文章编号:1672-3961(2009)03-0011-05

基于多粒度周期模式的时序离群点检测算法

罗玉盘,商琳*
(南京大学计算机软件新技术国家重点实验室,江苏 南京 210093)

摘要:目前研究时间序列离群点检测方法大都没有考虑到数据本身的周期性,有的只能处理名词性属性.针对实值性属性的时序数据,提出了多粒度周期模式的发现算法,该算法基于不同的时间间隔粒度来探测不同的周期模式,并利用得到的周期模式来发现那些偏离周期模式的离群点.该方法可有效避免将正常数据误报为异常值.通过实验验证了该算法既可正确找出数据在不同粒度下的周期模式,又可有效探测时序数据中的异常值,并与不用周期模式发现的离群点检测算法比较,减少了对特殊事件的离群点误报.

关键词:周期性分析;时间序列;粒度;离群点检测
中图分类号:TP181 **文献标志码**:A

Detect outliers in time series data with multi-granule periodic patterns

LUO Yu-pan, SHANG Lin*
(National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China)

Abstract: Contributions on outlier detection in time series data have seldom taken into account the data cyclical nature and numerical attributes values. An algorithm to find periodic patterns under different granularities was proposed, which could be used to detect outliers in time series data with numerical attributes. This method could avoid a false alarm, and experimental results showed that it could not only correctly identify multi-granule periodic patterns but also effectively detect outliers in data. Compared to outlier detection methods without periodic patterns, the results showed that it could reduce false alarms.

Key words: periodicity analysis; time series ; granularity; outlier detection

0 引言

时间序列数据(time series data)是由不同时间重复测量得到的值或事件的序列组成[1].这些值通常是在相等的时间间隔(例如:每小时,每天,每周)里测量.时间序列数据在现实生活中很常见,例如:股票市场中的价格数据、产品的周期销售量、气象数据等.在数据挖掘中,周期性分析在时序数据的研究中占重要地位.周期性分析(periodicity analysis)是周期模式(periodic pattern)的挖掘,即在时间相关序列数据中搜索重复出现的模式[1].同时,在许多重要的领域(例如:潮汐现象观测、日常电力消耗等)中,数据都呈现特定的周期模式.如果准确地把握了周期模式,就可以了解产生时间序列数据的机制,这将有利于预测数据或者监测数据异常.

离群点检测是数据挖掘中的4类知识发现任务之一,其目的在于发现与其他数据的一般行为或模型不一致的数据对象,也就是离群点[1].离群点可能是测量或记录时的误差所导致的,也可能是特殊数据行为的表现.许多数据挖掘算法力图使离群点的影响最小化,或者排除他们.然而,这可能导致丢失重要的潜在信息,因为离群点本身可能表示了数据的某种特殊行为模式.对时序数据进行离群点检测在许多领域中都有现实意义[2-3].在现实生活中,存在着各种包含周期性事件的时间序列数据,例如

收稿日期:2009-05-20
基金项目:国家自然科学基金资助项目(60775046)
作者简介:罗玉盘(1987-),女,江西吉安人,硕士研究生,主要从事数据挖掘研究.E-mail: luoyupan@ai.nju.edu.cn
*通讯作者:商琳(1973-),女,河北曲阳人,副教授,博士,主要从事数据挖掘、粒计算、粗糙集等研究.E-mail: shanglin@nju.edu.cn

心电图[4]、太阳黑子[5]等. 如果能对这些时序数据进行正确的周期性分析,便可以有效地帮助人们发现异常事件,例如心率异常,黑子爆发等. 基于周期模式分析的时序离群点检测可以引导人们去研究离群点产生的原因,有利于及时解决与应对突发或异常事件,例如抢救病人、探寻黑子爆发原因等. 目前在时序离群点检测方法中,有的使用统计学方法[6-7];有的使用聚类方法[8],但这些方法大都没有考虑到时序数据本身的周期性. 在针对周期性时序数据研究的方法中,处理的大都是名词性属性的序列数据[9-10],无法对实值性属性的时序数据进行处理. 针对实值性属性的时序数据,提出了多粒度周期模式的发现算法,该算法基于不同的时间间隔粒度来探测不同的周期模式,并利用得到的周期模式来发现那些偏离周期模式的离群点,该算法可以避免将此类正常数据误报为异常值.通过实验验证了算法探测周期模式的正确性以及基于多粒度周期模式的时序离群点检测算法的有效性.

1 多粒度周期模式的时序离群点检测算法

与人类生活相关的时序数据带有明显的周期性,最有代表性的就是每一重复周期的数据都具有类似的行为模式,如马路上的车流量. 于是,可以将周(week)作为周期模式分析的基本粒度. 但是还存在这样的一些数据,对于周作为时间粒度来分析,它是异常的,但从另外的时间粒度来看,如每一年作为周期模式分析的基本粒度,它的出现是正常的,如圣诞节前商店销售额的猛涨. 相比平常的销售情况,这肯定会被认为是一个离群点,但从多年的数据观察,这种猛涨每年固定时间就会发生,其实是一个正常的现象. 相反,如果销售额在圣诞节前没有猛涨,这时就应该认为是离群点,要加以分析原因,思考相应的对策. 因此,可以认为这些时序数据符合不同的周期模式,一种是以周为单位的周期模式,它符合数据中的大部分;另一种是以年为单位的个别周期模式,它符合数据中的特定部分. 找出不同粒度下的周期模式,便可利用它们找出时序数据中的离群点,也就是那些偏离周期模式的数据对象.

提出的算法主要包括三个步骤(其伪代码如图1所示)(以周和年作为不同的时间粒度):

第一步,找出每周的数据行为模式. 将训练数据按周一到周日划分,分别对它们进行聚类(这是为了除去个别异常值对模式判断的影响). 根据聚类结果中最多的一类,也就是正常值所在的那一类,来判定数据的模型. 假设数据将服从两种分布之一:均匀分布或正态分布.

```
Input: training dataset D1, testing dataset D2
Output: weekly pattern P1, yearly pattern P2, outliers set O

//Step 1: find P1
Partition D1 into 7 sets: {Mon., Tue., …, Sun.};
For(every sets in D1)
K-means(3);//clustering to 3 parts
Model(max cluster)//model with the max cluster
  {
If(values are all equal)
P1.add(0,value,0);
    //the parameters represent model type, expectation and standard deviation
Else
Calculate the expectation μ and deviation σ²;
P1.add(1,μ,σ);//model type = 1, represent Gaussian distribution
  }
//Step 2: find P2
Find all outlying days OutDays in D1 with P1;
PeriodicPattern(OutDays)//find yearly pattern in OutDays
{
  Partition OutDays into several sets with the same month and day;
  For(every sets in OutDays)
Model(the set);
  If(there is a cycle)
P2.add(cycle,date,model,solar);
//the parameters represent the cycle, starting time, data model and calendar type
}
Switch OutDays to lunar calendar;
PeriodicPattern(OutDays);//the forth parameter is lunar now
//Step 3: detect outliers in D2
For(every record t in D2)
If(t does not meet neither P1 nor P2)
O.add(t);//t is an outlier
```

图1 算法的伪代码

Fig.1 The pseudocode of algorithm

第二步,找出特殊的周期模式. 根据上一步得到的星期模式,找出训练数据中所有的异常值,然后将这些异常值按照同月同日划分,并分别求出该天各时段的数据分布模型. 接着查找符合这个模型的异常值中,是否存在固定的规律(例如每年此时都发生同类事件),如果存在,这就是一类特殊的周期模式,将其记录下来.

第三步,利用上面得到的周期模式探测离群点. 先判断数据是否符合星期模式,如果偏离,再看这一天是否有特殊的周期模式存在,若没有,则认定为离群点.

2　实验

实验运行在 Pentium(R) D CPU 2.66 GHz 的计算机中,其操作系统为 Windows XP,内存为 1G. 所有程序由 java 语言实现,并运行在 jrel 1.6.0_03 上,同时引用了 Weka 3-5 中的相关包. 实验所用的 3 个数据集均由人工设计生成.

实验中的数据集有 6 个数值型属性,分别为:年(year)、月(month)、日(day)、小时(hour)、星期(week)、值(value),其中星期的取值分别表示周一到周日. 在第一步中,选用 K-means 聚类算法(簇的均值作为簇中心),k(簇的个数)取值为 3,因为训练数据一般可以分为正常值,偏高值与偏低值. 在判断数据是否符合正态分布时,选取阈值为 4 * sigma(标准差),因为同一正态分布的数据与期望偏离小于 4 倍标准差的概率大于0.999 9. 如果与此正态分布的期望偏离大于该阈值,我们可以认为它们不是来自同一数据模式.

2.1　实验一

本实验使用的数据集表示的是学校教学楼里的人数,其构造方法如表 1 所示. 这是结合实际情况设计的,在 6 点前教学楼不开门,因此人数为 0,正常工作日的人数比周末的明显要多,而且平常上课在上下午各有一个人数高峰期. 表中的 $N(\mu,\sigma)$表示数据满足以 μ 为期望,以 σ 为标准差的正态分布.

表 1　教学楼人数数据集的构造方式

Table 1　The description of dataset people-in-teaching-building

时间段	周六、周日	周一至周五
[0,5]	0	0
[6,7]	$N(5,1)$	$N(20,1)$
[8,11]	$N(30,5)$	$N(200,10)$
[12,14]	$N(10,1)$	$N(20,1)$
[15,17]	$N(50,5)$	$N(200,10)$
[18,21]	$N(20,2)$	$N(100,5)$
[22,23]	$N(5,1)$	$N(20,1)$

同时,假定每年的 1 月和 8 月分别为寒假和暑假,以及“五一”3 d 和“十一”7 d 假期教学楼都不开放,即人数为 0,这将作为特殊的周期模式设计.

在实验中,生成了 60 a 的数据,其中前 50 a 作为训练数据(1948 年 1 月 1 日至 1997 年 12 月 31 日),后 10 a 作为测试数据(1998 年 1 月 1 日至 2007 年 12 月 31 日). 表 2 和表 3 分别给出了算法从训练数据中得出的星期模式与特殊周期模式. 表 2 显示了,以周日和周一前 12 h 为代表的数据分布情况,其中分布类型 0 代表均匀分布,1 代表正态分布. 由表 2 可知,实验结果与构造函数非常接近,即准确地计算出了数据的分布情况. 表 3 以 1 月 1 日为例说明了特殊模式的表示形式,周期即事件发生的固定间隔,模型即这一天的数据分布. 实验中,算法将预先假定的特殊周期模式全部都找出,即每年的假期教学楼里的人数均为 0.

表 2　教学楼人数数据集的星期模式表示

Table 2　The weekly pattern of dataset people-in-teaching-building

时间	周日数据分布			周一数据分布		
	分布类型	期望	方差	分布类型	期望	方差
0:00	0	0	0	0	0	0
1:00	0	0	0	0	0	0
2:00	0	0	0	0	0	0
3:00	0	0	0	0	0	0
4:00	0	0	0	0	0	0
5:00	0	0	0	0	0	0
6:00	1	4.532 951	1.044 064	1	19.496 89	1.102 882
7:00	1	4.513 372	1.098 483	1	19.489 25	1.08 748
8:00	1	29.534 86	25.395 58	1	199.434 3	100.699
9:00	1	29.464 66	25.431 86	1	199.320 1	96.673 77
10:00	1	29.623 69	23.218 09	1	199.399 9	101.505 9
11:00	1	29.545 85	25.252 32	1	199.268 5	104.719 5

表 3　教学楼人数数据集的特殊周期模式表示

Table 3　The periodic pattern of dataset people-in-teaching-building

周期/a	代表日期	模型		
		分布类型	期望	方差
1	1948.1.1	0	0.0	0.0

表 4 给出了对测试数据进行离群点探测的实验结果,并且对比不带特殊周期模式和带特殊周期模式的结果,前者将有规律的特殊事件作为离群点,而后者正确判断了并没有离群点存在. 因此,特殊周期模式能够减少对此类事件的离群点误报.

表 4　实验一结果对比

Table 4　The comparison of experiment results

不带特殊周期模式的日期	带特殊周期模式的日期
1998-01-01	—
1998-01-02	—
1998-01-03	—
1998-01-04	—
1998-01-05	—
1998-01-06	—
…	…

实验一说明本算法能准确计算出星期模式及特殊周期模式,并且基于这两种模式可以有效避免对有规律的特殊事件进行误报.

2.2　实验二

本实验中使用的数据集为超市销售额数据,它的构造方式如表5所示. 假设超市每天都9点开门,24点关门. 一个常见的规律就是销售额周末多于正常工作日,晚上多于下午,下午多于上午. 基于星期模式,加入了一个阳历下的特殊周期模式:每年圣诞节前10 d(12月15日~12月24日)超市的销售额全天都增长到 $N(2000,5)$ 的分布,当然从开门后开始;同时加入一个阴历下的特殊周期模式:每年春节前半个月(阴历12月15日~12月29日)超市的销售额全天也都增长到 $N(2000,5)$ 的分布. 在中国,传统节日都是阴历记法,因此很多事件虽然是固定模式,但从阳历看来却是在不同时间发生. 于是,算法中加入了阴历与阳历之间互相转换的程序,这样既可以查找阳历时间中的周期模式,又可以查找出阴历时间的周期模式. 实验结果表明算法可以有效探测阴历的特殊周期模式.

表5 超市销售额数据集构造方式

Table 5 The description of dataset sale-of-supermarket

时间段	周六、周日	周一至周五
[0,8]	0	0
[9,14]	$N(50,5)$	$N(50,5)$
[15,19]	$N(500,5)$	$N(200,5)$
[20,23]	$N(1000,5)$	$N(500,5)$

生成了25年的数据,其中前20年(1985年1月1日至2004年12月31日)作为训练数据,后5年(2005年1月1日至2009年12月31日)作为测试数据. 表6同样以周日和周一的数据模型为例,给出了算法对星期模式的计算结果.由表6可知,其计算结果与构造函数相差无几,同样准确的描述了数据分布情况.

表6 超市销售额数据集的星期模式表示

Table 6 The weekly pattern of dataset sale-of-supermarket

时间	周日数据分布			周一数据分布		
	分布类型	期望	方差	分布类型	期望	方差
0:0	0	0	0	0	0	0
1:0	0	0	0	0	0	0
2:0	0	0	0	0	0	0
3:0	0	0	0	0	0	0
4:0	0	0	0	0	0	0
5:0	0	0	0	0	0	0
6:0	0	0	0	0	0	0
7:0	0	0	0	0	0	0
8:0	0	0	0	0	0	0
9:0	1	50.118 01	24.851 93	1	50.081 34	24.623 25
10:0	1	49.994 06	24.188 66	1	49.886 05	27.317 59
11:0	1	49.777 49	24.289 19	1	50.180 02	24.157 45

表7将特殊周期模式在阳历和阴历时间下各举了一个例子(为节省空间,类似分布的时间段用一个模型表示),算法将加入的特殊模式都找到了,而且周期和分布类型全都正确. 表8以2005年为例给出了3种带不同特殊周期模式的结果,可以看出当加入了阴历周期模式,便不再存在任何误报问题.

表7 超市销售额数据集的特殊周期模式表示

Table 7 The periodic pattern of dataset sale-of-supermarket

时间系统	周期/a	代表日期	模型			
			时间	分布类型	期望	方差
阳历	1	1985-12-15	[0,8]	0	0	0
			[9,23]	1	2 000.945	31.833 57
			…			
		1985-12-24	[0,8]	0	0	0
			[9,23]	1	1 998.861	19.785 83
阴历	1	1984-12-15	[0,8]	0	0	0
			[9,23]	1	2 000.495	26.786 49
			…			
		1984-12-29	[0,8]	0	0	0
			[9,23]	1	2 000.689	29.281 15

表8 实验二结果对比

Table 8 The comparison of experiment results

不带特殊周期模式的日期	带阳历周期模式的日期	带阳历和阴历周期模式的日期
2005-01-24	2005-01-24	—
2005-01-25	2005-01-25	—
…	…	…
2005-02-06	2005-02-06	—
2005-02-07	2005-02-07	—
2005-12-15	—	—
2005-12-16	—	—
…	…	…
2005-12-23	—	—
2005-12-24	—	—
…	…	…

2.3 实验三

为了验证算法的探测离群点能力,在实验二的数据集人为加入了一些离群点. 考虑金融危机、经济下滑等原因,假设2008和2009年的圣诞节和春节的购物高峰下滑为 $N(1000,5)$,这相对于正常情况来说就是离群点. 利用算法得到的离群点在表9中列出,可以看出该算法有效的找出了所有离群点.

因此,实验三说明了本算法具有探测真正离群点的能力.

表 9　实验三的离群点探测结果
Table 9　The experimental results of outlier detection

日期	值
2008-01-23,星期 4,9 点	999.940 7
…	…
2008-01-23,星期 4,23 点	1 001.698
…	…
2008-02-06,星期 4,23 点	998.6133
2008-12-15,星期 2,9 点	1 003.136
…	…
2008-12-15,星期 2,23 点	996.430 9
…	…
2008-12-24,星期 4,23 点	997.1
…	…

3　结束语

在与日常生活密切相关的众多数据集中,通常存在一些周期性规律. 为了不将此类正常数据误报为异常值,提出了基于多粒度周期模式的时序离群点检测算法,并在 3 个人工数据集上进行实验,验证了算法探测周期模式的正确性以及离群点探测的有效性. 进一步的工作将使用增量学习的方法来改进算法,以便及时调整周期模式.

参考文献:

[1] HAN J W, KAMBER M. Data mining concepts and techniques [M]. 2nd ed. Beijing: China Machine Press, 2007.

[2] BASU S, MECKESHEIMER M. Automatic outlier detection for time series: an application to sensor data[J]. Knowledge and Information Systems, 2007, 11:137-154.

[3] FEBRERO M, GALEANO P, GONZAÁLEZ-MANTEIGA W. A functional analysis of NOx levels: location and scale estimation and outlier detection[J]. Computational Statistics, 2007, 22:411-427.

[4] BOUCHEHAM B. Matching of quasi-periodic time series patterns by exchange of block-sorting signatures[J]. Pattern Recognition Letters (PRL), 2008, 29(4):501-514.

[5] GERSCH W, KITAQAWA G. Systems and computers, pacific grove, CA, 1995[C]// Smoothness Priors Analysis of Quasi-periodic Time Series: Proceedings of the 29th Asilomar Conference on Signals. Washington: IEEE Computer Society, c1995.

[6] ABRAHAM B, BOX G E P. Bayesian analysis of some outlier problems in time series[J]. Biometrika, 1979, 66(2):229-236.

[7] ABRAHAM B, CHUANG A. Outlier detection and time series modeling[J]. Technometrics, 1989, 31(2):241-248.

[8] BLENDER R, FRAEDRICH K, LUNKEIT F. Identification of cyclone-track regimes in the north atlantic[J]. Quarterly Journal of the Royal Meteorological Society, 1997, 123(539):727-741.

[9] MA S, HELLERSTEIN J L. Mining partially periodic event patterns with unknown periods[J]. International Conference on Data Engineering, 2001, 205-214.

[10] YANG J, WANG W, YU P S. Mining asynchronous periodic patterns in time series data[J]. IEEE Transactions on Knowledge and Data Engineering, 2003, 15:613-628.

(编辑:陈燕)