第32卷 第22期 计 算 机 工 程 2006年11月
Vol.32 No.22 Computer Engineering November 2006
·软件技术与数据库· 文章编号：1000—3428(2006)22—0061—03 文献标识码：A 中图分类号：TP311

一种时态近似周期的数据挖掘研究

姜 华[1,3]，孟志青[2]，肖建华[1]，彭丽芳[3]，田 密[3]
（1. 湖南省第一师范学校信息技术系，长沙 410002；2. 浙江工业大学经贸管理学院，杭州 310032；
3. 湘潭大学信息工程学院，湘潭 411105）

摘 要：研究了时态近似周期的挖掘问题，提出了近似周期模式，引进了近似精度、近似周期模式覆盖等概念及性质，提出了一个基于SOM(自组织特征映射)聚类来寻找近似周期模式的算法，实验表明算法是有效的。

关键词：数据挖掘；时态型；近似周期模式；SOM

Study on Data Mining for Temporal Approximate Periodicity

JIANG Hua[1,3], MENG Zhiqing[2], XIAO Jianhua[1], PENG Lifang[3], TIAN Mi[3]
(1. Dept. of Information & Technology, Hunan First Normal College, Changsha 410002; 2. College of Business and Administration, Zhejiang University of Technology, Hangzhou 310032; 3. School of Information & Engineering, Xiangtan University, Xiangtan 411105)

【Abstract】 This paper discusses a problem of temporal approximate periodicity. It presents an approximate periodic pattern on the basis of temporal type, introduces the concepts of approximate precision and approximate periodic pattern mantle, and proves some relative properties. The paper discusses an algorithm based on self-organizing map to find approximate periodic pattern. Experiments show that proposed algorithms are efficient.

【Key words】 Data mining; Temporal type; Approximate periodic pattern; SOM

1 概述

在时态数据库中周期是一个非常有趣的用于理解时态数据、预测未来趋势的特征，周期挖掘有着广泛的应用领域，如股票价格、市场营销、天气变化等。目前周期模式挖掘问题可以分为3类：(1)挖掘全周期模式，它是指每一时间点都影响着时序上的循环行为。例如一周中的每一天的销售量都会对一周中的销售量发挥作用。(2)挖掘部分周期模式，它描述部分时间点的时序周期。例如小李每天早晨7：00—7：30读《人民日报》，而其他时间没有什么规律。(3)挖掘周期关联规则，这种规则是周期出现的事件的关联规则。例如，设时间单位t为小时的关联规则：牛奶$\rightarrow$面包具有周期$c(24,7)$，即每天7AM—8AM有关联规则牛奶$\rightarrow$面包成立。这些周期模式挖掘问题都是寻找严格意义上的数学周期，然而，现实生活的周期往往不是严格固定的，而是在固定时间K的一个时间范围δ内波动。例如，顾客A买了某种商品B(如生活日用品)，当他们在一定时间内消费完该产品后，又需要重新购买该产品，而消费产品的时间可能是一个月左右，并不是刚好一个月。因此，考虑这样一种周期现象，即事件发生后，相隔一个时间周期(该时间周期可以在一定范围内波动)事件重复发生，称之为近似周期，这是一个很有意义的研究方向，因为很多实际问题看上去没有严格的周期规律，但有可能存在近似周期，例如股票价格变化，基本上没有周期规律，如果能够发现股票的近似周期，这对研究股票价格变化是非常有意义的。

在时态数据挖掘方面，文献[1]系统地研究了时态型与时间粒度的有关理论。在时态数据的周期方面，近年来有了大量的研究。文献[2]首次研究了完整的周期模式，目的是寻找重复发生的关联规则；文献[3]引入了部分周期的概念并提出了一个在时间序列中寻找部分周期的高效算法——“最大子模式命中集算法”。前面的研究虽然准许模式在某些部分出现缺失，但没有考虑现实世界中的周期往往在一定范围内波动，精确周期并不常见。文献[4]提出了异步周期的概念和挖掘算法，即考虑周期之间可能插入噪音而发生平移的现象，当平移的时间间隔在一定范围内仍认为这种现象符合周期模式。但异步周期是一种完美周期，不允许模式的循环在某些部分出现缺失，置信度必须是100%。

目前有关周期模式和周期关联规则的大部分研究都是基于Apriori特性启发式和采用了变通的Apriori挖掘方法，文献[7]中根据事务发生的频度利用聚类分析来实现对时间段的划分，但并没有采用聚类方法发现周期。本文与传统方法不同，提出了一种新的挖掘周期的思路，针对时态数据数据量大、属性多、维数高等特点，采用SOM网络聚类的方法从时态数据库中挖掘近似周期。

本文的主要贡献为：(1)考虑周期可能在一定范围内波动，引进近似精度和近似周期模式等概念；考虑到非完美周期的情况，引进支持度和置信度的概念；并证明了相关性质。(2)给出一个SOM网络自组织聚类挖掘近似周期模式算法。

2 近似周期定义

下面给出近似周期的严格数学定义：

设A表示具有有限个属性/特征的集合，记$A=\{A_1,A_2,\ldots,A_m\}$。设E表示有限个状态构成的集合，记

作者简介：姜 华(1980－)，女，硕士生，主研方向：数据挖掘，神经网络；孟志青、肖建华，教授；彭丽芳、田 密，硕士生

收稿日期：2006-01-07 **E-mail：**jianghua_cl01@126.com

$E=\{e_1,e_2,\ldots,e_m\}$，每个状态$e\in E$表示某一属性/特征的发生状况或一种描述。

设v是 1 个时态型，用$(A_i,e,v(t))$表示属性/特征A_i在时态因子$v(t)$处发生e值的事件[1,6]。

定义 1 符号$(A_i,e,v(t_k),p)$表示属性/特征A_i在时态因子$v(t_k)$处发生e值的事件$(A_i,e,v(t_k))$，从当前时态因子$v(t_k)$开始紧后的(间隔)p个时态因子处事件$(A_i,e,v(t_{k+p}))$重复发生，其中$p\geq 1$为整数。

定义 2 符号(A_i,e,p)表示在时间段$[T,T']$中存在时态因子$v(t)$使得在每次发生后紧后p个时态因子的事件$(A_i,e,v(t),p)$重复发生的事件。记

$$N^{V}[T,T'](A_i,e,p)=\sum_{k=1}^{n}E((A_i,e,v(t_k),p))$$

表示(A_i,e,p)在时间段$[T,T']$中所有时态因子$v(t)$的事件$(A_i,e,v(t_k),p)$重复发生的次数，其中若事件$(A_i,e,v(t_k),p)$发生，记$E((A_i,e,v(t_k),p))=1$；否则$E((A_i,e,v(t_k),p))=0$。

定义 3 设给定波动值 $\sup(p)>\inf(p)$($\inf(p)$可以取 0)，定义一种近似周期模式的表示：

$$P=((A_i,e):[\inf(p),\sup(p)])=\{(A_i,e,p_k)\,|\,\inf(p)\leq p_k\leq\sup(p)\}$$

P表示事件(A_i,e,p_k)每隔时态因子p_k的重复发生所有的事件构成的集合，其中p_k可以在一定范围$[\inf(p),\sup(p)]$内波动，称$[\inf(p),\sup(p)]$为属性A_i出现状态e (表示为(A_i,e))的近似周期。$\sup(p)-\inf(p)$称为近似周期波动阈值或近似精度。当$\sup(p)-\inf(p)=0$时，表示精确周期。

定义 4 (支持度和置信度)近似周期$((A_i,e):[\inf(p),\sup(p)])$的支持度为

$$\begin{aligned}&\operatorname{support}((A_i,e):[\inf(p),\sup(p)])\\&=\frac{N^{V}[T,T']((A_i,e):[\inf(p),\sup(p)])}{n}\end{aligned}$$

近似周期$((A_i,e):[\inf(p),\sup(p)])$的置信度为

$$confidence((A_i,e):[\inf(p),\sup(p)])=\frac{N^{V}[T,T']((A_i,e):[\inf(p),\sup(p)])}{N^{V}[T,T'](A_i,e)}$$

其中符号$N^{V}[T,T']((A_i,e):[\inf(p),\sup(p)])$表示在时间段$[T,T']$中所有时态因子$v(t)$满足模式

$$P=((A_i,e):[\inf(p),\sup(p)])$$

的个数，记

$$N^{V}[T,T']((A_i,e):[\inf(p),\sup(p)])=\sum_{k=1}^{n}E(A_i,e,v(t_k),p_j)$$

其中P_j：(1)$\inf(p)\leq p_j\leq\sup(p)$；(2)若存在时态因子$v(t)$，使得

$$(A_i,e,v(t),p_{j1}),(A_i,e,v(t),p_{j2}),\ldots,(A_i,e,v(t),p_{js})$$

均发生，其中

$$\inf(p)\leq p_{j1},p_{j2},\ldots,p_{js}\leq\sup(p)$$

则取

$$p_j=\min\{p_{j1},p_{j2},\ldots,p_{js}\}$$

n表示在时间段$[T,T']$中所有时态因子$v(t)$的个数。符号(A_i,e)表示在时间段$[T,T']$中存在时态因子$v(t)$使得事件$(A_i,e,\ v(t))$发生，记

$$N^{V}[T,T'](A_i,e\ )=\sum_{k=1}^{n}E(A_i,e,v(t_k))$$

表示(A_i,e)在时间段$[T,T']$中所有时态因子$v(t)$的事件$(A_i,e,\ v(t))$的重复发生次数。

定义 5 模式

$$P=((A_i,e):[\inf(p),\sup(p)])$$

模式

$$P'=((A_i,e):[\inf(p'),\sup(p')])$$

若

$\inf(p)\leq\inf(p')$，且

$\sup(p)\geq\sup(p')$

称模式 P 覆盖模式 P'。

定义 6 模式

$p=((A_i,e):[\inf(p),\sup(p)])$，称

$$u(P)=\frac{[\sup(p)-\inf(p)]}{\sup(p)}$$

为模式 P 的近似周期波动幅度。

该定义表明 $u(p)$越大，则周期波动范围越大。当 $u(p)=0$，即 $\sup(p)=\inf(p)$时，表示精确周期，周期没有波动。

性质 1 $u(p)$为模式 P 的近似周期波动幅度，则 $0\leq u(p)\leq 1$。

证明 从定义显然得证。

性质 2 对于时间段$[T,T']$中的对象 A，若模式 P 覆盖模式 P'，则 $support(p\geq support(P)$，$confidence(P)\geq$confidence(P')，$u(p)\geq u(p')$。

证明 设模式

$$P=((A_i,e):[\inf(p),\sup(p)])$$

模式

$$P'=((A_i,e):[\inf(p'),\sup(p')])$$

因为

$\inf(p)\leq\inf(p')$，且

$\sup(p)\geq\sup(p')$，据定义

$$\operatorname{support}(P)=\frac{N^{V}[T,T']((A_i,e):[\inf(p),\sup(p)])}{n}=$$

$$\frac{N^{V}[T,T']((A_i,e):[\inf(p'),\sup(p')])+N^{V}[T,T']((A_i,e):[\inf(p),\inf(p')]+N^{V}[T,T']((A_i,e):[\sup(p'),\sup(p)])}{n}$$

$$\geq\operatorname{support}(P')$$

同理可证

$confidence(P)\geq$confidence(P')

下证 $u(p)\geq u(p')$，令

a=sup(p)，b=inf(p)，c= sup(p')，d= inf(p')，则 $a\geq c$，$b\leq d$

$$u(P)-u(P')=\frac{a-b}{a}-\frac{c-d}{c}=\frac{ad-bc}{ac}\geq 0$$

故 $u(p)\geq u(p')$，证毕。

3 近似周期模式挖掘算法

这一节将用 SOM 自组织神经网络发现近似周期模式。

3.1 SOM 网络(自组织特征映射网络)

SOM网络结构如图 1 所示，它由输入层和竞争层组成。输入层神经元数为n，竞争层由$M=m^2$个神经元组成的二维平面阵列，输入层与竞争层各神经元之间实现全互连接。

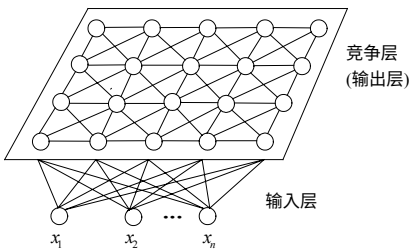

图 1 SOM 网络结构

3.2 SOM 训练算法

将网络各结点权赋予[0，1]区间的随机数作为初始值$w_{ij}(0)$，$i=1,\ldots,n$，输入为n维，$j=1,\ldots,M$，输出为$M=m^2$。确定

学习率$\eta(t)$的初始值$\eta(0)$(0＜$\eta(0)$＜1＝；确定邻域$N_C(t)$的初始值$N_C(0)$，邻域$N_C(t)$的值表示在第t次学习过程中邻域所包含的神经元的个数；确定总的学习次数T。

(1)在q个学习模式中随机选取一个模式$x(t)$=(x_1(t), x_2(t),…x_n(t)))提供给网络的输入层。

(2)在输出节点寻找最佳匹配节点(获胜神经元)C，可用欧氏距离，则 C 应为

$$\|x(t)-w_C(t)\|=\min_j\{\|x(t)-w_j(t)\|\}$$

(3)确定邻域函数并修正邻域内节点的权值

$$w_j(t+1)=\begin{cases}w_j(t)+\eta(t)[x(t)-w_j(t)], j\in N_C\\ w_j(t), j\notin N_C\end{cases}$$

(4)选取另一个学习模式提供给网络的输入层，转(2)，直到 q 个学习模式全部提供给网络。

(5)$t\leftarrow t+1$，转(1)，直到 t=T。

3.3 挖掘算法

挖掘算法处理的输入为：

(1)给定对象(如某支股票)和时间段$[T,T']$，选定时态型v，不妨设时态型v将$[T,T']$划分成n段，即存在n个$t_1,t_2,\ldots,t_n$，使得

$$[T,T']=\bigcup_{i=1}^{n}v(t_i)$$

其中

$$t_1<t_2<\ldots<t_n$$

$$v(t_i)\bigcap v(t_j)=\varnothing$$

$$i\neq j$$

$$i,j=1,2,\ldots,n$$

在此基础上符号化属性状态，得到一个符号化的时间序列；

(2)周期长度阈值L，支持度阈值$\underline{s}$，置信度阈值$\underline{c}$和近似精度。要找到在$[1,\ldots,L]$范围内符合$\underline{s}$，$\underline{c}$和近似精度的近似周期模式 P。

下面们给出相应的算法：

(1)对于每个$v(t_i)$，计算(A_i,e_j,p)，$1\le p\le L$。

(2)初始化SOM网络，包括输入层结点数、输出层结点数、学习次数T、权值矢量w_{ij}(t)、邻域函数$N_C(t)$和学习速率$\eta(t)$的初始值选取。

(3)为保证聚类效果，以便将属性状态相同的近似周期聚为一类。根据输入的各分量在聚类划分时重要性的大小，对输入向量各分量分配不同的权重值。即将输入向量(A_i,e_j,p)表示成$(\alpha A_i,\beta e_j,\gamma p)$，其中α,β,γ为权重值，把每个$v(t_i)$对应的$(\alpha A_i,\beta e_j,\gamma p)$逐一输入网络的输入层，得到输入向量

$$x(t)=(x_1(t),x_2(t),x_3(t))$$

(4)计算欧氏距离

$$d_j=\sum_{i=1}^{3}(w_{ij}(t)-x_i(t))^2$$

找出距离最小的输出结点作为获胜结点。

(5)调整获胜结点及其邻域内的结点权值

$$w_j(t+1)=\begin{cases}w_j(t)+\eta(t)[x(t)-w_j(t)], j\in N_C\\ w_j(t), j\notin N_C\end{cases}$$

(6)降低学习率$\eta(t)$和邻域函数$N_C(t)$，转(4)，直到满足条件(达到指定的学习次数)。

(7)聚类结束，输出结点的个数即为类的个数。类对应的每个输入向量x_i=(x_{i1},x_{i2},x_{i3})，如果x_{i1}=x_{j1}且x_{i2}=x_{j2}，其中$i\neq j$，则认为该类达到了聚类目的，记录最小x_{i3}和最大x_{i3}，得到$((A_i,e)$:[inf(p),sup(p)])。

(8)计算支持度s，置信度c和近似精度；并输出满足给定的近似精度，s≥$\underline{s}$，$c\ge\underline{c}$的模式。

4 实验

算法用 C++实现，运行在 CPU 为 P4 1.8GHz，256MB 内存，操作系统为 Windows2000 的 PC 上。

应用本文的算法对多年来的几支股票数据进行挖掘。以天为时态型，将股票交易数据的价格属性按如下规则符号化为 5 种状态：设 x=(当天的开盘价-前一天的开盘价)/前一天的开盘价，若 $x\in(-\infty,-0.02]$，意指开盘价大幅下跌，置状态值为 1；$x\in(-0.02,-0.01]$意指开盘价小幅下跌，置状态值为 2；$x\in(-0.01,0.01]$，意指开盘价正常波动，置状态值为 3；$x\in(0.01,0.02]$，意指开盘价小幅上涨，置状态值为 4；$x\in(0.02,+\infty]$，意指开盘价大幅上涨，置状态值为 5。设置阈值 $\underline{s}=10\%,\underline{c}=50\%$，$L=30$，近似周期的近似精度分别取 0，1，2，3，4。

下面给出 St 中华，深发展，St 石化，深深宝 4 支股票的实验结果。在试验中发现了许多满足要求的模式的近似周期，其中周期短、置信度高、满足模式覆盖的近似周期是最有意义的，列出部分试验结果于表 1 中。其中的空白行表明没有发现满足要求的近似周期模式。

表 1 几支股票的近似周期挖掘实验结果

股票名称	时间段	近似周期	支持度	置信度	近似精度 sup(p)-inf(p)
St中华	19920331-20020919	((开盘价,5):[12,13])	13.557 994	50.510 948	1
		((开盘价,5):[12,14])	17.750 784	66.131 386	2
		((开盘价,5):[11,14])	20.297 806	75.620 438	3
		((开盘价,5):[11,15])	22.021 944	82.043 793	4
深发展	19910102-20020919	((开盘价,3):[12,12])	22.490 514	50.818 394	0
		((开盘价,3):[11,12])	30.907 209	69.836 319	1
		((开盘价,3):[11,13])	35.701 97	80.124 710	2
		((开盘价,3):[11,14])	38.289 065	86.515 976	3
		((开盘价,3):[11,15])	40.117 282	90.646 919	4
St石化	19920506-20020919	-	-	-	1
		((开盘价,5):[7, 9])	15.994 624	63.297 871	2
		((开盘价,5):[6, 9])	18.324 373	72.517 731	3
		((开盘价,5):[6,10])	20.161 29	79.787 231	4
深深宝	19921012-20020919	((开盘价,1):[13,14])	14.600 666	52.861 446	1
		((开盘价,1):[12,14])	18.802	68.072 289	2
		((开盘价,1):[11,14])	21.173 045	76.656 624	3
		((开盘价,1):[11,15])	23.336 106	84.487 953	4

从表 1 中可以看出：

(1)大多数股票在相应的时间段内并没有严格意义上的周期存在(即精确周期，sup(p)-inf(p)=0)，但是如果放宽近似精度这一约束条件，却能找到近似周期规律。例如：拿 St 中华的股票来说，没有发现精确周期，但是当近似精度>0 时，就能发现近似周期。如：((开盘价,5):[12,14]))表明开盘价大幅度上涨后有 66.131 386%的可能性在每隔 12～14 天大幅度上涨一次。

(2)随着近似精度数值增大，支持度和置信度普遍提高。例如：近似精度为 1，2，3，4 时，深深宝股票的置信度分别为 52.861 446%，68.072 289%，76.656 624%，84.487 953%。当近似精度为 4 时，周期的置信度几乎都达到 80%以上，而深发展的则达到 90%。

(3)对于同一股票对象，若模式 A 覆盖模式 B，则模式 A 的支持度和置信度必然大于模式 B。例如：St 中华股的((开盘价,5):[11,14]))覆盖((开盘价,5):[12,13])，显然，((开盘价,5):[11,14])支持度和置信度都大于(开盘价,5):[12,13])。也就是说，St 中华股开盘价大幅度上涨后 11～14 天内上涨的可能性大于 12～13 天内上涨的可能性。

(下转第 83 页)