中国农业大学学报 2001, 6(4): 76～80
Journal of China Agricultural University

时态数据库周期规律与关联规则的挖掘

陶 兰① 唐玉荣
(中国农业大学电气信息学院)

摘 要 提出了一种简单有效、抗干扰的周期规律挖掘算法; 研究了关联规则提取过程中的连续属性离散化, 并用Apriori算法发现有效的规则。对电信话务量时态数据库的挖掘测试结果表明, 该算法实现较简单, 执行效率较高, 具有实用性和有效性。

关键词 知识发现; 时态数据库; 周期; 关联规则
中图分类号 TP 311

Study on Mining Periodicity and Association Rules From Temporal Database

Tao Lan Tang Yurong
(College of Electricity and Information, CAU)

Abstract The method for mining the periodicity and association rules from temporal database is presented. To discover periodicity, a simple, efficient and anti-jamming algorithm was proposed. A method for discretization in the processing of mining association rules and Apriori algorithm to discover efficient rules was presented. The results on a group of traffic data showed that the method is useful and efficient.

Key words KDD; temporal database; periodicity; association rule

知识发现(Knowledge Discovery in Database, KDD)亦称为数据挖掘(Data Mining, DM), 是由人工智能、模式识别、统计学、数据库、数据可视化等众多学科分支相互交叉、融合所形成的一个新兴的且具有广阔应用前景的研究领域。KDD 试图从大量数据中发现有效的、新颖的、潜在有用的以及最终可理解的知识。其挖掘对象有关系数据库、面向对象数据库、空间数据库、时态数据库、多媒体数据库、异质数据库、遗产数据库以及环球网WWW。

众多实际领域(如气象、电信、金融等)的数据都是时态数据(即数据具有时间相关性), 存储时态信息的数据库称为时态数据库。对时态数据库中可能存在的周期性和特征属性的相关性进行挖掘, 具有广泛的实用意义。笔者以电信部门的长途话务量时态数据库为研究对象, 对其进行了周期规律和关联规则挖掘的实验研究。

1 周期规律挖掘

实际领域中的许多时态数据具有一定的周期性, 但并非严格的数学意义上的周期现象。它

收稿日期: 2000-12-26
①陶 兰, 北京清华东路 17 号 中国农业大学(东校区) 64 信箱, 100083

们可能在时间上发生了不规则的伸缩, 在幅度上迭加了干扰信号。话务量(电信业务流量的简称, 也称为电信负载量)属于具有这类特性的数据。传统上, 对话务量的研究采用基于 Poisson 过程的流量模型, 没有对其所隐含的内在规律进行研究, 而挖掘话务量数据的周期规律及特征规则, 可为调整各个中继电路群组织提供所必需的基础资料。

1.1　周期规律挖掘算法

Step1　数据预处理。预处理的目标是使处理后的数据更完整、更合理, 有利于提高挖掘效率, 得出更有意义的挖掘结果。本文中采用了缺值处理、有效属性的选择变换、清理和过滤记录等处理方法。

Step2　趋势挖掘。时态数据库的数据序列中存在着一个隐含的变化模式, 实际数据可看作该变化模式和随机干扰的迭加, 通过趋势挖掘可以使该变化模式同随机干扰区别开来。本文中采用滑动平均法[1]消除随机干扰, 使数据呈现大致的趋势, 然后对趋势进行形式化描述, 为下一步提取周期提供基础。滑动平均法简述如下:

记原始数列为$\{x(t)\}$, $t=1,2,\cdots,n$。中间点(即 $1<t<n$ 时)滑动平均值计算公式为

$$Y(t)=[x(t-1)+2x(t)+x(t+1)]/4$$

两端点的计算公式为

$$Y(1)=[3x(1)+x(2)]/4 \quad Y(n)=[x(n-1)+3x(n)]/4$$

图 1 示出平滑处理前后一周的话务量数据曲线。横坐标序号每点为一个时间段(1 h), 6 个点为一天。新数据序列的方差小于原始非负序列的方差, 从而新序列的随机性弱于原始序列。可以看出, 平滑处理后的数据点具有峰、谷、上升、下降和边界(起点和终点为边界)等 5 种状态, 因此可对每一个数据赋予一个状态值。具体方法如下, 其中 t2 为当前数据值, t1 为 t2 前的值, t3 为 t2 后的值。

```
If t2>=t1 And t2>=t3          Then tendency="peak"
  Else If t2<=t1 And t2<=t3   Then tendency="valley"
    Else If t2>t1 And t2<t3   Then tendency="up"
      Else If t2<t1 And t2>t3 Then tendency="down"
```

得到指定时刻的势态值 peak(峰)、valley(谷)、up(上升)、down(下降)、edge(边界)等, 从而获得属性的趋势信息。

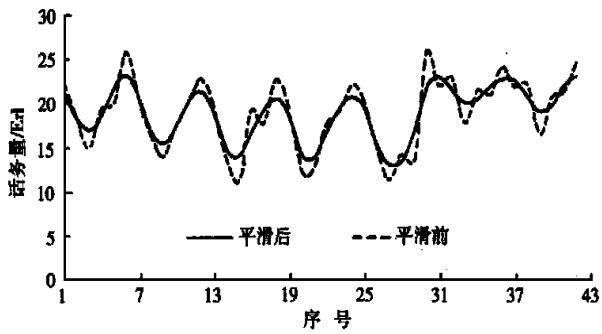

图 1　平滑处理前后的话务量曲线

Step3 周期获取。利用属性的趋势信息,用峰谷统计分析法[2]把属性趋势中值为peak 或valley 的记录挑选出来,形成一条峰谷链;构造 k 跨度数组和 k 跨距统计数组;k 跨距统计数组中离散差 $S_{k,d}$ 最小的平均值 $S_{k,av}$ 即为所求的周期。

k 跨度数组:设 $k>1$, s_{j1}, s_{j2}, …, s_{jk} (若峰谷链长度为 l,则 $1<j<z$, z 表示由 l/k 得到的向上舍入为最接近的整数)为峰谷链中连续的 k 个峰点(或谷点),时间差 $s_{jk}t-s_{j1}t$ 称为从 $j1$ 开始的 k 峰距(谷距),k 峰距或 k 谷距称为 k 跨距,所有 k 跨距按起点为序排成的数组称为 k 跨度数组,记为 S_k,当 k 为 0 和 1 时,规定 $S_k=\{0\}$

k 跨距统计数组:对固定的 k,记 $S_{k,max}=\max S_k$, $S_{k,min}=\min S_k$, $S_{k,av}$ 等于对 S_k 取平均, $S_{k,d}=\max S_k-\min S_k$。4 元组 $(S_{k,max}, S_{k,min}, S_{k,av}, S_{k,d})$ 称为 k 跨距统计记录。

1.2 周期规律挖掘实验

以某电信部门的长途话务量作为实验数据,对上述算法进行了测试。每日采集 6 个时间段的数据,每个时段历时 1 h,共收集了 4 个月的数据。用峰谷统计分析法得出的分析结果见表 1。注意到 $k=2$ 时有最小的离散差,此时平均值 $S_{k,av}=5.982$,向上取整数 6,本文中所用为 1 d 中 6 个时间段的数据,所以周期为 1 d。

表 1 峰谷统计分析结果

k	$S_{k,max}$	$S_{k,min}$	$S_{k,av}$	$S_{k,d}$
2	7	5	5.982	2
3	12	8	9.018	4
4	14	11	12.019	3
⋮	⋮	⋮	⋮	⋮

2 关联规则挖掘

在实际生活中,由于某些事件的发生而会引起另外一些事件的发生,从数据库中采掘出具有这种形式的规则就是挖掘关联规则。它在决策支持系统和智能信息系统等各个方面起着重要的作用。如果一条关联规则的发生次数(即支持度和可信度同时超过指定值的次数)达到用户指定的比率(最小支持度和最小可信度),则称此规则是一条有效的“规则”。

2.1 关联规则的形式定义[2]

令 $I=\{i_1, i_2, \cdots, i_m\}$ 为项目集(item set),D 为事务数据库,其中每个事务 T 是一个项目子集($T\subseteq I$),并具有一个唯一的标识符 ID。关联规则是形如 $X\Rightarrow Y$ 的逻辑蕴含式,其中 $X\subset I$, $Y\subset I$,且 $X\cap Y=\varnothing$ 有 2 个因子与规则相关:如果事务数据库中有 $s\%$ 的事务包含 $X\cup Y$,称关联规则 $X\Rightarrow Y$ 的支持度(support)为 s;如果事务数据库中包含 X 的事务中有 $c\%$ 的事务同时也包含 Y,称关联规则 $X\Rightarrow Y$ 的可信度(confidence)为 c。

2.2 关联规则挖掘算法

step1 连续属性值离散化。从关联规则的定义可知,可将挖掘关联规则看作是在一个所有属性均为布尔类型的关系表中寻找“1(T)”值之间的关联。在关系表的一个记录中,某个属性的值为“1(T)”则表示在相应的事务中包含了相应的项目,否则属性值为“0(F)”。这就需要使每个属性都对应一个新的布尔型属性。离散型属性或定量型属性只取少数几种值,对应很容易,而本文中的话务量“TRAFFIC”属性取值范围大、数值多,有必要对属性进行划分,再对应到布尔型属性上。这个问题就是连续属性离散化,即在特定连续型属性的值域范围内,根据某种评价规则,设定若干个划分点,用这些划分点将属性的值域范围划分成一些子区间(离散化

区间), 最后用特定的符号或整数值来代表每个子区间。划分的方法对关联规则提取的质量起着决定性的作用。通常采用的方法有等宽度离散法、等频离散法、基于信息熵的方法和基于粗糙集的方法等。本文中采用的是 χ^2 归并离散化方法[2], 步骤如下:

1) 将连续值排序去重, 然后将每一个值划分为一个间隔;

2) 计算每 2 个相邻间隔的 χ^2 值;

3) 合并具有最小 χ^2 值的 2 个间隔, 直到所有 χ^2 值都小于 χ^2 阈值(χ^2-threshold);

4) 最后赋予每个间隔一个唯一的离散值。

在本文中, 间隔的相似度就定义为 χ^2 值

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(A_{ij}-E_{ij})^2}{E_{ij}}$$

式中: m ——间隔数目, $m=2$;

k ——分类属性值个数;

A_{ij} ——在第 i 个间隔中属于第 j 个类的事例个数;

E_{ij} ——A_{ij} 的期望频率, $E_{ij}=(R_i\times C_j)/N$;

R_i ——在第 i 个间隔中的事例个数;

C_j ——属于第 j 个类的事例个数;

N ——事例个数。

step2　布尔转换。经过连续属性离散化后, 每个区间对应一个离散值, 这样所有的属性均已成为离散属性, 然后再将其对应成布尔性属性, 完成"布尔转换表"。

step3　调用 Apriori 算法[3], 挖掘可信度不低于用户最小可信度阈值的规则。首先引入了若干记号和定义:

1) 频繁项目集(频繁集) ——所有支持度不低于用户规定的最小支持度阈值的项目集;

2) k-项目集——具有 k 个项目(属性)的项目集, 其长度为 k;

3) k-项目序列集——由 k-项目集构成的集合;

4) L_k ——由频繁 k-项目集构成的集合;

5) C_k ——由候选 k-项目集构成的集合。

Apriori 算法描述如下:

Step1　遍历目标数据库 1 次, 记录项目(属性)的出现次数, 即计算每个属性的支持度。

Step2　收集所有支持度不低于用户支持度阈值的项目构成频繁1-项目序列集 L_1。

Step3　链接 L_1 中所有的元素对形成候选2-项目序列集 C_2。

Step4　再次遍历目标数据库, 计算 C_2 中每个候选2-项目集的支持度, 得到 L_2, 形成 C_3。

Step5　重复执行上述过程, 直到没有新的候选生成为止。

Step6　从 Step1～5 得到的频繁集中构造可信度不低于用户最小可信度阈值的规则。

2.3　关联规则挖掘实验

利用关联规则挖掘算法, 首先对话务量属性 TRAFFIC 采用 χ^2 归并离散化方法进行离散化处理, 然后对所获得的离散区间进行布尔转换, 最后调用 Apriori 算法挖掘可信度不低于用户最小可信度阈值的规则。从话务量时态数据库中发现的特征规则如下:

(TIME=6) ⇒ (TRAFFIC=8)　　(6% support, 95% confidence)

(WEEK= 6) ⇒ (TRAFFIC= 7) (5% support, 87% confidence)

(WEEK= 7) ⇒ (TRAFFIC= 7) (5% support, 85% confidence)

本实验数据取自内蒙古某县城电信局, 2000 年 3 月。每日采集 6 个小时的数据, 时间分别为每日的 8∶00—9∶00, 10∶00—11∶00, 12∶00—13∶00, 14∶00—15∶00, 16∶00—17∶00, 21∶00—22∶00。TIME 对应 6 个时间段,WEEK 对应星期, TRAFFIC 对应话务量属性被离散后的 8 个区间。

实验结果表明: 每日的第 6 时段, 即 21∶00—22∶00 为一天中话务量最大时段, 星期六和星期日的话务量较其他几天为大, 这正好对应于当时国内长途电话周末和每日 21∶00 点以后半价计费的收费政策。由于所采集数据的地区政府办公地点搬迁, 办公电话急剧减少, 再加上经济比较落后, 所剩单位经济不景气, 导致长途话务量主要依靠个人电话, 而星期六和星期日是半费日且为双休日, 所以这两天话务量比较大。

参 考 文 献

1 牛东晓, 曹数华, 赵 磊, 等. 电力负荷预测技术及其应用. 北京: 中国电力出版社, 1998. 12～17

2 李 爽. 基于高阶特征聚类的数据库知识发现的研究: [学位论文]. 北京: 中国科技大学, 1998, 10～40

3 梁曼君, 张 瑞, 熊范纶. 从数据库中发掘定量型关联规则. 计算机科学, 1999(8): 71～73