非同步多时间序列中频繁模式的发现算法*

李 斌, 谭立湘, 解光军, 李海鹰, 庄镇泉

(中国科学技术大学 电子科学与技术系,安徽 合肥 230026)
E-mail: binli@ustc.edu.cn
http://www.ustc.edu.cn

摘要: 从多个时间序列中发现频繁模式在实际应用中具有非常重要的价值.已知文献所提供的方法均假设多时间序列是同步的,但是,在现实世界中,这一条件并不总能满足,许多情况下它们是非同步的.提出了一个从非同步多时间序列中发现频繁模式的算法.该算法首先利用线性化分段表示法和矢量形态聚类实现时间序列的特征分割与符号化转换,然后通过将 Agrawal 关联模式发现算法的核心思想与时间序列最短实现表示方法相结合,实现了非同步多时间序列中多种结构频繁模式的发掘.与已有算法相比,该算法更简单、更灵活,并且不要求序列严格同步.实验结果证明了该算法的有效性.

关 键 词: 数据挖掘;时间序列;频繁模式;最短实现;符号化
中图法分类号: TP18 文献标识码: A

时间序列是现实世界中最常见的数据形式之一,对时间序列进行分析,可以揭示事物运动、变化和发展的内在规律,对于人们正确认识事物并据此作出科学的决策具有重要的现实意义.数据挖掘利用机器学习等方法,从大量历史数据中发现局部的、频繁出现的行为模式,是一种新的、很有前途的时间序列分析方法.

在对金融领域的多个时间序列(如各种价格数据和指标数据)进行分析时,经常希望能够发现不同时间序列间可能存在的关联关系,这种关联关系一般表现为不同序列中频繁地同时或依次出现的变化模式.发现这种多时间序列中的频繁结构模式对于人们认识金融系统内在的相互影响并据此作出合理的决策具有重要的参考价值.

本文提出了一个从多个时间序列中发现多种结构的频繁模式的数据挖掘算法.该算法首先利用线性化分段和矢量形态聚类方法实现时间序列中基本变化模式的分割与提取,将多个时间序列转换成离散的、非同步的多个符号序列;然后利用 Agrawal 关联模式发现算法的核心思想[1],结合“最短实现”表示方法[2],实现了非同步多符号序列中频繁模式的发现.该算法简单、直观,具有较高的实用价值.

1 相关工作

频繁模式的发现研究始于 Agrawal 提出的关联规则的发现研究[1],一直是数据挖掘研究中的一个重要课题.在文献[1]中,Agrawal 给出了关于频繁模式的一个重要定理,即“任何频繁模式的子模式必定也是频繁的”.由该定理可以得到一个更为实用的推论,即“可以由已知频繁模式集产生更大长度的候选频繁模式”.

Heikki Mannila 将 Agrawal 关联规则发现算法的核心思想推广到事件序列,提出了事件序列中频繁情节的

* 收稿日期: 2000-06-15; 修改日期: 2000-09-26
基金项目: 国家重点基础研究发展规划 973 资助项目(G1998030413);国家教育部博士点基金资助项目(1999035808)
作者简介: 李斌(1970－),男,安徽合肥人,博士,讲师,主要研究领域为数据挖掘,神经网络,遗传算法;谭立湘(1970－),女,山东青岛人,讲师,主要研究领域为数据库,数据通信,多媒体;解光军(1970－),男,安徽合肥人,博士生,讲师,主要研究领域为神经网络,量子计算;李海鹰(1968－),男,安徽合肥人,博士生,讲师,主要研究领域为神经网络,电子商务;庄镇泉(1938－),男,福建泉州人,教授,博士生导师,主要研究领域为智能信息处理.

发现算法[3],事件序列可看作是一种离散的时间序列.

Tim Oates 等人提出了从多个数据流中搜索关联模式的数据挖掘算法(MSDD)[4],其中,多数据流表示为严格同步的多个符号序列.Oates 等人给出了候选模式的产生和强关联模式的启发式搜索算法,但该算法要求数据序列必须是严格同步的.

与 Oates 的挖掘算法相比,本文提出的从多个时间序列中发现频繁结构模式的挖掘算法对不同序列间是否同步没有限制,并且能够发现多种结构形式的频繁模式,具有更大的灵活性和较低的计算复杂度.

2 时间序列的符号化转换

本文研究从时间序列中发现各种频繁出现的结构模式,而以连续数值形式表示的时间序列不便于描述和计算,为此,需要将以数值形式表达的时间序列转换成以离散的、相对抽象的符号表示的符号序列.以后的挖掘算法都要在这个符号序列上展开,最后发现规则的有效性在很大程度上取决于符号表达的有效性.因此,我们希望在作符号化转换时所形成的符号种类数不要太多,每一个符号都尽可能代表一种基本的、相对独立的变化模式.我们称这种模式为“元模式”,它是构成模式及规则表达式的基本元素.

本文采用了一种基于线性化分段和矢量形态聚类的时间序列符号化方法[5].该方法分两步进行(如图 1 所示):首先,利用线性化分段方法实现时间序列的特征分割与表示,将连续的时间序列转换成离散的、形态各异的线性分段序列;然后利用矢量形态相似性度量和神经网络模糊聚类算法实现各分段的聚类,以类标识符替代所有属于该类的分段,即得到离散的符号序列.

经符号化处理以后,多个连续时间序列变成多个离散的符号序列,每个符号代表一个基本的、相对独立的变化模式(元模式),这些符号序列将被用来定义待搜索的模式空间.基于变化形态特征的时间序列分割决定了不同序列的元模式之间必然是非同步的(如图 2 所示,其原始序列见图 3 左半部分),即同一时间序列,不同元模式的起止时间间隔不相等,且垂直方向不同,时间序列各元模式的起止时间不同步,我们称这样的多符号序列为非同步多符号序列.

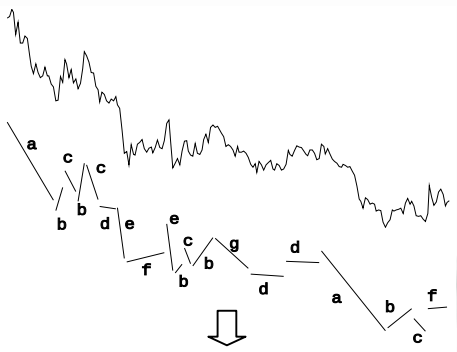


Fig.1 Transform time series into discrete symbol sequenc

图 1 将时间序列转换为离散的符号序列

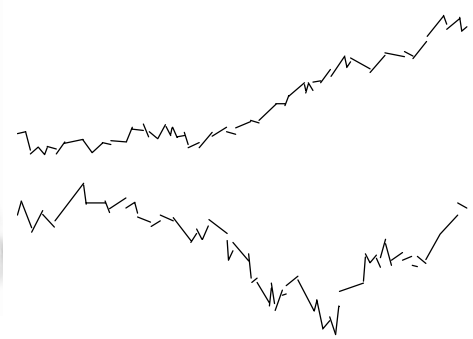


Fig.2 Linear segment representation of multiple time series

图 2 非同步多时间序列的线性化分段表示

不同序列之间模式起止时间的不同步,给垂直方向模式的表达与搜索增加了难度,文献[4]的同步多序列的挖掘算法将不再适用.根据非同步多符号序列的特点,本文提出将关联模式发现算法的核心思想与时间序列中模式“最短实现”表示方法相结合,以实现非同步多符号序列中频繁结构模式的发现算法.该算法不要求元模式之间严格同步,利用第 1 次遍历数据集所获得的各个元模式的“最短实现”集,通过专门的合并算法即可求得各种结构的频繁模式及其在序列中的“最短实现”,并且,在不同条件下获得的各种频繁模式的“最短实现”集可重复利用,随着应用的增多,可以显著地减少计算开销.

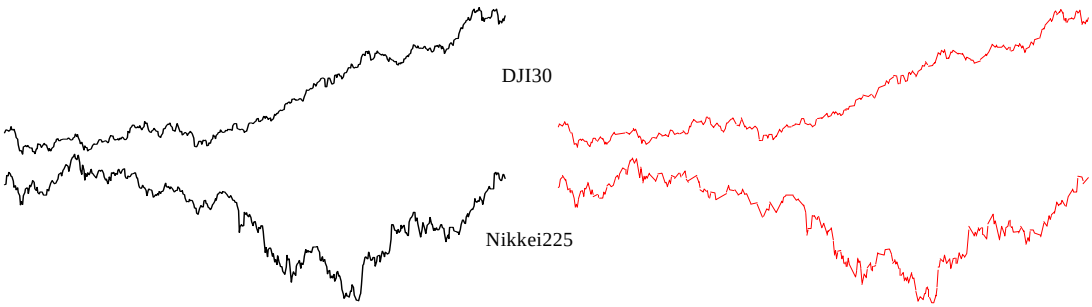


Fig.3 Time series of two stock indexes (left) and their linear segment representations (right)
图 3 两个股市指数的时间序列(左)及其分段线性化表示(右)

3 相关描述

本节给出将要用到的一些概念的形式化描述.

定义 **1.** 给定序列集合 $S=\{S_1,S_2,...,S_m\}$和元模式类型集合 $C=\{C_1,C_2,...,C_n\}$.元模式 $PM=(S_i,C_j,T_s,T_e)$,其中 $S_i\in S$ 为序列标识,$C_j\in C$ 为类标识,T_s 和 T_e 为模式的开始和结束时间.

定义 **2.** 模式 $P=(V,\leq,T_s,T_e)$.其中 V 为元模式 PM 的非空集合.$\leq$为定义在 V 上的某种偏序关系,如果这种偏序关系是非严格顺序的,如对 $\forall x,y\in V$,$x\neq y$,则 $x\leq y$ 不成立,则称模式 P 是并行的;如果该偏序是严格顺序的,如对 $\forall x,y\in V$,均有 $x\leq y$ 或 $y\leq x$,则称模式 P 是串行的.T_s,T_e 为模式的开始和结束时间.模式 P 的长度为集合 V 的大小$|V|$,即构成 P 的元模式的个数.

定义 **3.** 模式 $P_s=(V',\leq',T_s',T_e')$为模式 $P=(V,\leq,T_s,T_e)$的子模式,记为 $P_s\subseteq P$,如果 $V'\subseteq V$,且遵循相同的偏序关系(即$\leq'=\leq$),$T_s=<T_s',T_e'=<T_e$.

定义 **4.** 称模式 $P=(V,\leq,T_s,T_e)$在区间$[t,t']$实现一次,如果在区间$[t,t']$内存在一组元模式 $V'=V$,它们之间满足偏序关系$\leq$,且 $t=<T_s,t'>=T_e$.

定义 **5.** 称模式 P 在区间$[t,t']$的一次实现为最短实现,如果不存在任何子区间$[u,u']\subset[t,t']$,从中可以发现 P 的一次实现.以数组 $Si(P)$记录模式 P 的所有最短实现.

$$Si(P)=\{[t,t']|[t,t']\text{为模式 }P\text{ 的一次最短实现}\}.$$

记模式 P 的出现频率为其在序列集 S 中的最短实现的个数,即数组 $Si(P)$的长度$|Si(P)|$,发现算法的任务是发现所有频率大于给定阈值 $f_{\min}$ 且时间跨度$(t'-t)$小于给定阈值 T 的模式.

4 发现算法

令 C_k 表示长度为 k 的候选频繁模式集,L_k 表示长度为 k 的频繁模式集,则从非同步多符号序列中发现频繁模式的算法总体框架如下:

算法 **1.** Find_Freq_Pattern($S,T,f_{\min}$)

输入:序列集合 S,时间宽度 T,最低频率阈值 $f_{\min}$.

输出:全部时间跨度在 T 以内的频繁模式集 L_k 及其最短实现集$\{Si(P)|P\in L_k\}$.

程序:

(1) C_1=长度为 1 的候选频繁模式集=全部元模式集合; //为便于计算,C 中元素分别按起始时间和序列号进行了排序.

(2) For all $P\in C_1$,计算 $Si(P)$; //为便于后面计算,$Si(P)$中的元素均按起始时间先后排序.

(3) $L_1=\{P\in C_1|\ length(Si(P))>=f_{\min}\}$;

(4) $k=1$;

(5) while $|L_k|>=2$

(a) $k=k+1$;

(b) C_k =C_generator(L_{k-1}); //由 L_{k-1} 产生长度为 k 的候选频繁模式集 C_k,其中每个候选频繁模式的所有子模式都属于 L_{k-1}.

(c) For all $P\in C_k$ do //计算 $Si(P)$,由算法 2 实现;

(d) 选择子模式 P_1 和 P_2,由它们的最短实现集 $Si(P_1)$和 $Si(P_2)$计算 $Si(P)$;

(e) $L_k=\{P\in C_k|\ \text{length}(Si(P))>=f_{\min}\}$;

(6) 输出全部频繁模式集 $L_k(k=1,2,...)$及每个频繁模式的全部最短实现 $Si(P)$.

候选频繁模式集 C_k 的产生(函数 C_generator 实现)原理同文献[1].为了加快算法的搜索效率,本文将所有已获得的频繁模式(表示为元模式标识序列)按字母顺序进行排序.为便于区分,规定不同时间序列的元模式标识集之间不存在重叠.然后,根据前$(k-2)$个标识是否相同将该序列分割成若干段,根据 Agrawal 关于频繁模式的定理,长度为 k 的候选模式只会由前$(k-2)$个字符相同的长度为$(k-1)$的频繁模式合并而成,因而搜索将只在各个分段中进行,搜索空间大为缩小.

在已获得全部长度为 $k-1$ 的频繁模式及其所有最短实现 $Si(P)$以后,对每一个长度为 k 的候选频繁模式 c_k,就可以根据其子模式的最短实现 $Si(p_{k-1})$求得自己的全部最短实现.以下是通过子模式最短实现集之间的合并计算求取串行候选频繁模式最短实现的算法:

算法 **2.** join(L_{k-1},Si,C_k)

输入:长度为 $k-1$ 的频繁模式集 L_{k-1} 及其最短实现集$\{Si(P)|P\in L_{k-1}\}$,候选频繁模式集 C_k.

输出:全部候选频繁模式的最短实现集$\{Si(P)|P\in C_k\}$.

程序:

(1) for all $P\in C_k$ do

(a) find $P_1\in L_{k-1}$,where $P_1(1)=P(1)$, $P_1(2)=P(2)$,…,and $P_1(k-1)=P(k-1)$; //$P(i)$表示构成模式 P 的第 i 个元模式.

(b) find $P_2\in L_{k-1}$,where $P_2(1)=P(2)$,$P_2(2)=P(3)$,…,and $P_2(k-1)=P(k)$;

(c) $Si(P)=\{[t,t']$|存在$[t,t1]\in Si(P_1)$和$[t2,t']\in Si(P_2)$,有 $t<t2$,$t1<t'$,$t'-t=<T$,且$[t,t']$是最短的$\}$;

(2) end.

对于并行候选频繁模式,情况稍有不同.在寻找子模式 $P_1\in L_{k-1}$ 和 $P_2\in L_{k-1}$ 时,它们各自省略的属于 $P\in C_k$ 的元模式必须不同,且 $Si(P)$的计算公式如下:

$Si(P)=\{[t,t']$|存在$[t_1,t_1']\in Si(P_1)$和$[t_2,t_2']\in Si(P_2)$,有 $t=\min\{t_1,t_2\}$,$t'=\max\{t_1',t_2'\}$,$t'-t=<T$,且$|Si(P_1)|+|Si(P_2)|$最小$\}$;

$|Si(P_1)|+|Si(P_2)|$最小是为提高计算效率而提出的,因为$|Si(P_1)|+|Si(P_2)|$的大小决定了合并算法搜索空间的大小,在其他条件相同的情况下,选择$|Si(P_1)|+|Si(P_2)|$最小的 P_1 和 P_2 所构成的搜索空间最小.由于对于模式 $P\in C_k$ 的任一最短实现$[t,t']$,只要 P_1 是它的一个子模式,则必存在 P_1 的一个实现$[t_1,t_1']$($t_1>=t$ 和 $t_1'=<t'$),且$[t_1,t_1']\in Si(P_1)$,否则必然还存在$[t_{11},t_{11}']\in Si(P_1)$,使得 $t_{11}>=t_1$ 和 $t_{11}'=<t_1'$.同理,也必存在其子模式 P_2 的一个最短实现$[t_2,t_2']\in Si(P_2)$($t_2>=t$ 和 $t_2'=<t'$).因而,$|Si(P_1)|+|Si(P_2)|$是否最小对并行候选频繁模式的产生没有影响,不会遗漏候选模式.

算法 1 的每次执行都将产生满足给定频率阈值 $f_{\min}$ 和时间宽度 T 的一组频繁模式集 $L_k(f_{\min},T)$,$k=1,2,...,m$,及其每个频繁模式的所有最短实现 $Si(P)$,$P\in L_k$.由最短实现的定义可以得到如下的关于频繁模式集及其最短实现之间关系的定理.

定理 **1.** 设已有满足条件 $f_{\min}=f_{\min 1}$ 和 $T=T_1$ 的一组频繁模式集 F_1 及其最短实现,由最短实现的定义可得,该频繁模式集必为满足条件 $f_{\min 2}=<f_{\min 1}$ 和 $T_2>=T_1$ 的频繁模式集 F_2 的一个子集.其中,如果 $T_2=T_1$,则每个频繁模式的最短实现集不变;如果 $T_2>=T_1$,则已有的最短实现集是新的最短实现集的一个子集.

由上述定理,我们可以得到如下一个定义在已有频繁模式集上的计算.

设已有两个频繁模式集 $F_A(f_{\min\ A},T_A)$和 $F_B(f_{\min\ B},T_B)$,满足 $f_{\min\ A}<f_{\min\ B}$ 和 $T_A<T_B$.则 $F_A\cap F_B=L_{A\cap B}(f_{\min\ B},T_A)=\{P|P\in F_A\&P\in F_B\}$,且对 $\forall\ P\in L_{A\cap B}(f_{\min\ B},T_A)$,$Si(P)=Si_A(P)$,其中,$Si_A(P)$表示频繁模式集 F_A 中模式 P 所对应的最短实现集.

可见,基于最短实现的频繁模式的发现结果可被用于计算或直接构成新的频繁模式集,随着挖掘的多次进行,可以显著降低计算开销.

5 实验结果

在金融投资决策中经常需要考虑不同国家或地区的证券市场之间的相互影响.本文的实验试图从美国DJI30 指数与日本 Nikkei225 指数 1994 年 3 月 10 日~1996 年 1 月 18 日的价格变化中发现频繁出现的结构模式.图 3 给出了原始时间序列(左)及其线性化分段表示(右).经符号化处理之后,两个时间序列被转换成两个离散的符号序列,包括 59 类共 480 个元模式.

本文分别考察了最小频率阈值 f_{min} 和时间宽度 T 取不同值时的挖掘结果,表 1 给出了 f_{min}=3 而 T 取不同值时所发现的不同长度候选频繁模式集和频繁模式集大小的变化.表 2 给出了 T=40 而 f_{min} 取不同值时所发现的不同长度候选频繁模式集和频繁模式集大小的变化.从表中可见,随着频率阈值的减小或时间宽度的增大,所发现的频繁模式的最大长度越来越长,同一长度频繁模式的发现数量也越来越多.

Table 1 Number of candidate patterns and frequent patterns of different lengths (K) discovered when different spans of time (T) are adopted

表 **1** 时间宽度(T)取不同值时所发现的不同长度(K)候选及频繁模式的数量

f_{min}=3		K=1	K=2	K=3	K=4	K=5	K=6
T=10	Candidate pattern set①	59	2 500	556			
	Frequent pattern set②	50	179	4			
T=20	Candidate pattern set	59	2 500	5 734	66		
	Frequent pattern set	50	569	266	8		
T=30	Candidate pattern set	59	2 500	13 274	1 165	17	
	Frequent pattern set	50	846	1 465	256	9	
T=40	Candidate pattern set	59	2 500	20 315	8 647	378	8
	Frequent pattern set	50	1 054	4 087	2 073	201	7

①候选模式集,②频繁模式集.

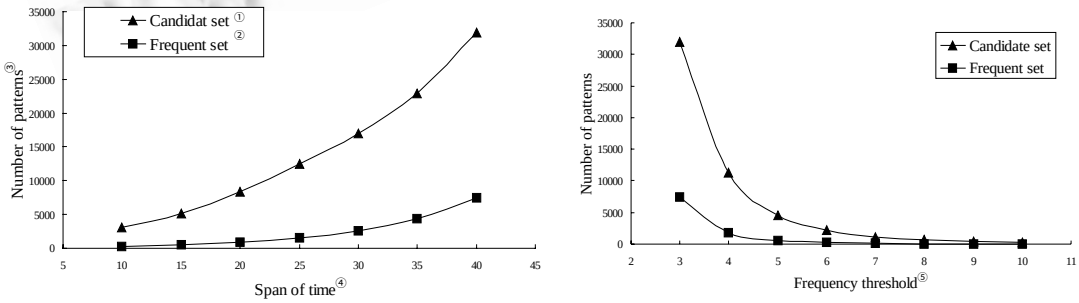
Table 2 Number of candidate patterns and frequent patterns of different lengths (K) discovered when different frequency thresholds (f_{min}) are adopted

表 **2** 频率阈值(f_{min})取不同值时所发现的不同长度(K)候选及频繁模式的数量

T=40		K=1	K=2	K=3	K=4	K=5	K=6
f_{min}=3	Candidate pattern set①	59	2 500	20 315	8 647	378	8
	Frequent pattern set②	50	1 054	4 087	2 073	201	7
f_{min}=5	Candidate pattern set	59	1 600	2 824	82	4	
	Frequent pattern set	40	332	224	8	1	
f_{min}=7	Candidate pattern set	59	900	173	6		
	Frequent pattern set	30	66	10	1		
f_{min}=9	Candidate pattern set	59	361	15	2		
	Frequent pattern set	19	13	2			

①候选模式集,②频繁模式集.

图 4 给出了 f_{min} 和 T 取不同值时所发现的候选频繁模式和频繁模式的数量变化曲线,由图可见,随着 f_{min} 的减小或 T 的增大,候选频繁模式与频繁模式的数量均呈增长态势,其中,随时间宽度的变化相对比较均匀,而随频率阈值的变化则近似呈指数形式变化.



①候选模式,②频繁模式,③模式数量,④时间宽度,⑤频繁阈值.

Fig.4 Relationship of the number of patterns with the span of time (left) and frequency threshold (right)

图 4 模式数量与时间宽度 T(左)和频率阈值 f_{min}(右)的关系

图 5 给出了从上述时间序列中发现的几个不同长度频繁模式的线性化分段表示(坐标尺度仅对单个片段有意义).

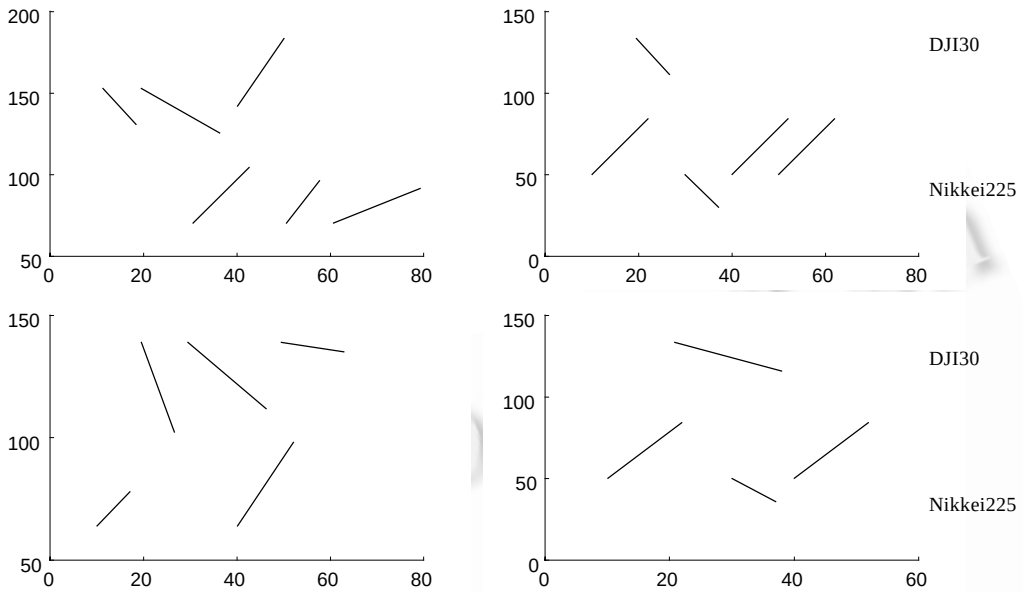


Fig.5 Some discovered frequent patterns of different lengths represented in linear segments

图 5 几个已发现的以线性分段表示的不同长度频繁模式

6 结 论

本文提出了一个非同步多时间序列中频繁模式的发现算法,在实际金融时间序列上的实验表明:

(1) 本文采用的基于线性化分段和矢量形态聚类方法的时间序列分割与符号化方法,可以最大程度地保留原时间序列的变化特征,所产生的每一个符号均可代表一个基本的、相对独立的时间序列变化模式,因而从根本上保证了挖掘结果的有效性.

(2) 基于"最短实现"的频繁模式发现算法对于非同步时间序列具有独特的优点.该算法实现简单,且发掘的结果可重复利用,对于需要多次重复挖掘的应用来说,可以有效地降低计算开销.

References:

[1] Agrawal, R., Imielinski, T., Swami, A. Mining association rules between sets of items in large databases. In: Buneman, P., Jajodia, S., eds. Proceedings of the ACM SIGMOD Conference on Management of Data (SIGMOD'93). Washington, DC: ACM, 1993. 207~216.

[2] Mannila, H., Toivonen, H. Discovering generalized episodes using minimal occurrences. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96). Portland, Oregon: AAAI Press, 1996. 146~151. http://www.cs.helsinki.fi/~mannila/postscripts/genepisodes.ps.

[3] Mannila, H., Toivonen, H., Verkamo, A.I. Discovery of frequent episodes in event sequences. Data Mining and Knowledge Discovery, 1997,1(3):259~289.

[4] Oates, T., Cohen, P.R. Searching for structure in multiple streams of data. In: Proceedings of the 13th International Conference on Machine Learning. Morgan Kaufmann Publishers, Inc., 1996. 346~354. http://www-eksl.cs.umass.edu/papers/Oates96c.ps.

[5] Li, Bin, Tan, Li-xiang, Zhang, Jin-song, *et al*. The study of the data mining oriented method for the symbolization of time series. Journal of Circuits and Systems, 2000,5(2):9~14 (in Chinese).

附中文参考文献:
[5] 李斌,谭立湘,章劲松,等,面向数据挖掘的时间序列符号化方法研究.电路与系统学报,2000,5(2):9~14.

An Algorithm for Discovering Frequent Patterns in Non-Synchronous Multiple Time Series*

LI Bin, TAN Li-xiang, XIE Guang-jun, LI Hai-ying, ZHUANG Zhen-quan

(*Department of Electronic Science and Technology*, *University of Science and Technology of China*, *Hefei* 230026, *China*)

E-mail: binli@ustc.edu.cn

http://www.ustc.edu.cn

Abstract: Discovering frequent patterns in multiple time series is important in practices. Methods appeared in literatures assume that the multiple time series are synchronous, but in the real world, that is not always satisfied, in most cases they are non-synchronous. In this paper, an algorithm for discovering frequent patterns in non-synchronous multiple time series is proposed. In this algorithm, first, the time series is segmented and symbolized with the linear segment representation and the vector shape clustering method, so that each symbol can represent a primitive and independent pattern. Then, the minimal occurrence representation of time series and the association rule discovery algorithm proposed by Agrawal is combined to extract frequent patterns of various structures from non-synchronous multiple time series. Compared with the previous methods, the algorithm is more simple and flexible, and does not require time series to be synchronous. Experimental results show the efficiency of the algorithm.

Key words: data mining; time series; frequent pattern; minimal occurrence; symbolization

* Received June 15, 2000; accepted September 26, 2000

Supported by the National Grand Fundamental Research 973 Program of China under Grant No.G1998030413; the National Research Foundation for the Doctoral Program of Higher Education of China under Grant No.1999035808