挖掘方法

何占军[1]　邓　敏[1]　蔡建南[1]　刘启亮[1]

1　中南大学地理信息系,湖南 长沙,410083

摘　要:事件序列关联规则挖掘旨在发现序列中不同事件在邻近时间域内的相互依赖关系,对于理解事件间的交互作用机制具有重要意义。然而,当前事件序列关联规则挖掘方法忽略了序列中事件的分布特征,支持度与置信度阈值参数设置困难,进而造成了挖掘结果的冗余或遗漏问题。充分考虑序列中事件的固有分布特征,定义了新的规则度量指标,并给出了一种顾及背景知识的多事件序列关联规则挖掘算法。实验结果表明,与当前经典的 MOWCATL 算法比较,此方法挖掘结果更加准确,且规则度量指标间的一致性更好,可有效改善挖掘规则冗余或遗漏问题。应用此方法对 2013 年冬季北京市 $PM_{2.5}$ 浓度与气象因素的多序列进行挖掘,发现 $PM_{2.5}$ 浓度与空气相对湿度的联系最为紧密,高湿、低温和弱风环境最容易导致高浓度 $PM_{2.5}$ 的形成。

关键词:背景知识;数据挖掘;关联规则;多事件序列;$PM_{2.5}$

中图法分类号:P208　　**文献标志码**:A

随着地理空间数据获取技术的迅速发展,地理空间数据的数据类型、覆盖范围及获取速度已呈爆炸式增长,如何从这些海量的地理空间数据中发现知识已成为空间信息技术的研究热点。地理空间数据挖掘是地理空间知识发现的核心内容,旨在从海量地理空间数据中发现潜在的、有意义的模式,对地理现象的理解、分析和预测有重要的指导意义。关联规则是地理空间数据挖掘中的一个重要内容,可以发现不同要素、现象或事件间的关联关系,其已被广泛应用于土地覆盖变化分析、生态物种分布、气象预报、疾病相关因子分析等领域[1-5]。

关联规则挖掘的概念最早由文献[6]提出,其一般形式为 $A \rightarrow B(s, c)$。其中,A 称为规则前件,B 为规则后件,s 和 c 为规则度量的两个重要指标,前者称为支持度(或频繁度),表示规则在数据库中出现的频率,后者为置信度,表示在前件发生前提下后件发生的概率。文献[7-8]提出了时间序列模式挖掘的问题,并在随后给出一种通用序列模式挖掘方法 GSP。此后,为进一步提升算法效率,SPADE[9]、Prefixspan[10] 等类似算法陆续被提出。然而,这些算法主要针对序列集合数据,其目标在于找出序列集合中频繁出现的子序列。实际中,时间序列观测数据通常是对同一地区某种变量的长期重复观测,如环境监测、气象监测数据。这种序列数据的一个基本特征是序列较长,记录的是同一现象随时间演化的不同现象或状态,如降水序列数据可体现干旱程度(干旱、潮湿等)。这些不同的现象或状态可视为事件,则观测序列也就是记录了不同事件的事件序列。则针对单个事件序列,文献[11]定义不同类型事件的有序组合为频繁事件集,并发展了两种频繁事件集挖掘算法 WINEPI 和 MINEPI。随后,文献[12]发展了基于事件约束的频繁事件集挖掘算法 REAR[12],以精简挖掘结果从而提升结果的易解译性。针对多个事件序列的频繁事件集挖掘,文献[13]提出了一种新的算法 MOWCATL。该方法允许不同事件存在时间滞后,并被成功应用于分析美国内布拉斯加州干旱与其他海洋大气指数间的关联关系[14]。在国内,一些学者同样利用时序关联规则挖掘的技术研究了太平洋暖池和中国内陆降水之间的遥相关现象[15-16]。

上述序列关联规则挖掘方法大都是在频繁度-置信度的框架下进行,该框架下算法的不足在于需要人为设置最小支持度和置信度阈值。这些阈值的设置不仅影响着挖掘算法的效率,而且影

收稿日期:2016-06-08

项目资助:湖南省自然科学杰出青年基金(14JJ1007);国家自然科学基金(41471385)。

第一作者:何占军,博士生,主要研究方向为时空关联模式挖掘方法及应用。hezhanjun000@126.com

通讯作者:邓敏,博士,教授。dengmin208@tom.com

响着挖掘结果的解译[17-18]。阈值设置太低会导致巨大的计算量,同时使得最终所得规则中包含大量冗余、非显著的规则,造成规则进一步筛选的难题;而阈值设置太高却又难以保证得到所有的有效规则。尽管这些阈值的设置至关重要,但实际中却鲜有关于这些阈值的先验信息,因此造成实际挖掘过程的困难。本文给出了一种顾及背景知识的多事件序列关联规则挖掘方法,旨在降低挖掘结果对人为设置参数的依赖性,提升挖掘过程中参数设置的自适应性和挖掘结果的可解译性。

1　顾及背景知识的多事件序列关联规则挖掘算法

1.1　研究策略

在传统支持度-置信度框架下的关联规则挖掘中,主要存在两个问题:(1)最小支持度等阈值的设置缺乏先验信息,造成阈值选取的困难;(2)最小支持度阈值仅仅是从全局的角度指定了不同类别事件需满足的最小数目,并未考虑不同事件所在背景序列中的分布特征差异,自适应性较差,难以适应序列中事件频率差异较大的情形。由此带来的后果是挖掘结果对参数依赖性太强,易发生规则冗余或遗漏的问题。因此,需要设计一种新的度量指标,以使其更好适应不同分布特征的事件序列,以降低挖掘结果对度量指标的依赖。

对事件序列关联规则挖掘而言,用户通常需要发现包含特定后件(感兴趣事件)的关联规则,例如挖掘重浓度污染相关联的气象条件。因此,可利用感兴趣事件作为后件约束。另一方面,前件对后件的影响作用只存在于一定的时间范围内,从而可将这种事件间的时间影响域视为邻近域。进而,将感兴趣后件邻近域内的前件分布特征视为对应前件的局部特征,而将整体序列中各前件的固有分布特征则可以视为一种全局特征。最后,从相对的视角出发,研究其局部分布特征和全局分布特征的差异。若某前件在感兴趣后件邻域内发生个概率远大于其在整体序列中平均的发生概率,则认为该前件与感兴趣后件之间存在较强的关联。这种思路的优势在于:(1)可以将事件在整个序列中的分布特征作为一种背景知识,从而使得度量指标设置带有了一定的先验知识;(2)度量指标不再局限于一个全局设置的固定参数,反映的是全局分布特征和局部分布特征的差异,刻画的是一种相对特征。因此,给定度量指标最小阈值时,当不同前件的全局分布特征存在较大差异时,算法对其在兴趣后件邻域内分布数目的要求随之发生变化,从而提升了阈值参数的自适应性。

本文提出了一种顾及背景知识的多事件序列关联规则挖掘算法(density ratio based association rules mining for multiple sequences, DR-BARMS)。为详细说明算法,首先给出相关定义。

1.2　度量指标定义

在多事件序列关联挖掘中,关联事件间相互影响的时间范围定义为影响域。影响域内的不同事件视为时间邻近。这种邻近关系可借助时间窗口来表示,即发生在同一时间窗口的事件视为互相邻近。算法以感兴趣事件作为后件约束,潜在关联事件作为前件,且认为前件对后件的影响仅存在给定的时间邻域内。换言之,前件和后件发生时刻允许存在时间滞后,但二者发生时刻必须足够相近(发生于同一时间窗口内)。

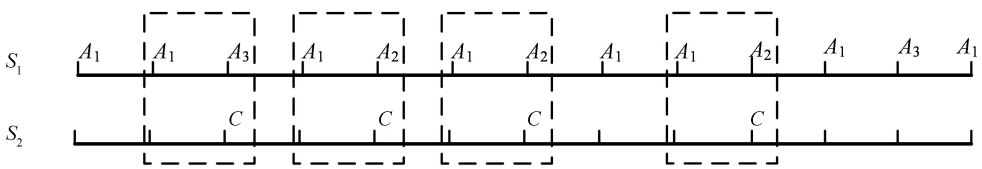

图 1　时间邻域示意图

Fig.1　Example of Temporal Neighboring Relationship

定义 1 全局计数(global count, GC):给定序列集 $S=\{S_1,S_2\cdots S_i\cdots S_n\}$,其中,$S_n$ 为后件序列,其他序列为前件序列。序列 $S_i(i\neq n)$ 中,事件 A_i 总计发生的数目称为全局计数。以图 1 为例,S_1 为前件序列,S_2 为后件序列,则有 $GC(A_1)=8$。

定义 2 局部计数(local count, LC):局部计数是指感兴趣后件邻域内前件总计发生的数目。本文认为兴趣后件是在前件作用下的结果,因此后件邻域不包含晚于后件发生的时刻。以如图 1 为例进行说明,图 1 中虚线框表示时间邻近域(时间窗口宽度为 2)。则以事件 C 作为后件约束的 A_1 局部计数:$LC(A_1\rightarrow C)=4$。同理,有 $LC(A_2\rightarrow C)=3$,$LC(A_3\rightarrow C)=1$。

定义3 全局密度(global density, GD):该指标用于刻画在某前件A_i在其所在序列S_i中的分布特征(视为全局分布特征),定义为全局计数与序列长度之比,具体表示为:

$$GD(A_i)=\frac{GC(A_i)}{length(S_i)} \tag{1}$$

式中,A_i为前件;S_i为该事件所在序列。图1中,序列长度为13,则$GD(A_1)=8/13$,$GD(A_2)=3/13$,$GD(A_1)=2/13$。

定义4 局部密度(local density, LD):该指标用于刻画感兴趣后件邻近时间域内前件A_i的分布特征(视为局部分布特征),具体表示为:

$$LD(A_i\rightarrow C)=\frac{LC(A_i\rightarrow C)}{width\times GC(C)} \tag{2}$$

式中,$LC(A_i\rightarrow C)$表示后件C邻域内的A_i的局部计数;$GC(C)$表示兴趣事件的全局计数,width表示时间窗口宽度。图1中,后件C邻域内发生的前件类型有$\{A_1,A_2,A_3\}$,且各自局部计数分别为4、3、1,则有$LD(A_1\rightarrow C)=4/8$。类似地,$LD(A_2\rightarrow C)=3/8$且$LD(A_3\rightarrow C)=1/8$。

定义5 密度比(density ratio,DR):局部密度与全局密度之比称为密度比,具体计算公式表示为:

$$DR(A_i\rightarrow C)=\frac{LD(A_i\rightarrow C)}{GD(A_i\rightarrow C)} \tag{3}$$

该指标用来衡量某前件在后件邻域内的局部分布与其在序列中全局分布特征的差异,若密度比大于给定密度比阈值(默认值取1),即局部密度远高于全局密度时,则认为这种关联关系是强关联关系。如图1中,$DR(A_1\rightarrow C)=13/16$,$DR(A_2\rightarrow C)=13/8$,$DR(A_3\rightarrow C)=13/16$。若仅根据局部计数(对应传统方法中的支持数指标)进行判断,则后件C主要与前件A_1相关联。然而,尽管后件C邻域内的A_1事件的局部计数更高,但导致该现象发生的原因是A_1事件在整个序列中本身的发生频率较高。因此,A_1和C之间的强关联实际是因为没有充分考虑A_1的全局分布特征而导致的一种伪关联关系。根据密度比指标则可以发现,事实上,后件C的与A_2间的关联程度更高。

定义6 相对密度比(relative density ratio, RDR):当同一后件与多个前件相关联时(大于给定密度比阈值),则需要从中进一步筛选出较为显著的关联前件,即需要比较各前件与后件之间关联程度的显著性。为此,进一步引入相对密度比指标,计算公式表示为:

$$RDR(A_i\rightarrow C)=\frac{n\times LD(A_i\rightarrow C)}{\sum_{i=1}^{n}LD(A_i\rightarrow C)} \tag{4}$$

式中,n为后件C邻域内前件类型数目。相对密度比是指同一后件约束下,某关联前件局部密度与所有关联前件局部密度均值的比率。该指标的意义在于比较不同规则间的相对重要程度,从而对规则进行适当过滤和筛选。相对密度比取值越大,说明其重要程度越高(默认值为1,即大于平均水平)。图1中,满足密度比指标的只有事件A_2,因此不需要继续计算相对密度比指标。

定义7 置信度(confidence ,Conf):对于关联规则挖掘结果的评价,应采取不同于关联规则挖掘过程中的指标参数。传统关联规则挖掘算法中的置信度指标度量的是给定前件是后件的条件概率,忽略了后件本身的分布特征。为此,本文将关联规则的置信度指标定义为:

$$Conf(A_i\rightarrow C)=\frac{LC(A_i\rightarrow C)}{\max(GC(A_i),\ LC(C))} \tag{5}$$

置信度定义为全置信度格式,指标综合考虑了前件与后件的分布特征,度量的是前件和后件的关联程度,取值范围为[0,1]。

1.3 DRBARMS算法描述

总的来说,DRBARM算法是一种基于后件约束的算法,将后件邻域内前件的分布特征视为局部分布特征,而将前件所在序列的整体分布特征作为一种背景知识。算法采取基于密度的思想,用密度指标衡量前件的局部和整体分布特征。同时,采取密度比指标,从相对的视角度量事件的局部分布特征和全局分布特征的差异。若感兴趣事件邻域内某前件密度远大于其在整个序列中的分布密度,则认为该前件与兴趣后件相关联,以此来削弱分布于全局的、高密度、不相关前件的影响,进而发现真正有意义的关联规则。算法流程描述如下。

输入:前件序列集$S=\{S_1,S_2\cdots S_i\cdots S_{n-1}\}$、后件序列$S_n$、滑动窗口长度width,最小密度比阈值min_density_ratio

输出:含多事件的关联规则

步骤1:计算前件序列$S_i(i=1\cdots n-1)$中各事件类型的全局计数和全局密度。

步骤2:选择后件序列中感兴趣事件C,并以C作为后件约束,计算其邻近时间域width内前件的局部计数和局部密度。

步骤3:计算前件事件$A_j(A_j\in S_i)$关于后件C的密度比,并选择密度比大于min_density_ratio的前件集合;若所得集合中前件类型数目大于1,进一步计算对应前件的相对密度比,并选择相

对密度比大于平均水平的前件集合,作为最终与兴趣后件 C 关联的一项集合 L_1。

步骤 4:对 L_1 中来自不同前件序列的事件进行组合,重复步骤 1～3,依次生成二项集 L_2、L_3 …直至 L_k 为空。

步骤 5:计算规则的置信度,输出规则,算法结束。

2　关联规则挖掘实验与分析

2.1　模拟实验

首先,采取模拟数据对本文算法的有效性和正确性进行验证。如图 2 所示,设置序列 S_1、S_2、S_3 为前件序列,序列 S_4 为后件序列。其中,序列 S_1、S_2 中各事件均以相同的概率随机生成,S_4 序列由 S_1、S_2 生成,关系式为 $S_4=0.5S_1+0.5S_2$。具体而言,事件 A_1、A_2、A_3 编码为 −1、0、1,事件 B_1、B_2、B_3 编码为 0、1、2,编码值大小表示对 S_4 的影响权重。最后,将所得序列 S_4 按得分由低到高的顺序等概率分为 3 个等级,分别记为 D_1、D_2、D_3。S_3 为独立的随机序列,其中事件 C_1、C_2、C_3 的发生概率为 0.7、0.2、0.1。显然,D_3 的产生仅可能与 A_3、B_2、B_3 相关,且 B_3 的影响作用一定大于 B_2。为验证本文算法,滑动窗口取为 1(S_1、S_2 中的事件仅仅影响同一时刻的 S_4 序列),密度比阈值设为 1.2。算法挖掘结果中包含后件 D_3 的规则列于表 1。同时,选择 MOWCALT 算法挖掘中较为显著的前 6 条规则列于表 2。

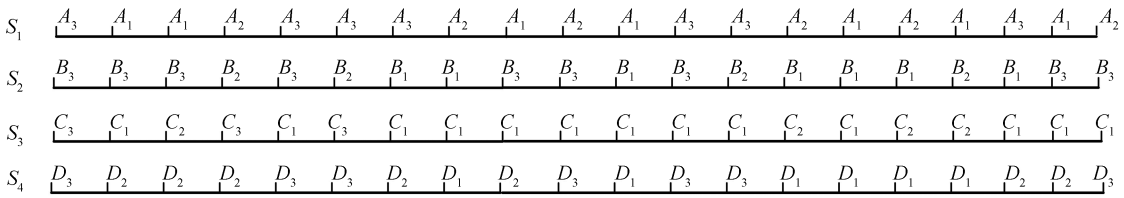

图 2　含干扰序列的多序列关联规则挖掘

Fig.2　Multiple Sequences Including an Uncorrelated Series

表 1　DRBARMS 算法挖掘所得规则

Tab.1　Association Rules Obtained by DRBARMS

规则编号	规则	密度比	置信度
1	$\{A_3\rightarrow D_3\}$	2.04	0.71
2	$\{B_3\rightarrow D_3\}$	1.59	0.56
3	$\{A_3+B_3\rightarrow D_3\}$	2.86	1

表 2　MOWCALT 算法挖掘所得规则

Tab.2　Association Rules Obtained by MOWCALT

规则编号	规则	支持数	置信度
1	$\{A_3+C_1\rightarrow D_3\}$	3	0.6
2	$\{A_3+B_3\rightarrow D_3\}$	3	1
3	$\{B_3+C_1\rightarrow D_3\}$	4	0.57
4	$\{C_1\rightarrow D_3\}$	5	0.38
5	$\{B_3\rightarrow D_3\}$	5	0.56
6	$\{A_3\rightarrow D_3\}$	5	0.71

由表 1 结果可知,本文算法挖掘结果与预期结果一致,所得规则表明事件 D_3 只与 A_3 和 B_3 呈明显关联。若采用 MOWCAL 算法,结果较大程度依赖于支持度阈值的设置,若设置太小,重要的规则可能遗漏。例如,尽管表 2 中规则 2 有着最高的置信度,但若支持度设置不合理(如大于 3),该规则将被遗漏。再者,规则的支持度和置信度之间存在不一致性,高支持度规则会同时伴随较低置信度,从而造成规则判读、选取和识别的困难,如表 2 中规则 2 和规则 3。此外,所得规则中包含一些关于高频率事件的冗余规则,这些规则实际对应的是相互独立情形,如规则 4 中事件 C_1 和 D_3。换而言之,此类规则实为一种虚假关联,不具备实际意义。相反,本文算法由于较好地顾及了序列的背景知识,可有效消除上述冗余或虚假规则。同时,本文算法避免了支持度阈值的设置,较好地降低了阈值参数依赖性,且所得结果与实际情形完全符合。

2.2　实例分析

采用真实数据对本文算法进行进一步分析。数据采用中国北京市 35 个空气质量观测站点 2013 年 12 月至 2014 年 2 月的 $PM_{2.5}$ 日均观测值序列及同期湿度、温度、风力观测序列,用于发现不同程度 $PM_{2.5}$ 水平与湿度、温度、风力之间的定量关联关系。主要关注两个问题:(1)不同气象因素与 $PM_{2.5}$ 水平的关联关系如何,哪种因素是高浓度 $PM_{2.5}$ 形成的主要因素?(2)哪些气象因素的组合是高浓度 $PM_{2.5}$ 形成的最佳环境?

首先,对不同观测数据进行离散化,这是由于关联规则挖掘主要针对类别型数据。其中,温度取值范围为 −5～15 ℃,等间距分为 4 级;相对湿度取值范围为 0～100,等间距分为 5 级;风力按照气象领域常用分级,分为轻风、和风等 4 级;$PM_{2.5}$ 浓度参考环境空气质量标准,按空气质量指

数(air quality index, AQI)分为6级,不同要素具体离散化结果如表3所示。

表3 不同因子的离散化结果

Tab.3 Discretization of Different Factors

温度/℃	分级	湿度	分级	风力	分级	$PM_{2.5}$	分级
<0	1	0~20	1	0	1	0~50	1
0~5	2	20~40	2	1~2	2	50~100	2
5~10	3	40~60	3	3~4	3	100~150	3
>15	4	60~80	4	5~6	4	150~200	4
		80~100	5			200~300	5
						300~500	6

进而,分别采用本文算法及MOWCATL算法对上述序列进行关联规则挖掘,均以$PM_{2.5}$浓度观测序列作为后件序列。需要指的是,气象要素与空气质量变化可能是非同步的,即不同要素间相关关联,且存在一定时间滞后。滑动窗口大小的选择取决于前件对后件影响作用的持续时间,设置太大,则会削弱前件对后件的影响作用,从而有可能导致某些较弱规则的遗失。本文中认为温度等因素会对同一时刻及下一时刻的AQI产生影响,故窗口大小设置为2。以湿度和$PM_{2.5}$观测序列为例说明,不同算法挖掘结果分布列于表4和表5。

表4 MOWCATL算法挖掘所得规则

Tab.4 Association Rules Obtained by MOWCATL

规则编号	前件	后件	支持数	置信度(%)
1	{湿度>80}	{ $PM_{2.5}$>300 }	272	35
2	{60<湿度<80}	{ 200< $PM_{2.5}$<300}	210	27
3	{湿度>80}	{150< $PM_{2.5}$<200}	167	22
4	{60<湿度<80}	{150< $PM_{2.5}$<200}	176	19
5	{40<湿度<60}	{100< $PM_{2.5}$<150}	137	25
6	{60<湿度<80}	{ 100< $PM_{2.5}$<150}	203	22
7	{40<湿度<60}	{ 50< $PM_{2.5}$<100}	160	30
8	{湿度<20}	{$PM_{2.5}$<50}	83	93
9	{20<湿度<40}	{$PM_{2.5}$<50}	99	35

表5 本文算法挖掘所得规则

Tab.5 Association Rules Obtained by DRBARMS

规则编号	前件	后件	密度比	置信度(%)
1	{湿度>80 }	{ $PM_{2.5}$>300 }	2.11	35
2	{湿度>80}	{200< $PM_{2.5}$<300}	1.50	27
3	{湿度>80}	{ 150< $PM_{2.5}$<200}	1.29	22
4	{40<湿度<60}	{ 100< $PM_{2.5}$<150}	1.35	25
5	{40<湿度<60}	{ 50< $PM_{2.5}$<100}	1.88	30
6	{20<湿度<40}	{$PM_{2.5}$<50}	2.54	35

实验中,用MOWCATL算法挖掘共得到规则14条,从中选择9条(选择指标为支持数和置信度)。具体而言,当同一后件对应多种前件时,选择支持数和置信度最大的作为最终结果。可以发现,同一后件经常有不同的前件与之对应,且支持度和置信度之间并不一致。如规则8和规则9,若仅考虑支持数,较显著的规则应该是规则9;然而,规则8的置信度却远远高于规则9。正是由于所得规则中支持度和置信度之间的不一致性,从所得规则中选择最佳规则就成为一个难题。本文算法所得结果和MOWCATL所得结果并无冲突,同时,密度比指标和置信度指标间有着较好的一致性,因而在规则取舍和选择上更为简单。

进一步地,对于$PM_{2.5}$浓度和温度、风力等多因子间的关联,主要针对高浓度污染事件(AQI>150)进行挖掘,选取其中12条较显著的规则,列于表6。

综合表5、表6可知,湿度与高浓度$PM_{2.5}$形成的关联度最高,其次是风力,最后是温度。结合大气污染相关研究可知[19-21],大气颗粒物粒子主要有3种模态结构,爱根核模、积聚模和粗粒子模,其中积聚模在大气中停留周期最长,也是大气中最稳定的粒子。积聚模的来源主要是爱根核模的凝结以及大气化学反应所产生的各种气体分子转化成的二次颗粒物等。从化学组成上看,主要

表6 包含重度污染事件的关联规则

Tab.6 Association Rules Concerning High Level of Pollutants

编号	前件	后件	密度比	置信度(%)
1	{5<温度<10 }	{ $PM_{2.5}$>300 }	1.28	21
2	{5<温度<10}	{ 200< $PM_{2.5}$<300}	1.01	18
3	{0<温度<5}	{150< $PM_{2.5}$<200}	1.04	17
4	{风力=0 }	{ $PM_{2.5}$>300 }	1.35	22
5	{1<风力<2}	{ 200< $PM_{2.5}$<300}	1.24	23
6	{1<风力<2}	{150< $PM_{2.5}$<200}	1.04	17
7	{湿度>80,5<温度<10,1<风力<2 }	{ $PM_{2.5}$>300 }	2.54	42
8	{湿度>80,0<温度<5,1<风力<2 }	{ 200< $PM_{2.5}$<300}	1.49	27
9	{湿度>80,0<温度<5,2<风力<4 }	{150< $PM_{2.5}$<200}	1.48	25
10	{40<湿度<60, 1<风力<2}	{100< $PM_{2.5}$<150}	1.63	30
11	{40<湿度<60, 2<风力<4 }	{ 50< $PM_{2.5}$<100}	2.01	32
12	{20<湿度<40,0<温度<5,2<风力<4 }	{ $PM_{2.5}$<50}	2.83	40

含无机粒子 SO_4^{2-}、NO_3^-、NH_4^+ 和有机物 OC，而这些无机粒子主要由 SO_2、NO_x 等气体与水蒸汽的二次化学反应而来。因此，$PM_{2.5}$ 浓度与相对湿度呈现较强的关联。而风力是污染物粒子在空间传播和扩散的主要动力，同时也对污染物浓度稀释有着重要作用，故高浓度污染大多发生在风力较弱的环境。从所挖掘结果来看，温度并没有污染物浓度呈较强的关联。究其原因，可能是温度呈较强的季节特征，而本文主要针对冬季观测数据，整个季节中温度变化并不显著，因而并未与污染物浓度表现出较强的关联。同时，所得关联规则置信度偏低，这是由于这些气象因子只是高浓度污染形成的孕育环境，而并非污染物的直接来源。由实验结果可推测，高浓度 $PM_{2.5}$ 形成的最佳生成环境为高湿、低温和弱风环境，其关联程度由高到低依次为湿度、风力和温度。

3　结　语

本文提出了一种顾及背景知识的多事件序列关联规则挖掘算法。该算法将事件在整体序列中的分布特征作为背景知识，从相对视角定义了新的规则度量指标，其目的在于提升度量指标的自适应性，降低结果的参数依赖性。通过模拟实验和实例分析可以发现：(1)本文算法同时顾及序列中事件局部分布和整体分布特征，可较好地消除分布于序列中的高密度、非相关事件影响，从而发现数据有意义的真实关联，降低了规则中的冗余；(2)本文算法所挖掘结果与经典算法 MOWCATL 结果并无冲突，但规则重要性度量指标(密度比、置信度)间的一致性更好，从而使得规则筛选较为简单；(3)通过实例分析 $PM_{2.5}$ 浓度与温度等气象因子间的关联关系，结果显示：不同因子间的关联程度从高到低依次为湿度、风力和温度。需要指出的是，尽管本文算法允许前件和后件非同时发生，但算法假设前件对后件的影响是在一定时间范围内的持续作用，并不侧重于挖掘前件、后件间详细的滞后效应。此外，$PM_{2.5}$ 等污染物的浓度可能存在一定的季节差异，更详细的规则还需针对不同季节的观测进一步分析。

参　考　文　献

[1] Mennis J, Liu J W. Mining Association Rules in Spatio-Temporal Data: An Analysis of Urban Socioeconomic and Land Cover Change[J]. *Transactions in GIS*, 2005, 9(1): 5-17

[2] Sha Zongyao, Li Xiaolei. Algorithm of Mining Spatial Association Data under Spatially Heterogeneous Environment [J]. *Geomatics and Information Science of Wuhan University*, 2009, 34(12): 1 480-1 484 (沙宗尧,李晓雷. 异质环境下的空间关联规则挖掘[J]. 武汉大学学报·信息科学版, 2009,34(12): 1 480-1 484)

[3] Cai Siyue, Sui Fenzhen, Zhou Chenghu. Period Table Based Spatio-Temporal Association Rules Mining [J]. *Journal of Geo-Information Science*, 2011,13(4): 455-464 (柴思跃,苏奋振,周成虎. 基于周期表的时空关联规则挖掘方法与实验[J]. 地球信息科学学报, 2011,13(4): 455-464)

[4] Chen Jiangping, Huang Bingjian. Application and Effects of Data Spatial Autocorrelation on Association Rule Mining [J]. *Journal of Geo-Information Science*, 2011,13(1): 109-117 (陈江平,黄炳坚. 数据空间自相关性对关联规则的挖掘与实验分析[J]. 地球信息科学学报, 2011,13(1): 109-117)

[5] Feng L, Dillon T, Liu J. Inter-transactionalAssociation Rules for Multi-Dimensional Contexts for Prediction and Their Application to Studying Meteorological Data[J]. *Data and Knowledge Engineering*, 2001, 37(1): 85-115

[6] Agrawal R, Imieliński T, Swami A. Mining Association Rules Between Sets of Items in Large Databases[C]. ACM SIGMOD International Conference on Management of Data, Washington D C, 1993

[7] Agrawal R, Srikant R. MiningSequential Patterns [C]. The 6th International Conference on Data Engineering, Taipei, China, 1995

[8] Srikant R, Agrawal R. MiningSequential Patterns: Generalizations and performance improvements[M]. New York: Springer, 1996

[9] Zaki M J. SPADE: An Efficient Algorithm for Mining Frequent Sequences [J]. *Machine Learning*, 2001, 42(1/2): 31-60

[10] Pei J, Han J, Mortazavi-Asl B, et al. Prefixspan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth[C]. The 20th International Council for Open and Distance Education, Dusseldorf, Germany, 2001

[11] Mannila H, Toivonen H, Verkamo A I. Discovery ofFrequent Episodes in Event Sequences[J]. *Data Mining and Knowledge Discovery*, 1997, 1(3): 259-289

[12] Harms S K, Deogun J, Saquer J, et al. Discovering Representative Episodal Association Rules from Event Sequences Using Frequent Closed Episode Sets and Event Constraints[C]. IEEE International

Conference on Data Mining, San Jose, California, USA, 2001

[13] Harms S K, Deogun J, Tadesse T. Discovering Sequential Association Rules with Constraints and Time Lags in Multiple Sequences[M]. New York: Springer, 2002

[14] Tadesse T, Wilhite D A, Harms S K, et al. DroughtMonitoring Using Data Mining Techniques: A Case Study for Nebraska, USA[J]. *Natural Hazards*, 2004, 33(1): 137-159

[15] Shi Yan, Deng Min, Liu Qiliang, et al. Discovering Sequential Association Rules Between SingleOcean Climate Index and Land Abnormal Climate Events [J]. *Journal of Geo-Information Science*, 2014,16(2): 182-190 (石岩,邓敏,刘启亮,等. 海陆气候事件关联规则挖掘方法[J]. 地球信息科学学报, 2014,16(2): 182-190)

[16] Zhang X W, Su F Z, Shi Y, et al. Association Rule Mining Based on Spatio-Temporal Processes of Spatial Distribution Patterns[C]. The 4th International Conference on Wireless Communications, Networking and Mobile Computing, Dalian, 2008

[17] Yoo J S, Bow M. Mining Spatial Colocation Patterns: A Different Framework[J]. *Data Mining and Knowledge Discovery*, 2012, 24(1): 159-194

[18] Qian F, He Q, Chiew K, et al. Spatial Co-location Pattern Discovery Without Thresholds[J]. *Knowledge and Information Systems*, 2012, 33(2): 419-445

[19] Whitby K T. ThePhysical Characteristics of Sulfur Aerosols[J]. *Atmospheric Environment*, 1978, 12(1-3): 135-159

[20] Huang X, He L, Hu M, et al. AnnualVariation of Particulate Organic Compounds in PM2.5 in the Urban Atmosphere of Beijing[J]. *Atmospheric Environment*, 2006, 40(14): 2449-2458

[21] Song Y, Tang X, Xie S, et al. SourceApportionment of PM2.5 in Beijing in 2004[J]. *Journal of Hazardous Materials*, 2007, 146(1/2): 124-130

A Context-Based Association Rules Mining Method for Multiple Event Sequences

HE Zhanjun[1] *DENG Min*[1] *CAI Jiannan*[1] *LIU Qiliang*[1]

1 Department of Geo-informatics, Central South University, Changsha 410083, China

Abstract: Association rules mining of event sequences aims to discover interesting patterns of different neighboring events and plays an important role in understanding their mutual relationship. However, for most existing methods, the distribution characters of events in the sequences are usually ignored and selecting proper thresholds is really a tough task, which brings about the problems of redundant results or interesting rules missing. Thus, new measuring indexes were defined and a context-based method for multiple event sequences mining was proposed. Results of both the simulated experiment and practical cases emphasized that the proposed method could effectively reduce the redundancy in the results in comparison with the classic MOWCATL method. Moreover, there was good consistency between the measuring indexes, which eases the selection of generated rules. Finally, the proposed method was applied to mine association rules between and $PM_{2.5}$ concentration and several meteorological factors. Results indicated that the most associated meteorological factor with $PM_{2.5}$ concentration was the humidity and an eligible environment for high $PM_{2.5}$ concentration were high humidity, low temperature and weak winds.

Key words: context; data mining; association rules; multiple event sequences; $PM_{2.5}$

First author: HE Zhanjun, PhD candidate, specializes in spatial-temporal association patterns mining and applications. E-mail: hezhanjun000@126.com

Corresponding author: DENG Min, PhD, professor. E-mail: dengmin208@tom.com

Foundation support: The Hunan Provincial Science Fund for Distinguished Young Scholars, No.14JJ1007; The National Natural Science Foundation of China, No.41471385.