海陆气候事件关联规则挖掘方法

石 岩[1],邓 敏[1*],刘启亮[2],杨文涛[1]

(1. 中南大学地理信息系,长沙 410083; 2. 香港理工大学土地测量与地理信息资讯学系,香港红磡 999077)

摘要:近年来,异常气候事件的频发对人类的生活环境和经济发展带来严重负影响。气象学家研究表明:海洋气候异常对陆地气候异常事件的发生具有重要的诱发作用,因此,对海陆气候间的内在关联机制进行深入挖掘具有重要研究价值。本文提出了一种关联规则挖掘方法,以探索单一海洋气候指数与陆地异常气候事件间存在的关联。首先,针对陆地气候要素,采用顾及空间邻近关系的层次聚类方法进行有效气候分区,通过对各层分区结果进行相关统计分析得到有效的各区域气候序列;然后,进行顾及多重约束进行时序关联规则挖掘,以探索海陆气候要素间的关联机制;最后,通过实际算例分析得到的各气候指数与我国陆地区域异常降水事件间的关联机制结果,与实际情况高度吻合。

关键词:气候指数;异常气候事件;气候分区;时序关联规则挖掘

DOI:10.3724/SP.J.1047.2014.00182

1 引言

近年来,全球气候变化导致异常气候事件频发,对人类的生活环境和经济发展带来严重负影响,研究表明海洋气候异常对陆地气候异常事件的发生具有重要的诱发作用[1]。海陆气候数据通常以时间序列的形式记录海陆气候要素随时间的变化趋势,具有海量、多维、异质等特性,并隐含着大量未知的海陆气候关联模式[2]。许多学者为了发现海陆气候时间序列间隐藏的关联模式进行了大量研究,所提出的方法可大致分为两类:一类是特征值统计方法,该类方法通过对海陆气候数据进行主成分分析[3]、奇异值分解[4]等统计手段得到特征序列,并进一步挖掘海陆气候序列间隐藏的关联模式;另一类是数据挖掘方法[5],旨在从海量数据集中获取潜在的、有用的知识和模式,包括聚类分析[6-8]、关联规则挖掘[9-13]、数据建模预测分析[14-15]、混合[16-18]等内容。其中,时序关联规则挖掘技术可有效地从多维时间序列中发现事件间隐藏的关联模式,较有代表性的工作有:Mannilia等[11]对时间序列中的事件类型按照项集间发生的时间顺序进行了详细分类,进而引入时间窗口的概念计算事件发生频数,并提出WINEPI和MINEPI两个算法从时间序列中挖掘得到频繁事件集,以及事件间有效关联模式;Das等[12]对时间序列进行聚类分析,提取序列中的不同变化趋势,进而针对其变化趋势,利用时间延迟因子挖掘序列内和序列间隐藏的关联规则;Harms等[13]在传统频度和支持度约束基础上,分别对前件和后件施加时间延迟约束,提出MOWCATL算法,得到有效关联规则。

分析现有的研究工作可发现:(1)特征值统计方法对于海量数据稳定性不高,且对噪声敏感,此类方法现已很少使用;(2)数据挖掘方法虽然可弥补特征值统计方法存在的缺陷,但海量时空数据的自相关性导致大量冗余、无意义的规则出现。聚类技术可有效地顾及时空数据的相关性和异质性,并将海量数据集划分为若干有意义的簇,从而可有效压缩数据量,有利于进一步的分析工作。然而,现

收稿日期:2013-06-08;**修回日期**:2013-07-06.

基金项目:教育部新世纪优秀人才资助计划(NECT-10-0831);高等学校博士学科点专项科研基金项目(20110162110056);江苏省资源环境重点实验室开放基金项目(JS201101);现代工程测量国家测绘地理信息局重点实验室开放基金项目(TJES1102)。

作者简介:石 岩(1988-),男,山东济南人,博士生,主要从事时空数据挖掘及其应用的研究。E-mail:CSU_ShiY@126.com

*通讯作者:邓 敏(1974-),男,江西临川人,博士,教授,主要从事时空数据挖掘、推理与分析方面教学与研究工作。E-mail:dengmin028@yahoo.com

有聚类方法大都在单一尺度下进行,忽视了时空数据的尺度特征,使得聚类结果无法反映尺度变换过程中时空信息的特征渐变规律,由此所得结果的实际有效性难以判别;(3)现有的时序关联规则挖掘方法大都未充分融合气象领域相关知识,且缺乏相应约束条件,从而难以得到有意义的规则。例如,大多时序关联规则挖掘方法一方面未顾及应用背景,难以提取有意义事件;另一方面,虽然对时间窗口和时间延迟进行了必要约束,但并未对规则前件和后件之间的充分度和必要度施加必要限制条件,从而使得到的规则缺乏可信度。因此,本文结合层次聚类和时序关联规则挖掘思想,顾及多尺度效应,针对气象问题提出一种关联规则挖掘方法,以探索海洋气候要素对陆地异常气候事件的影响机制。

2　海陆气候关联规测挖掘的研究策略

鉴于气候时间序列数据具有海量、多维、相关、异质、多尺度等特性,特征值统计方法对海量数据中的噪声敏感,传统时序关联规则挖掘方法缺乏对此类数据的先验知识约束,不可避免地得到大量冗余和无效的规则,而结合聚类技术和时序关联规则挖掘的混合方法[17-18]则大都未顾及气候时间序列数据内在的多尺度特性,得到的聚类结果无法反映时空信息随尺度变化的渐变规律,进而得到的规则亦缺乏一定客观性。针对以上问题,为有效挖掘海陆气候间的关联特性,其研究策略为:首先,针对陆地气象要素采用顾及空间邻近关系的层次聚类方法进行多尺度气候分区,并通过统计量分析选取合适的分区结果作为进一步关联规则挖掘的对象;然后,采用顾及多约束的时序关联规则挖掘方法深入探析各海洋气候指数与陆地异常气候事件间的关联机制,进而抽象为知识。本文研究策略的理论框架如图1所示。

3　海陆气候单因子关联规则挖掘方法

3.1　基本概念

针对上述研究策略,首先给出几个基本定义[11-12]。

定义1——事件:给定事件类型ET,用$<A, t>$二元数组表示一个事件,其中,$A \in ET$,t为事件A发生时间。

定义2——事件序列:对给定事件类型ET上的

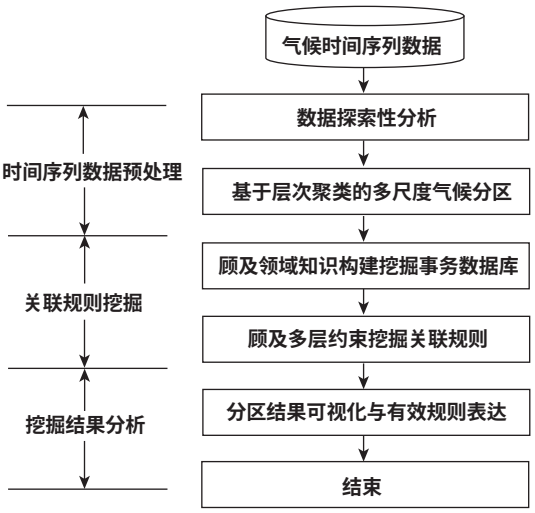


图1　海陆气候关联规则挖掘理论框架

Fig.1　The framework of association rules mining between ocean climate index and land abnormal climate events

事件序列ES用$<s, T_s, T_e>$三元数组表示,其中,s为一系列事件的有序集合,表示为:$s=<(A_1, t_1), (A_2, t_2), \cdots, (A_n, t_n)>$,其中,$A_i \in ET$;$t_i<=t_i+1$,$i=1, 2, \cdots, n-1$;$T_s$和$T_e$分别为$S$的起始时间和结束时间,且对于所有的$i=1, 2, \cdots, n-1$,满足$T_s \leqslant t_i < T_e$。

定义3——事件集:给定任一事件序列ES,ES上任意n个不同事件类型构成一个n元事件集$EP=<ET_1, ET_2, \cdots, ET_n>$,其中分别隶属于这$n$个事件类型的$n$个事件构成此事件集的项$EPI$。

定义4——时间窗口宽度:给定任意一个事件集的项,其中包含的所有事件发生的最早时间与最迟时间的时间差即为该事件集的时间窗口宽度win_width。

定义5——时序关联规则:给定一系列的事件集$EP=<ET_1, ET_2, \cdots, ET_n>$,若$EP_1$发生后,$EP_2$亦发生,即可描述为$EP_1=>EP_2$,并将这种模式称为时序关联规则$AR$,其中,$EP_1$为前件,$EP_2$为后件。并将所有分别隶属于$EP_1$和$EP_2$的项构成此规则的项$ARI$。

如图2所示,〈A, 24〉为一个事件;(D, 20), (C, 22),…, (D, 35), 20, 37>为事件序列;〈B, C〉为一个二元事件集,〈(B, 23), (C, 22)〉、〈(B, 23), (C, 34)〉、〈(B, 32), (C, 22)〉、〈(B, 32), (C, 34)〉为〈B, C〉的项;〈(B, 32), (C, 34)〉的时间窗口宽度为2;〈B, C〉=>〈<F, E〉为一个时序关联规则。

3.2　基于层次气候分区和气象背景的事件序列

现有的层次聚类方法[7-8]大都没有考虑实体间

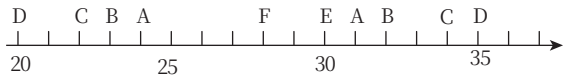


图2 事件序列

Fig.2 The sequence of events

的空间邻近特性，故难以较好地进行气候分区。层次聚类思想可以详细描述空间实体从完全离散到完全聚合的层次结构，地理现象的多尺度效应隐藏其中，本文借鉴文献[19]和文献[20]的思想，采用顾及空间邻域的层次聚类方法进行多尺度气候分区，另用相关统计量分析得到层次结构中的折点以获得层次变化中从量变到质变的转折，即有效的特征尺度分区结果。聚类分析方法仅适用于呈正态分布的数据集，然而，诸如降水等专题属性的数据往往呈现一种偏态分布，使得聚类分析手段无法直接对这类数据进行有效分区。为此，首先需对这类数据进行探索性分析，在保持其数据间相对关系稳定的前提下使其近似服从正态分布。统计学中，通常采用平方根法、取对数法进行预处理，本文选用平方根法，即对时间序列的每个数值采用其开根号的值代替原始值，可表达为：

$$Z_i' = \sqrt{Z_i} \tag{1}$$

式中，Z_i 为原始时间序列在时间点 i 的数值；Z_i' 为对原始值开根号后的数值，其中，$i \in [1, d]$，d 为时间序列的长度。经式(1)处理后，即可得到一个新的时间序列数据。为检验处理后的数据分布情况，采用Q-Q图对样本数据进行检验。若Q-Q图上的点近似在一条直线附近，则样本数据近似服从正态分布。图3为某一时间序列的原始序列和转换后序列的Q-Q检验图，可以发现原始序列呈曲线分布，而经平方根转换后的序列则近似在一条直线附近小幅度摆动。因此，处理后的数据近似服从正态分布，并可进一步对其进行聚类分析。

针对转换后的数据进行顾及空间邻域的层次气候分区，具体算法如下：

(1)对初始散点数据构造Delaunay三角网，进而借助文献[19]的策略对三角网施加整体和局部约束，从而精化每个点的空间邻域；

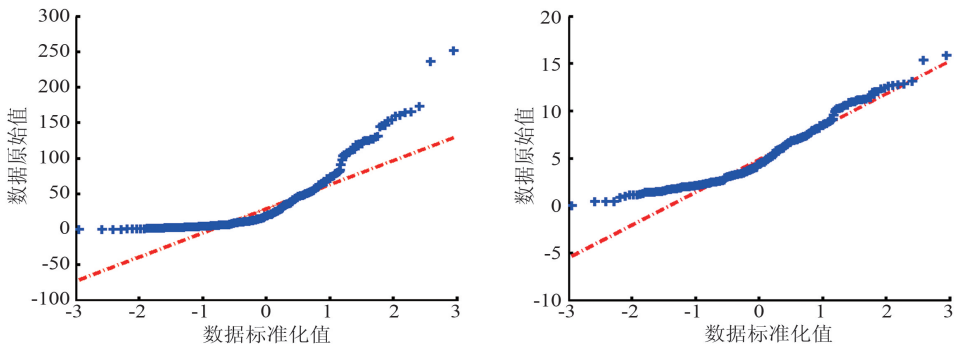
(2)针对每个实体点，用WARD法度量与其空间邻域点之间的相似性；

(3)对数据集中最相似的两个实体进行聚合成簇，用簇内所有实体属性均值作为簇的属性；

(4)用聚合成的簇作为新实体代替簇内实体，重复步骤(2)和(3)，直到所有点聚合为一个整体，得到层次树和每一层的聚合结果；

(5)从层次树中选择合适的区间结果进行伪T统计量分析，并从中选取合适的聚合结果。

海洋气候指数和陆地气候分区得到的是观测数据分析处理后的时间序列，仍然无法反映一些具体气象特征(如干旱、洪涝等)。在气象学领域，已有相关研究对气候指数、气象要素(降水、气温等)进行分类处理[21]，其中对SOI(Southern Oscillation Index)指数进行分类得到的结果列于表1。从表1中可看出，若SOI值在0.5和1之间的所有时间点均属于事件类型C。据此可对海洋气候指数和陆地气候分区结果进行离散化，从而得到具有明显气象特



(a) 原始数据的QQ图 (b) 平方根后数据的QQ图

图3 时间序列Q-Q图检验

Fig.3 The Q-Q diagram test of time series

表1 SOI指数分类

Tab.1 Classification of the SOI indices

事件类型	A	B	C	D	E	F	G
SOI	[1.5,+∞)	[1,1.5)	[0.5,1)	(−0.5,0.5)	(−1,−0.5]	(−1.5,−1]	(−∞,−1.5]

征的事件序列。

3.3 顾及多约束的时序关联规则挖掘

针对海洋气候指数和陆地气候分区经离散化得到的事件,本文采用顾及多约束的关联规则挖掘策略,有效挖掘海陆气候因子间的关联机制。下面阐述几个重要定义。

定义6——应用背景约束:给定任一事件序列*ES*,根据应用背景需要筛选出感兴趣事件,并称之为有效事件。例如,对于降水量时间序列,通常只保留降水异常多或异常少的事件。

定义7——时间窗口宽度约束:给定任一事件集*EP*的项*EPI*,若其时间窗口宽度$win_width \leq min_win$,则认为此项有效,否则视其为无效,其中,*min_win*为给定的最小时间窗口宽度阈值。

定义8——时间延迟约束:给定任一规则*AR*的一个项*ARI*,其前件和后件中最早事件发生时间分别记为t_s和t_s',最晚事件发生时间分别记为t_e和t_e',若$0<t_s'-t_s \leq time_lag$且$t_e'-t_e>0$,则认为*ARI*有效,否则视其为无效,其中,*time_lag*为给定最大时间延迟阈值。

定义9——充分度约束:给定任一规则*AR*,记其前件的有效项的数目为*n*,规则的有效项的数目为*m*,则需满足$m/n \geq min_Suf$,其中,*min_Suf*为给定的最小充分度阈值。

定义10——必要度约束:给定任一规则*AR*,记其后件的有效项的数目为n',规则的有效项的数目为m',则需满足$m'/n' \geq min_Nec$,其中,*min_Nec*为给定的最小必要度阈值。

基于上述相关定义,下面进一步给出顾及多约束的时序关联规则挖掘算法流程:

(1)针对相关应用背景需要和领域知识,从气候时间序列中提取感兴趣事件;

(2)参数初始化:*min_win*,*time_lag*,*min_Suf*,*min_Nec*;

(3)针对步骤(1)提取的有效事件,探索性的对其施加时间窗口宽度约束,从而得到一系列的有效前、后件事件集;

(4)根据步骤(3)得到的有效后件事件集的数目*n*,以及*min_Suf*和*min_Nec*,对有效前件事件集进行筛选,保留数目位于区间[n*min_Nec,n/min_Suf]的有效前件事件集,以减少无效规则的产生;

(5)根据步骤(3)和(4)处理得到的有效前后件事件集,进而对其施加时间延迟约束,以及充分度和必要度约束,提取有效的关联规则;

(6)根据相关领域知识对规则进行有效性分析,最终从规则里面提取出潜在的模式和知识。

综上所述,本文提出的关联规则挖掘方法一方面充分顾及了气候时间序列的特性和气象学背景需要,采用层次气候分区的策略从空间范围上对时间序列进行有效聚类划分,对具有高度相似气候特征的区域聚合,同时引入气象学已有知识对气候时间序列进行基于分类的事件提取。这分别从空间和时间上有效地构造了挖掘事务表,从而得到关联规则挖掘的空间和时间因子。另一方面,在已有时序关联规则挖掘策略基础上引入多个约束,从而可以避免海量无效规则的产生,有利于从规则里面提取模式和知识。下面,通过一个具体实例来验证分析本文所提方法的实用性。

4 算法的实例与验证分析

本文采用的海洋气候指数SOI、PDO和MEI来源于美国气候诊断中心,中国陆地区域降水数据来源于国家气象信息中心气象资料室,二者时间跨度均为1982年1月至2007年12月,时间粒度均为月,用于分析海洋气候指数与中国陆地降水之间的关联机制。其中:(1)SOI为南方涛动指数,用以反映厄尔尼诺现象活跃程度。PDD(Pacific Decadal Oscillation)为太平洋十年涛动指数,用于表征10年周期尺度变化的太平洋气候变化现象,特征为北纬20°以北太平洋区域表层海温异常。MEI(Multivariate ENSO Index)为多变量ENSO指数,具有与ENSO事件(厄尔尼诺和拉尼娜事件)更好的相关性。三者均为多维时间序列,如图4所示(其中,横坐标代表时间点,纵坐标代表相应的气候指数数值);(2)中国陆地区域降水数据原始空间分布为756个气象站点,对数据缺失严重和西部地区较少站点进行必要删除后,可用站点为554站,如图5所示。

采用顾及空间邻域的层次聚类方法对中国陆地区域降水进行气候分区,文献[22]证明簇的数目小于$\sqrt{N}$(N为站点数目)时是有效的,因此,对簇数小于$\sqrt{N}$的各层次结果进行伪T统计量分析,并绘制折线图,如图6所示(其中,横坐标代表该层聚类结果的簇数,纵坐标表示该层聚类结果对应的伪T统计量数值),可发现选取的有效层次结构中包含6

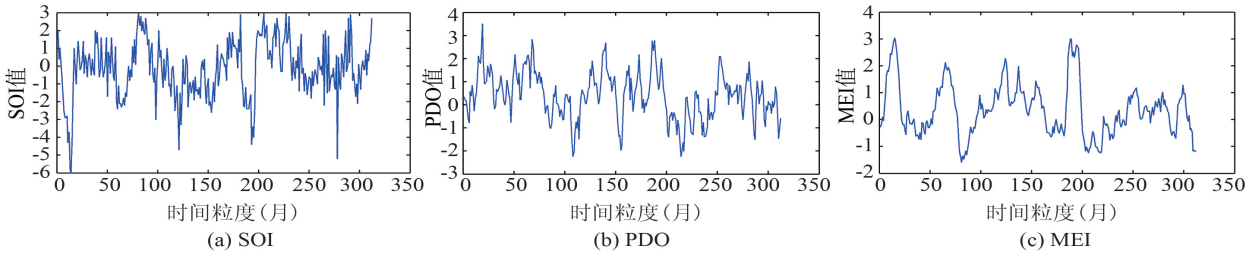


图4 海洋气候指数

Fig.4 Ocean climate indices

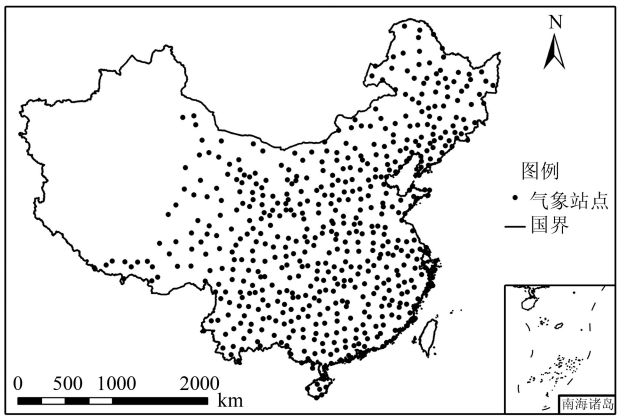


图5 中国陆地降水气象站分布

Fig.5 Distribution of land precipitation stations in China

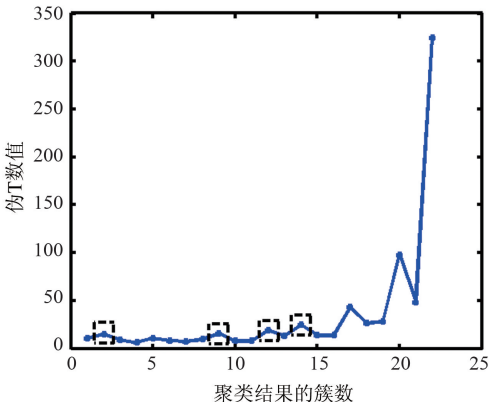
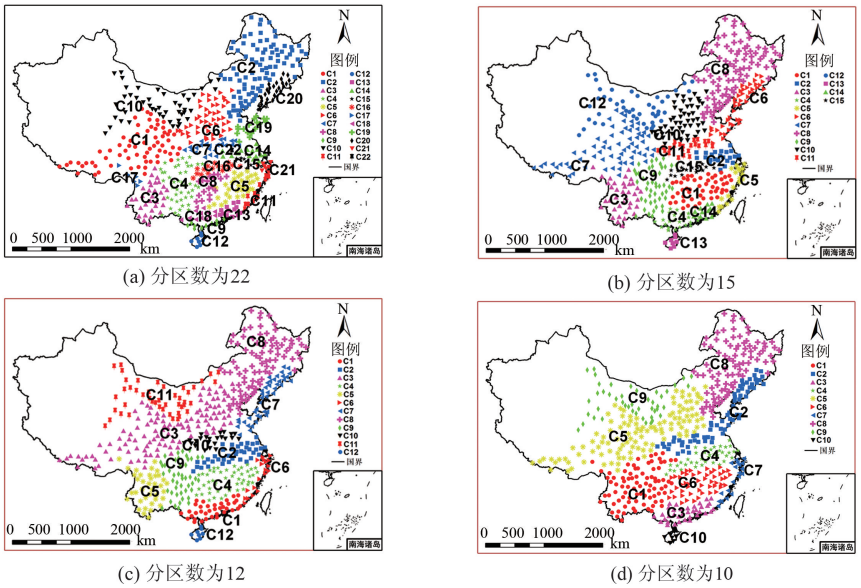


图6 伪T统计量

Fig.6 Pseudo-T statistic

个转折点，即层次聚类过程中在此6处发生了量变到质变的突变，因此，这6个聚类结果反映了6个特征尺度的有效分区。另外，尺度较大、簇数过少时说明已形成很多较大区域，并掩盖了某些细节和特殊模式，在这种情况下研究意义不大，为此忽略后面几处特征尺度分区结果，选择图6中前4处特征尺度对应的分区结果(图7)。由图7知：尺度较小、簇数过多使得某些区域之间仍然具有较大相似性，



(a) 分区数为22

(b) 分区数为15

(c) 分区数为12

(d) 分区数为10

图7 不同分区结果的提取

Fig.7 The extraction of different climate zones

故不能充分提取气候区域;尺度较大、簇数过少则掩盖了一些特征小区域,进而影响特殊关联模式的挖掘。因此,从4处特征尺度分区结果中进行折中,选择15个簇的分区结果作为下一步关联规则挖掘的对象。

由图7(b)可知:(1)从地理空间分布考虑,此尺度下的分区结果将我国降水区域自西向东大致分为西部、中部和沿海地区,而自北向南的区域划分也体现了我国降水的渐变特征;(2)从我国地形特征分布考虑,此分区结果体现了高原、盆地等不同地形对于降水的影响,例如,云南位于高原地区,使得此区域异于其他南方区域的多降水量,而时常发生干旱现象。基于以上两点,可证明此尺度下的分区结果完整地描述了我国不同区域的降水特征,因此,可作为下一步关联规则挖掘的对象。针对陆地降水分区得到的每个区域计算标准化降水指数(Standardized Precipitation Index,SPI),根据已有气象学研究对海洋气候指数和SPI指数的分类处理(表2、3),分别对各海洋气候指数和各陆地降水区域的SPI指数进行事件离散化,从而构建规则挖掘表,在

表2　气候指数离散化

Tab.2　The discretization of climate indices

事件类型	A	B	C	D	E	F	G
SOI	(−∞,−1.5]	(−1.5,−1]	(−1,−0.5]	(−0.5,0.5)	[0.5,1)	[1,1.5)	[1.5,+∞)
PDO	(−∞,−2]	(−2,−1.5]	(−1.5,−1]	(−1,1)	[1,1.5)	[1.5,2)	[2,+∞)
MEI	(−∞,−1.5]	(−1.5,−1]	(−1,−0.5]	(−0.5,0.5)	[0.5,1)	[1,1.5)	[1.5,+∞)

表3　标准化降水指数SPI分类

Tab.3　Classification of SPI indices

事件类型	极端干旱	中度干旱	轻度干旱	正常	轻度洪涝	重度洪涝	极端洪涝
SPI	(−∞,−2]	(−2,−1.5]	(−1.5,−1]	(−1,1)	[1,1.5)	[1.5,2)	[2,+∞)

此基础上,顾及多约束进行有效的关联规则挖掘。

本文仅针对干旱和洪涝事件,并从SPI指数中提取有效时间点,进而挖掘各气候指数对陆地异常降水事件间的影响机制。其中,关联规则挖掘前的各个阈值均是根据一定领域先验知识设置的,旨在进行探索性研究并试图发现未知知识。具体包括:前件时间窗口宽度阈值设置为6个月,后件采用时间尺度为3个月的SPI指数,时间延迟设置为6个月,充分度和必要度阈值设置为0.4,从而得到一系列的关联规则。以云南地区所在簇为例,得到的规则列于表4、5、6。对各区域挖掘得到的规则进行总结,并得到以下知识:

(1)SOI指数与我国陆地异常降水事件

① 仅与我国陆地区域轻度干旱和洪涝事件关联性较强;

② C(−1,−0.5]和D(−0.5,0.5)与我国大部分地区轻度异常降水事件关联较强;

③ A(−∞,-1.5]、B(−1.5,−1]和G[1.5,+∞)与我国部分地区轻度异常降水事件关联较强。

(2)PDO指数与我国陆地异常降水事件

① D(−1,1)和E[1,1.5)与我国大部分地区轻度异常降水事件关联较强;

② A(−∞,−2]与江苏、安徽地区重度干旱事件关联较强;

③ A(-∞,-2]、B(-2,-1.5]与广东、广西一带重度干旱事件关联较强;

④ C(−1.5,−1]和G[2,+∞)分别与内蒙、宁夏、山西一带中度干旱和洪涝事件关联较强;

⑤ G[2,+∞)与内蒙、新疆一带中度干旱事件关联较强;

⑥ C(−1.5,−1]与广东、福建一带中度干旱事件关联较强。

(3)MEI指数与我国陆地异常降水事件

① D(−0.5,0.5)和E[0.5,1)与我国大部分地区轻度异常降水事件关联较强;

② B(−1.5,−1]、C(−1,−0.5]和F[1,1.5)与我国部分地区轻度异常降水事件关联较强;

③ G[1.5,+∞)与云南地区中度干旱事件关联较强;

④ C(−1,−0.5]与山东、辽宁一带,以及长江湖北段区域中度干旱事件关联较强;

⑤ F[1,1.5)与四川地区、海南地区中度干旱事件关联较强。

此外,气象领域的已有知识表明,当SOI指数

表4 SOI指数与云南地区异常降水事件关联规则

Tab.4 Association rules between SOI and abnormal precipitation events in Yunnan Province

SOI指数	异常降水事件	充分度	必要度
(−∞,−1.5]	轻度干旱	0.45	0.82
(−1,−0.5]	轻度干旱	0.59	0.8
(−0.5,0.5)	轻度干旱	0.73	0.72
[1.5,+∞)	轻度干旱	0.49	0.71
(−1,−0.5],(−0.5,0.5)	轻度干旱	0.55	0.67
(−0.5,0.5),[1.5,+∞)	轻度干旱	0.49	0.56
(−∞,−1.5]	轻度洪涝	0.47	0.8
(−1,−0.5]	轻度洪涝	0.5	0.7
(−0.5,0.5)	轻度洪涝	0.82	0.72
[1.5,+∞)	轻度洪涝	0.45	0.59
(−1,−0.5],(−0.5,0.5)	轻度洪涝	0.58	0.61
(−0.5,0.5),[1.5,+∞)	轻度洪涝	0.44	0.56

表5 PDO指数与云南地区异常降水事件关联规则

Tab.5 Association rules between PDO and abnormal precipitation events in Yunnan Province

PDO指数	异常降水事件	充分度	必要度
[1,1.5)	轻度干旱	0.48	0.8
[1,1.5)	轻度洪涝	0.44	0.7

表6 MEI指数与云南地区异常降水事件关联规则

Tab.6 Association rules between MEI and abnormal precipitation events in Yunnan Province

MEI指数	异常降水事件	充分度	必要度
[1.5,+∞)	中度干旱	0.42	0.45
(−0.5,0.5)	轻度干旱	0.75	0.8
[0.5,1)	轻度干旱	0.45	0.82
(−0.5,0.5)	轻度洪涝	0.69	0.66
[0.5,1)	轻度洪涝	0.45	0.8

极大或极小时，容易造成我国降水的异常事件发生，本文得到的知识与此高度吻合，这亦在一定程度上证明本文方法是有效的。同时，本文方法还发现一些未知知识，尤其是PDO指数和MEI指数与我国某些区域的极端异常降水事件存在强关联，这可为气象学领域的深度研究，以及异常气候事件的预测提供必要依据。

5 结论与展望

本文针对异常气候事件发展了一种具有强针对性的的关联规则挖掘算法，通过实际算例证明了此方法具有良好的实用性。本文方法在2个方面具有一定优势：(1)充分顾及了多站点数据间存在的空间相关性，并基于气象数据的固有特征（如季节周期性等）对时间序列的自相关性进行了必要处理，以及气候分区预处理，有效地压缩数据并保留了原始数据重要特征；(2)专门针对气象领域问题，在挖掘关联规则的过程中引入多层约束处理，从而得到更加感兴趣的、有效的重要规则。

进一步的研究工作主要集中在3个方面：(1)对各气候指数联合分析，详细研究多因子与陆地气候事件间的关联机制；(2)采用相关技术从时间序列中直接提取有效事件，进而挖掘事件间的关联规则；(3)分别针对海洋和陆地气象要素进行必要分区，更加深入分析海陆气候关联机制。

参考文献：

[1] 姜世中.气象学与气候学[M].北京:科学出版社,2010.

[2] Tan P, Steinbach M, Kumar V, *et al*. Finding spatio-temporal patterns in earth science data[C]. Proceedings of KDD Workshop on Temporal Data Mining, San Francisco, U.S.A, 2001.

[3] Wold S. Principal component analysis[J]. Chemometrics and Intelligent Laboratory Systems, 1987, 2(1): 37-52.

[4] Klema V, Laub A. The singular value decomposition: Its computation and some applications[J]. IEEE Transactions on Automatic Control, 1980, 25(2): 164-176.

[5] Han J W, Kamber M. Data mining: Concepts and technique [M]. San Francisco: Morgan Kaufmann, 2005.

[6] 邓敏,刘启亮,李光强,等.空间聚类分析及应用[M].北京:科学出版社,2011.

[7] Fovell R, Fovell M. Climate zones of the conterminous United States defined using cluster analysis[J]. Journal of Climate, 1993,6(11),2103-2135.

[8] Fovell R. Consensus clustering of U.S. temperature and precipitation data[J]. Journal of Climate, 1997,10(6): 1405-1427.

[9] Agarwal R, Srikant R. Fast algorithms for mining association rules[C]. Proceeding of the 20th International Conference on Very Large Databases, 1994,487-499.

[10] Agarwal R, Srikant R. Mining sequential patterns[C]. Proceedings of the 11th International Conference on Data Engineering, 1995,3-14.

[11] Mannila H, Toivonen H, Verkanmo A. Discovery of frequent episodes in event sequences[J]. Data Mining and Knowledge Discovery, 1997,1(3):259-289.

[12] Das G, Lin K I, Mannila H, *et al*. Rule discovery from time series[C]. Proceedings of the 4th International Confer-

ence on Knowledge Discovery and Data Mining, New York, U.S.A, 1998, 16-22.

[13] Harms S K, Deogun J, Tadesse T. Discovering sequential association rules with constrains and time lags in multiple sequences[C]. Proceedings of the 2002 International Symposium on Methodologies for Intelligent Systems, Lyon, France, 2002,431-441.

[14] 王佳璆,邓敏,程涛,等.时空序列数据分析和建模[M].北京:科学出版社,2012.

[15] Cheng T, Wang J Q, Li X. A Hybrid framework for space-time modeling of environmental data[J]. Geographical Analysis, 2011,43(3):188-210.

[16] 孔令桥,秦昆,龙腾飞.利用二型模糊聚类进行全球海表温度数据挖掘[J].武汉大学学报(信息科学版),2012,37(2):215-219.

[17] Wu T, Song G, Ma X J, *et al*. Mining geographic episode association patterns of abnormal events in global earth science data[J]. Science in China Series E: Technological Sciences, 2008,51(1):155-164.

[18] Lin F, Jin X X, Hu C, *et al*. Discovery of teleconnections using data mining technologies in global climate datasets [J]. Data Science Journal, 2007,6(17):749-755.

[19] Deng M, Liu Q L, Cheng T, Shi Y. An adaptive spatial clustering algorithm based on Delaunay triangulation[J]. Computers, Environment and Urban Systems, 2011,35(4): 320-332.

[20] Guo D. Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP)[J]. International Journal of Geographical Information Science, 2008,22(7):801-823.

[21] Tadesee T, Wilhite D A, Harms S K, *et al*. Drought monitoring using data mining techniques: A case study for Nebraska, USA[J]. Natural Hazards, 2004,33(1):137-159.

[22] Bezdek J C, Nikhil R P. Some new indexes of cluster validity[J]. IEEE Transactions on Systems, Man, and Cybernetics-Part B, 1998,28(3):307-310.

Discovering Sequential Association Rules between Single Ocean Climate Index and Land Abnormal Climate Events

SHI Yan[1], DENG Min[1*], LIU Qiliang[2] and YANG Wentao[1]

(1. *Department of Geo-informatics, Central South University, Changsha* 410083, *China*; 2. *Department of Land Surveying and Geo-informatics, The Hong Kong Polytechnic University, Kowloon* 999077, *Hong Kong*，*China*)

Abstract: With the frequent occurrence of abnormal climatic events in recent years, social economic and people’s life are impacted more and more seriously. Meteorologists have found that ocean climate has an effect on land climate, such that the EI NINO can lead abnormal precipitation events on some land regions. Therefore, it is very critical to study the associations between ocean and land climate factors. At present, some researchers have done a series of work about this aspect and several representative methods have been proposed. The eigenvalue statistics and traditional sequential association rules mining are two main methods. However, the former is sensitive to noise and not suitable for huge amounts of data, while the latter dose not fully consider the correlation and multi-scale properties hidden in the climate time series data. In view of this, a method based on multi-constraints is proposed to discover sequential association rules between individual ocean and land climate factors in this paper. First, we took both time correlation and spatial correlation into account and a hierarchical clustering method with the consideration of spatial proximity is employed to find climate zones for the land climate factor. In this way, we not only preserve the effective information in the data, but also make the raw data simpler by removing time correlation and spatial correlation. Second, the land and ocean climate sequences are discretized based on domain knowledge and a series of events are also extracted. These are further used to construct the transactions mining table. Finally, a new method, which utilizes multiple constraints, is developed to mine sequential association rules. We only focus on the associations between ocean climate indices and abnormal land climate events, such as flood and drought. As a matter of fact, we need the frequent rules which can describe a law to a certain

extent. A practical example is used to explore the relationships between each climate index and unusual precipitation events in China, and the results obtained are very consistent with the actual situation. This to a large degree illustrates that the method proposed in this paper is rational. In addition, we also gain some unknown knowledge that can provide some information for meteorologists. Based on the information, meteorologists can study the internal mechanism deeply. Also, the information can guide the government to make related policy decisions. In summary, the method in this paper takes the spatial correlation, time correlation and multi-scale characteristics into account effectively, while considers multi-constraint to deal with climate problems more accurately. By experiments, it is proved that our method is correct and valid.

Key words: climate indices; unusual climate events; climate zone; sequential association rules mining

***Corresponding author:** DEND Min, E-mail: dengmin028@yahoo.com