ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.18, No.2, February 2007, pp.246−258
DOI: 10.1360/jos180246

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

时态数据挖掘的相似性发现技术*

潘 定[1,2+], 沈钧毅[2]

[1](暨南大学 管理学院,广东 广州 510632)

[2](西安交通大学 计算机科学与技术系,陕西 西安 710049)

Similarity Discovery Techniques in Temporal Data Mining

PAN Ding[1,2+], SHEN Jun-Yi[2]

[1](Management School, Jinan University, Guangzhou 510632, China)

[2](Department of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China)

+ Corresponding author: Phn: +86-20-85220180, E-mail: pandingcn@gmail.com

Pan D, Shen JY. Similarity discovery techniques in temporal data mining. *Journal of Software*, 2007,18(2): 246−258. http://www.jos.org.cn/1000-9825/18/246.htm

Abstract: Temporal data mining (TDM) has been attracting more and more interest from a vast range of domains, from engineering to finance. Similarity discovery technique concentrates on the evolution and development of data, attempting to discover the similarity regularity of dynamic data evolution. The most significant techniques developed in recent researches to deal with similarity discovery in TDM are analyzed. Firstly, the definitions of three categories of temporal data, time series, event sequence, and transaction sequence are presented, and then the current techniques and methods related to various sequences with similarity measures, representations, searching, and various mining tasks getting involved are classified and discussed. Finally, some future research trends on this area are discussed.

Key words: data mining; temporal data; similarities discovery; temporal rule

摘 要: 现实世界存在着大量的时态数据,时态数据挖掘(temporal data mining,简称 TDM)是近年来学术界关注的一个重要研究课题.相似性发现技术关注数据的发展变化,试图从时态数据中发现事物动态演化的相似性规律.分析和比较了近年来 TDM 研究中涉及的主要相似性发现技术.首先区分定义了 3 类时态数据:时间序列、事件序列和交易序列;然后分类并讨论了各种与序列相关的主要方法和技术,涉及相似性度量、序列抽象表示和搜索,以及各类挖掘任务及其算法操作;最后展望进一步研究的方向.

关键词: 数据挖掘;时态数据;相似性发现;时态规则

中图法分类号: TP311 文献标识码: A

传统上,大多数数据挖掘问题仅涉及静态数据.近几年来,主要关注数据动态特性的时态数据挖掘(temporal data mining,简称 TDM)成为学术界研究的热点之一.在现实生活中,时态数据随处可见,如股市交易指数、超市销售、Web 访问、气象观测、临床数据等.对时态数据的知识发现已有多方面的研究.统计学研究时间序列,主

* Supported by the National Natural Science Foundation of China under Grant Nos.60173058, 70372024 (国家自然科学基金)
Received 2005-06-19; Accepted 2006-01-11

要集中在实数或离散数量的预测上,时间序列分析方法可用于时态数据的表示、度量和预测;机器学习研究涉及离散序列的模式发现和预测;数据库研究涉及多维数据的存储、索引以及对大数据集的相似性查询.

1993 年,Agrawal 等人首先发表了关于时间序列相似搜索的研究论文[1].此后,相关研究项目和研究者不断增加.IBM 公司的 Agrawal 和 UC Irvine 的 Pazzani 研究小组较早且持续开展相关研究,UC Riverside 的 E. Keogh 和 UI Urbana-Champaign 的 J. Han 研究小组是目前 TDM 界最活跃的群体.还有美国的 UC Santa Barbara、Maryland 大学、Massachusetts 大学、南澳洲的 Flinders 大学等,也活跃着相应的研究小组.国内的相关研究大约从 2000 年开始起步,复旦大学、浙江大学、中国科技大学等[2]曾有相关的研究,但比较零散,缺少系统性.

传统数据挖掘研究通常按分析目的区分为:描述性方法、预测性方法和局部方法.在 TDM 研究中,将数据演化规律作为时态特性进行分析,传统的区分已难以适用.文献[3]主张将研究分为相似性发现和特征值预测.本文采用这种观点,集中讨论时态数据的相似性发现技术.时态数据相似性发现主要有 3 类任务:① 相似性搜索,它既是一类独立的挖掘技术,又是实现其他任务的基础;② 时态关联规则和序列模式的发现,涉及局部特征相似统计显著性;③ 全局特征综合分类,涉及全局特征相似性描述的聚类和有监督的分类.

高维度、高特征相关性和大量噪音是时态数据的独特结构.这种特征使许多经典算法难以发挥作用,增加了挖掘算法的研究难度.通常认为解决此类问题需要 3 种成分或技术:① 比较两个序列的距离度量;② 抽象表示数据形态的技术;③ 实现特定挖掘任务的搜索或优化算法.

本文第 1 节定义相似性发现问题.第 2 节讨论几种主要的距离度量.第 3 节归纳分析序列的 3 类表示和搜索方法.第 4 节详细讨论相似性发现的挖掘任务及其操作.第 5 节总结并说明未来的研究方向.

1 相似性发现问题

在 TDM 研究中,主要涉及 3 类时态数据:时间序列、事件序列和交易序列.对给定的有限时间集 T、非空状态属性集 $A=\{A_1,\ldots,A_m\}$及其对应各值域 D_{A_i},可分别定义如下:

定义 **1**. 一个时间序列 S 是有序项队列$\{S_1,S_2,\ldots,S_n\}$,其中,S_i 项是一个 $m+1$ 元组$(t,a_1,\ldots,a_m)$,$t\in T,a_i\in D_{A_i}$,D_{A_i}取实数集.

定义 **2**. 一个事件序列 E 是有序项队列$\{E_1,E_2,\ldots,E_n\}$,其中,E_i 项是一个 $m+1$ 元组$(t,a_1,\ldots,a_m)$,$t\in T,a_i\in D_{A_i}$,D_{A_i}取标称集,由字母表定义.

对于时间序列,每个元组中的属性值 a_i 取实数,典型的如每种股票的日收盘价构成的序列;而事件序列的属性值 a_i 取标称量,如客户浏览某个网站的事件可能是$(100,A),(110,B),(115,C),\ldots,(230,Q)$,其中:每个元组中的数字表示一种时序;字母 A,B,C 等是某种行为的标示.以上是对 m 维序列的定义.当 $m>1$ 时,应考虑各 a_i 间的相关性,增加了研究的难度.现有的多数研究是先基于一维序列研究后,再向多维情形扩展.

定义 **3**. 一个交易序列 C 是有序项队列$\{C_1,C_2,\ldots,C_n\}$,其中,C_i 项是一个 $m+2$ 元组$(p,t,a_1,\ldots,a_m)$,$p\in P$,P 是有限实体集,$t\in T,a_i\in D_{A_i}$,A_i 是交易项目的状态,D_{A_i} 可取实数(交易量)或二元量(是否成交).

交易序列的研究开始于对超市顾客购物行为的分析,即购买商品的关联规则分析,后来推广到其他交易类分析,如股票交易、电信服务等的交叉销售分析.

通常要求时间集 T 是升序的,为方便处理,常假设时间是等距的,即 $t[i+1]-t[i]=\Delta$为常数,这样就可以将 T 映射到自然数集.以上定义仅涉及一个时间维度,更一般地可定义多个时间维度,如交易时间、有效时间等.对两实体对象 Q_1,Q_2,其距离(不相似性)记为 $D(Q_1,Q_2)$.距离度量的一个重要性质是三角不等式,即对于 3 个对象 A,B,C,存在 $D(A,B)\le D(A,C)+D(B,C)$.

作为时态数据相似性发现的基本问题,相似性搜索可简单地描述为:给定一个序列集合 S、一个序列 q、一个距离度量 D 和容错阈值ε,要找出序列集合 $U=\{s\in S|D(q,s)\le\varepsilon\}$.进一步地,可定义搜索 S 中序列的子序列集合,即不仅比较 S 中的序列,而且比较序列中所有与 q 相似的子序列.如果实际搜索的结果 $U'\subset U$,则 $U-U'$为漏报(false dismissals);反之,若 $U'\supset U$,则 $U'-U$ 为多报(false alarms).保证无漏报的相似性匹配称为精确的,否则称为近似的.

时态关联规则是一对二元组(R,F),其中:R 是形如 $X\Rightarrow Y$ 的规则,且 $X\cap Y\neq\varnothing$;F 是时态特征(如有效时期、周期或特定日历).非形式地,时态关联规则的发现是对给定的时态数据集 S,找出所有满足指定支持度和可信度限制的时态关联规则的集合;序列模式的发现是对给定的时态数据集 S,找出所有满足指定支持度的序列的集合;分类是对给定的未标记时态序列 Q,指出 Q 属于两个或若干个预定义类之一;聚类是对给定的时态数据集合 S,找出满足某些相似性度量 $D(q,s)$的自然分组.

对于时态关联规则的发现,时态特征 F 是一个关键性的决定因素,而传统的关联规则是忽略时态的.序列模式的发现主要关注时态序列中按序出现的频繁模式,时间序列应转换成事件序列后再挖掘.

2 序列距离度量

时态数据序列的相似性是通过距离度量来确定的,而距离度量的选择与应用领域高度相关.有多种距离度量:对连续值的时间序列有 L_p 范数等;对事件序列有编辑距离.由于交易项目的稀疏性,交易序列通常不使用成对比较的距离度量.

对于时间序列,最流行的 L_p 范数距离定义如下:

$$L_p(X,Y)=\left(\sum_{i=1}^{l}|x_i-y_i|^p\right)^{1/p} \tag{1}$$

其中,$l=|X|=|Y|$;$p=1,2,\ldots,\infty$.在此,L_1 是 Manhattan 距离,L_2 就是最常用的 Euclidean 距离.L_2 仅适合于两个等长序列比较,且对时间轴变形很敏感,因而在应用中受到许多限制.

为了处理局部时间位移,1994 年,动态时间弯曲(dynamic time warping,简称 DTW)技术被引入数据挖掘领域,但直接用动态规划计算两序列距离的算法复杂性太大($O(mn)$),因而通常将其应用于经变换后的时间序列[4].DTW 通过压缩序列中的非显著特征,使预定义的距离最小化,可以比较两个任意长度序列之间的相似性,主要用于比较沿时间轴的波动模式.时间序列 X,Y 的 DTW 距离可递归定义为

$$\begin{cases}D_{tw}(<>,<>)=0\\ D_{tw}(X,<>)=D_{tw}(<>,Y)=\infty\\ D_{tw}(X,Y)=D_{base}(x_1,y_1)+\min\{D_{tw}(X,Rest(Y)),D_{tw}(Rest(X),Y),D_{tw}(Rest(X),Rest(Y))\}\\ D_{base}(a,b)=|a-b|\end{cases} \tag{2}$$

这里,$X=(x_1,x_2,\ldots,x_m)$,$Y=(y_1,y_2,\ldots,y_n)$,$Rest(X)=(x_2,\ldots,x_m)$,$Rest(Y)=(y_2,\ldots,y_n)$,D_{base} 可按需要选择 L_p 范数.

DTW 的灵活性使其获得大量的应用,但也伴随着可伸缩问题.已经证明 DTW 不满足三角不等式,并有研究表明:任何显式或隐含地规定了满足三角不等式的索引,都将对不符合要求的距离度量产生漏报[5].文献[6]介绍并证明了第 1 个支持时间弯曲且无漏报的索引结构,它是由序列的 4 个特征构成距离函数 D_{tw-lb} 和多维索引,但仅仅 4 个特征显然无法充分发挥多维索引的优势,且查询时若仅使用一个特征将会导致漏报.文献[7]提出并证明了支持 DTW 精确索引的新技术,这种距离度量可看作是将 DTW 转换成上下限序列的 Euclidian 距离.在一些现实的约束下,可为序列 Q 分别指定上/下限序列 U/L,则 Q 与序列 S 的距离为

$$D_{LB}(Q,S)=\sqrt{\sum_{i=1}^{n}\begin{cases}(s_i-U_i)^2, & s_i>U_i\\ (s_i-L_i)^2, & s_i<L_i\\ 0, & \text{否则}\end{cases}} \tag{3}$$

另外,DTW 度量距离采用逐点匹配方式,易受噪音、孤立点干扰.LCSS(longest common subsequence)距离度量[8]能够有效地克服这些缺陷,在序列比较时忽略噪音或孤立点,取两序列的最长公共子序列为其距离度量.ERP (edit distance with real penalty)距离[9]则结合 L_1 与编辑距离的优点,既能处理局部时间位移,又满足三角不等式,因而,可利用下边界和三角不等式策略修剪空间,并用 B+树索引.

对于事件序列的相似性,可以采用编辑距离进行度量[9],其基本思想是:两个序列越相似,则相互变换的代价就越少.可以为一组变换操作(如插入、删除、移动等)定义相应的代价函数,则事件序列的距离就是一个序列变

换为另一个序列所需操作的代价之和.

对于由时态数据建立的模型比较,可以通过比较对应模型及其参数来确定两序列的相似性,如参数的 Euclidean 距离度量,也可以按数据匹配模型的程度来确定序列的相似性.对于确定模型(图模型、确定文法等),相似度量是离散的,匹配结果确定;对于随机模型(Markov 链、随机文法等),匹配结果通常是序列由模型生成的概率.文献[10]介绍一种基于概率模型的相似距离度量,其基本思想是:观测值是由原型模板按照一个先验概率分布"变形"后生成的;距离模型由局部特征组合成全局形状序列,局部特征可能存在变形,其程度由先验概率分布来决定.

对于时间序列,Euclidean 距离的简单性使其成为最流行的距离度量,且对各种数据类型有很好的适用性.DTW 距离能够克服 Euclidean 距离在时间轴的弱点,但有 $O(n^2)$的时间复杂性且会产生漏报.D_{LB}的提出避免了 DTW 的这两个问题(时间复杂性为 $O(n)$),但它要求比较的序列是等长的,还存在弯曲路径的限制.对于事件序列,其整体相似性由序列的每个对应位置的事件相似性决定.模型的相似性则取决于模型参数的匹配.

3 序列表示和搜索

面对海量数据,直接去操作一个高维的数据空间是很困难的.一个具有 n 个点的序列可以看成是 n 维空间的点,若直接用 SAM(spatial access method)多维索引结构(如 R*树)来索引这种 n 维点,则容易导致维度灾难.因此,需要研究合适的数据表示形式,进行维度约简,在高效、方便的表示形式上进行有效的挖掘.衡量维度约简效果的重要标准之一是要满足"无漏报"原则[11],要求数据表示满足以下条件(下边界引理):

$$D_F(q,s)\le D(q,s) \tag{4}$$

即约简后的距离应不大于原先的距离.其中:q 是查询序列;s 是数据集中的任意序列;D_F 是约简空间中的两序列距离;D 是真实的两序列距离.

数据表示形式可分为基于变换、基于模型和其他方法.搜索方法与数据表示形式密切相关.几种主要技术,如离散傅立叶变换、小波变换和奇异值分解等是精确的方法.为提高相似匹配效率,也有学者提倡使用近似的方法,可采用有损耗的数据压缩模式,如分段线性方法、序列离散化、字符串匹配方法等.

3.1 基于变换的方法

基于变换的方法将时间序列从时间域映射到另一个特征空间,用特征空间的映像点表示原始序列,从而实现维度约简,并用多维索引结构存储映像点.另一方面,也可以将时间序列转换成离散的字符序列.

Agrawal 等人[12]使用离散傅立叶变换(discrete Fourier transform,简称 DFT)将时间序列从时域空间变换到频域空间.根据 Parseval 定理,在频域空间中,DFT 保持原序列的 Euclidian 距离,即满足下边界引理.可取频域的前 k 个系数形成一个 k 维点来表示原序列,相似度量就是 k 维点的 Euclidian 距离,并建立索引.DFT 匹配序列必须等长、系数相同,且仅对全序列匹配有效.为了解决子序列匹配的问题,Faloutsos 等人[11]提出设定滑动窗口,对窗口内的子序列进行 DFT 后形成 k 维特征点轨迹,并对轨迹按照 MBR(minimum bounding rectangle)来划分,建立相应索引.这种方法形成基于变换的 GENMINI(generic multimedia indexing)通用框架,其主要步骤是:① 建立距离度量;② 维度约简后适用于 SAM;③ 产生特征空间的距离度量,证明满足下边界引理.

DFT 作为经典的信号变换方法,全局性能良好,但丢失了时间局部化的重要特征.离散小波变换(discrete wavelet transform,简称 DWT)作为一种较新的线性变换技术,利用变换后生成的少数小波参数近似模拟原始信号.其小波参数具有时间/频率特性,可以保持比 DFT 还要多的信息,满足多分辨率的表示需求.文献[12]首次提出用 DWT 代替 DFT 对时间序列进行约简,并证明了 Harr 小波变换保持 Euclidian 距离且满足下边界引理,但未考虑小波的多尺度性.还可以使用全特征小波变换方式将一般小波变换应用于时间序列[13],实验显示,许多小波的性能优于 DFT 和 Harr 小波,且一般小波变换表示也满足下边界引理,但还有待数学证明.

奇异值分解(singular value decomposition,简称 SVD)可应用于压缩时间序列[14],即已知一组 n 维点,将其投影到 k 维子空间($k<n$),变换使得在投影维上的方差最大化.SVD 类似于主成分分析(prime component analysis,简称 PCA),其主要弱点是计算开销大,随着数据的增量加入,索引性能将产生退化,需要周期性地重组索引结构.

分段常数逼近方法[15,16]将序列分成等长的 k 个段,各段的平均值就构成该序列的 k 维特征向量.这种方法的优点是易于理解和实现、转换速度快、无漏报、线性建立索引开销、存在更灵活的距离度量等.对这种 k 维特征向量可建立 L_p 的多形式距离度量和综合索引机制,还可加入偏移变换、幅度/时间伸缩[15].PAA(piecewise aggregate approximation)方法[16]使用更灵活的距离度量,如加权 Euclidian 距离、DWT,并支持比索引项更短的查询,这种能力是 DFT,DWT 和 SVD 所没有的.APCA(adaptive piecewise constant approximation)方法[17]是 PAA 方法的改进,其序列分段是变长的、可索引的,即对波动较大区间划分多个短区段,对平稳区间划分少量长区段,优化表示性能.为了高效、高质量地拟合分段,算法首先利用小波变换,然后还原为变长分段.PAA 和 APCA 都能处理 L_p 范数,APCA 的性能比 DFT 和 DWT 提升了 1~2 个数量级.

以上变换方法都是遵循 GENMINI 框架进行的.DFT, DWT 和 SVD 都是利用多个基本函数的线性组合来表示序列的.DFT 利用正弦和余弦函数,DWT 利用小波基函数,SVD 利用特征波函数(eigenwaves).PAA 和 APCA 利用一系列箱型(box)基函数来表示序列,其中 PAA 的箱型基是定长的,而 APCA 的箱型基是变长的.从可使用的距离度量、性能和效果的综合评价来看,这些方法虽然各有长处,但多数情况下,PAA 是较优的.这些方法的共同不足是:当要求匹配的子序列长度超过滑动窗口时,算法性能明显下降;需要平衡维度约简与准确表示序列的矛盾;大多数算法不能处理小于滑动窗口的子序列匹配问题.

界标方法的理论基础是人类心理和认知科学.对于时间序列,界标是指序列中“最重要”的点(事件).界标形式依赖于应用领域,若曲线上某点处的 n 阶导数为 0,则此点为 n 阶界标,即局部极值点是一阶界标,拐点是二阶界标.文献[18]提出位移、均匀幅度伸缩、均匀时间伸缩、时间弯曲等 6 种界标的变换形式,具有强大灵活、效果稳定的特点.另一种基于界标的方法[6]是通过抽取序列开始、最后、最大和最小元素获得时间弯曲不变的特征向量,可应用于子序列或全序列匹配,与以前的方法相比,在性能上提高了若干数量级.

用字符串表示时间序列的方法已有多种:① 聚类变换方法[19],先由滑动窗口获得子序列集合,再对子序列聚类,并指定相应类符号表示,序列就由这些类符号构成;② 分段划分方法[20],将两点之间具有相同比率的连续点划分成段,并指定某个符号表示;③ 投影方法[21],使用 SOM(self-organizing map)将高维空间中的模式投影到图中的位置,这些位置对应于相应的符号,从而可以有效地减少噪声,降低序列维度;④ 聚合近似,SAX(symbolic aggregate approximation)方法[22]先将序列变换为 PAA 表示,然后符号化为字符串,实现了维度约简且满足下边界引理,这是其独特的优点;⑤ 限幅(clipping)方法[23]将一个实数值序列变换成一个二进制序列,原序列值大于总体均值的为 1,否则为 0,其离散化方式是 SAX 方法的特例.

界标模型将相似度量、数据表示和平滑技术集成于统一的框架内,其表现形式和变换方法的直观性较好.离散化成字符序列的方法可对原序列的实数值形成一种自然的划分表示,有利于应用许多成熟的字符串处理方法,如文法推理方法等.但这些方法的符号表大小、符号、形状选择等带有随意性,容易导致漏报.

3.2 基于模型的方法

基于模型的方法假设数据隐含某种强结构,试图通过数据建模,将模型看作是相应时态数据的发生器.

回归及其混合模型可用于时间序列建模[24].ARMA(auto-regression moving average)模型作为一个信息凝聚器,是时间序列分析中最基本、实际应用最广的模型,但难以实现子序列匹配.ARMA 混合模型可表示变长的时间序列[25].对于 ARIMA(auto-regressive integrated moving average)模型,应先用差分操作移去序列的非平稳性,再对独立 ARMA 模型建模,距离度量使用两序列模型的线性预言编码 cepstrum 的 Euclidean 距离[26].

Markov 链和隐 Markov 模型(hidden Markov model,简称 HMM)是常用的序列建模工具[27−29].HMM 能够同时处理时空维度的不确定性,建模能力更强.文献[30]提出采用基于分割的半 Markov 模型为时间序列建模,即将序列建模为 K 状态分割的隐 Markov 模型,每个分割对应一个状态,并用回归函数对应生成序列各成分;分割之间的距离允许沿时间轴变形;一个序列 S 的模型 M_S,相对于一个查询 Q,其相似度量定义为 $p(Q|M_S)$.

图模型是可以通过变量之间的条件独立关系直接说明的模型,用图形表示.图模型是常用的统计模型,Markov 链就是一种重要的有向图模型.Mannila 等人[31]采用有向图(递归图模型)直观地表示事件序列中的偏序关系,并通过顺序或并行方式组合事件成为复杂片段(episode).进一步的工作[32]将偏序看作是序列的生成

模型,并通过偏序的混合模型来描述序列集合.文法模型是表达字符序列的有力工具,已有许多研究涉及由序列推导出文法的方法.用于推导的文法包括:确定正则文法[33]、上下文无关文法[34]和随机文法[35].前馈神经网络由于缺少时间机制,不适于为时态事件建模.递归神经网络(recurrent neural network,简称 RNN)可以通过内部状态为序列的时态关系明确地建模,递归意味着网络状态不仅依赖于当前输入值,而且与此前的输出有关,还可以从 RNN 中学习字符序列的文法,找出序列演化规律,抽出确定有限状态机形式的规则[21].

在传统的时间序列分析中,已对回归及其混合模型进行了大量的研究,其建模技术已相当成熟,作为一种全局模型,不适于处理子序列.隐 Markov 模型具有坚实的基础,但可伸缩性差.图模型具有广泛的适用性,能灵活地处理序列的全局和局部问题.图模型还可以与神经网络以相互补充的方式来使用,如引入隐含变量减少图结构的复杂性.对于文法模型和 RNN,由于其复杂性,其中仅有少数可用于时间序列的预测.

3.3 其他方法

如前所述,基于三角不等式的索引结构可能产生漏报.后缀树存储序列的所有后缀,而仅保存一次相同前缀,通过深度优先的遍历就可取到每个后缀.作为索引的替代方法,后缀树不涉及距离函数,可实现子序列的精确匹配[36],但对如何实现最优离散化缺乏系统的策略,且庞大的后缀树易导致全序列匹配的性能问题.

用分段线性模型拟合曲线是经典的数学方法.分段线性表示是在最小二乘方误差ε的限制下,循环合并相邻的线性分段而得到的,各线段可增加重要性权重[37].在分段线性表示后,可建立基于散列法的搜索机制.其基本方法[38]是:让一个等距模板格窗沿序列滑动取得各个子序列的二进制位模式串(Hash 关键字),其中,每位的 1 对应于相应模板格中的上升线段,0 对应相应模板格中的不上升线段.相似搜索算法通过全箱修剪、箱排序和箱内修剪技术快速定位相似子序列,但可能导致漏报;通过将线性分段信息保存在 L-index 文件中,并定义一种乐观边界距离度量,可以实现无漏报搜索,并可以处理任意长度的子序列匹配[39].通常认为,取不连接的线性分段比较有利,这样可以实现下边界的 Euclidean 逼近.

另外,Povinelli 等人[40]提出使用时间延迟嵌入方式将时间序列映射到重构相空间,即 $X\to R^Q$.在此,时间序列 $X=\{x_t,t=1,\ldots,N\}$,R^Q 中的点 $x_t=(x_{t-(Q-1)\tau},\ldots,x_{t-2\tau},x_{t-\tau},x_t)^T$,$\tau$是时间延迟,$t\in[(Q-1)\tau+1,N]$,$t$ 是整数.已经证明,若 Q 足够大,则相空间与生成时间序列的状态空间同胚.

很少有文献涉及交易序列的表示.有关的表示方法[41]是:将交易项集对应字母表的某个符号串,这样,有序交易项集就转换成字符序列.为了实现数据压缩,可将相同滑动窗口内的交易当作一次交易[42].可以采用频繁模式树[43]来保存项集关联关系,这适用于在序列上寻找公共事件和显著的关联规则,但不适合分类.

大后缀树的问题在大量数据的挖掘中必须予以特别处理.分段线性方法直观、高比率压缩、噪声不敏感,可使用多种距离度量,但目前还是不可索引的表示方法.不同于上述必须预先假定序列形状的方法,重构相空间方式使用优化方法搜索最优的时态模式,但在序列表示时,确定合适的 Q 是一个难题.由于交易项目的稀疏性和不可比性,如何更有效地表示交易序列仍是值得探索的问题.

4 挖掘操作

4.1 时态关联规则的发现

典型地,关联规则发现的对象是交易序列,目标是从交易数据集(如超市的购物记录)中找出相似的交易关联规则.传统的关联规则发现不考虑时态关系,附加时态特征的时态关联规则可以更好地描述客观规律.

一种基于点和间隔时间模型的时态关联规则学习过程是[44]:① 生成初始关联规则(时态无关);② 剪裁规则,生成关联候选集;③ 计算支持度;④ 建立时态关联规则.在关联规则与时间间隔的约束关系中,发现规则有效周期是重要的,已相应地提出发现规则持续有效时间和周期性的算法[45].进一步的研究发现:交易项本身也有周期;引入项寿命(第一次与最后一次交易的期间)概念后,可将交易计数范围限于项寿命期间[46].

交易序列的关联规则常常带有规律性的循环,如季节性的购物高峰.一个循环规则是在一定时间间隔内周期性出现的规则.文献[47]提出了两种循环关联规则发现算法:一种是先用 Apriori 类算法产生传统关联规则,再找出规则中蕴含的循环关系;另一种是先找出循环的大项集再生成关联关系.后者更为有效.实际上,从不同的

时间间隔可以发现不同的关联规则.为了改进循环规则的表达能力,可以使用日历代数表达时态现象,算法基于用户定制的日历模式搜索近似的关联规则[48].更一般的情况是,使用日历图式作为框架(类型)发现时态模式(具体实例),相应的搜索算法使用时态 aprioriGen 和水平修剪两种优化技术[49].

对于时间序列的规则发现问题,可以通过离散化变换成事件序列形式,再用序列模式挖掘方法导出时态关联规则 $A\overset{T}{\Rightarrow}B$,其中,$T$ 是 A,B 事件的时间间隔约束[19].

先找出关联规则再寻找时态关系的方法效果显然有限;直接寻找时态关联规则的方法难度较大.可以先分时间段找出关联规则,再由各分段的距离推出其规则的时态关系,但如何分段是与领域知识相关的.关键问题之一是时间粒度问题,在不同的时间粒度下可能获得不同的规则以及规则集的结构关系.

4.2 序列模式的发现

如果考虑发生交易的先后顺序,传统关联规则挖掘可以扩充为序列模式挖掘[41].对于超市顾客的交易序列,一个顾客的所有交易按其发生时间的先后排列成一个交易序列,每个交易对应一个项目集.序列的支持度定义为支持该序列的顾客数与总顾客数之比.序列模式发现的目标就是找出所有频繁模式.序列模式挖掘算法可分为 3 类[50]:① 基于 Apriori 的水平格式方法,如 GSP(generalized sequential patterns)算法;② 基于 Apriori 的垂直格式方法,如 SPADE(sequential pattern discovery using equivalence classes)算法;③ 基于投影的模式生长方法,如 PrefixSpan 算法.Apriori 类方法基于反单调特性,使用逐步生成-测试候选集、广度优先搜索策略;模式生长方法不产生候选集,而是利用频繁模式树(FP-树)存储压缩的频繁关键信息,将频繁模式挖掘问题转换成 FP-树挖掘问题,并可使用广度优先或深度优先搜索策略.

Agrawal 等人[41]基于 Apriori 思路,在满足最小支持度的序列中发现最长序列.最长的频繁序列代表一个序列模式,满足最小支持度的序列称为大序列.AprioriAll 算法[41]对所有候选序列计数,然后修剪非最大序列.AprioriSome 和 DynamicSome 算法[41]只关注最大序列,采用先计算更长序列计数的方法,试图避免对那些包含在更长序列中的大序列计数.改进的 GSP 算法[42]减少每次生成的候选数,利用时间约束减少搜索空间.

Mannila 等人[31]将关联规则发现应用到事件序列上,提出片段的定义以及从事件序列中发现频繁片段的 WINEPI,MINEPI 搜索算法.片段是限于一个时间窗口内事件的偏序集,其是否频繁取决于片段出现的窗口数与序列总窗口数之比.进一步考察事件序列的全局偏序关系[32],如果将事件发生归于某个模型,则可使用混合建模技术获得偏序的描述,得到全局事件数据视图.频繁片段本质上是一个单项目交易序列的频繁序列.多数算法每增加一个频繁序列长度就需要扫描全数据库,而抽样又可能因数据偏斜使结果敏感.SPADE 算法[51]将挖掘频繁序列问题按格子理论分解为若干子格子,只需 3 次扫描数据库,就可以分别独立地在内存中使用格子搜索技术和简单连接操作解决频繁模式搜索问题.MOWCATL(minimal occurrences with constraints and time lags)方法[52]从序列中找出周期性片段的事件相关模式,并应用于预测其他序列的类似事件.

Apriori 类算法的性能瓶颈主要是生成大量候选集和多次扫描数据库,作为前缀树的扩充,FP-树方法为解决频繁模式挖掘问题提供了另一类解决途径.基于 FP-树的挖掘方法先后主要有 3 种技术:FreeSpan 方法[53]将频繁项目集循环投影到较小的投影数据库集,并在每个投影数据库中生长子序列片,每次测试被限制在投影数据库内.其缺点是投影必须保存整个序列,子序列的生长需探测候选序列的所有分裂点;PrefixSpan 方法[54]只检查前缀子序列,仅将它们对应的后缀子序列投影到投影数据库中,而且将一个基于内存的伪投影技术用于加速投影操作.在每个投影数据库,序列模式的生长仅限于局部频繁模式;gSpan 方法[50]在一个过程中结合了图模式生长和频繁计数,形成了有效的结构模式挖掘算法.

对于事件序列和交易序列,由于随时间变化的属性值无法简单比较,因而难以缩小搜索空间.为了提高模式发现的有效性和效率,除了利用其统计特性以外,还应使用约束条件.SPIRIT(sequential pattern mining with regular expression constraints)算法[55]基于 Apriori 框架,使用正则表达式表达约束,使用户的需求能合并到挖掘过程中.正则表达式提供了简单、自然的描述方法,更强大的上下文无关文法也已应用于描述约束[56].在模式生长方法的框架中,基于前缀单调特性,可以有效地融合约束与搜索方法,其效率和可伸缩性适合于挖掘大数

据集[57].

对于时间序列,通常先将连续值序列变换成有趣形状或模板序列,即转成事件序列后再进行序列模式的发现.文献[40]基于重构相空间的思想,提出了从时间序列中发现具有特征化和预测功能的有趣模式的算法.该算法将空间重构法与模式发现思想相结合,利用遗传算法实现了时间序列的模式聚类与模式提取,其基本步骤是:① 序列映射到重构相空间;② 形成扩充相空间;③ 形成聚类,获得满足最优目标函数的类.

对于交易项目为布尔值的交易序列,其搜索空间为$(2^A)^P$,其中,A 是交易项目集,P 是最大序列长度.因而,寻找序列模式的搜索空间是巨大的.Apriori 类算法的时间复杂性较高,模式生长方法试图用频繁模式树降低搜索空间,但会大量占用存储,增加内外存交换开销.利用格子理论的搜索空间分区方法,可以有效分解问题,并使用并行算法提高搜索效率,但生成垂直格式数据需要相应的开销.对于搜索的约束条件,使用分类法(taxonomies)也是一种有效的方法.作为分层的领域知识,分类法可以有效地消除冗余规则,还支持多层次的模式发现.传统的支持-置信度框架容易导致组合爆炸,但目前尚未出现公认、通用的度量值.

4.3 分 类

在模式识别、机器学习、数据挖掘等领域中,对分类问题的研究一直很活跃,但时态数据的独特结构使得许多经典算法难以发挥作用.对序列的分类研究相对较少.

时间序列的分类研究的典型特征是:① 用离散技术变换搜索空间;② 注重规则的可解释性;③ 按绝对时间获取规则.多数基于时域的分类方法都假定存在一些简单的序列基本形状或模板,它们是事先给定或能从数据中学习到的.对于以分段线性表示的时间序列,可以通过合并算子获得两个原始输入时间序列的折衷序列.这种基于合并算子的分类方法[37]对某个类中的正例迭代合并,建立一般模型,并使用影响因子来控制合并算子,正影响因子意味着对输入序列的泛化;也可以使用负例来关注形状差异,负影响因子有助于加大正负形状的差异.另一种用决策树组合局部模式的方法[58],可以获得较好的可解释分类结果.它用分段常数逼近[16]约简候选模式的搜索空间,使用回归树递归离散化序列,获得有判决力的局部模式后可构造出二分类规则,最后由决策树组合规则形成分类能力.对于以模型表示的时间序列,因为容易获得序列是否属于某个模型生成的,所以,确定或随机模型可以直接用于获得时间序列分类.

基于重构相空间的方法来源于对动态系统和混沌理论的研究,有坚实的理论基础,避免了序列基本形状的假设.这种新的分类方法[59]将时间序列置于重构相空间中,用全协方差高斯混合模型(Gaussian mixture model,简称 GMM)对其直接建模,EM (expectation maximization)算法估计 GMM 参数.其基本过程是:① 数据分析;② 每个序列类学习 GMM 概率分布;③ 分类序列,使用 Bayesian 最大似然分类器.

在事件序列的分类研究中,FeatureMine 算法[60]运用序列模式挖掘和分类两种技术范式,使用 SPADE 算法[51]快速抽取序列特征,约简潜在的有用特征作为事件序列分类的预处理器.序列特征是一个事件或项的子序列,以特征-值对向量的形式作为分类器的输入.

时态序列的高维度、高特征相关和大量噪音特点,使许多经典方法失效,极大地增加了序列分类的难度.通常,分类方法主要有两步:先获得数据的特征;然后,将这些特征作为输入拟合一个分类模型.实际上,时态序列的特征由一系列与时间有关的特征构成,即是一种序列模式,所以,可利用以上的序列模式发现方法来获得时态数据特征.重构相空间方法由于能够较好地捕捉序列的动态等价性,因而可以获得较好的分类结果.另外,支持向量机(support vector machine,简称 SVM)方法对处理高维空间问题具有良好的特性,但时间复杂度为 $O(n^3)$,较少有相关报道.

4.4 聚 类

聚类序列的目的是找出具有相似演化方式的序列类,其中心问题是:确定序列类数并初始化参数;确定序列之间有意义的相似度量.聚类方法主要有两类:基于距离的方法和基于模型的方法.前者假设数据中仅含弱结构关系;后者假设数据中存在强结构关系.基于距离的划分方法通常假定事先已知聚类个数;而基于距离的层次方法可以通过分裂或凝聚获得聚类个数.层次方法由于计算复杂性较高,显然不能直接适用于大数据集.从数据挖

掘的观点来看,聚类的技术要求是:可伸缩性、可用性、结果可解释性和有效处理高维数据.

4.4.1 基于距离的方法

对时间序列实施聚类,先以限幅方式离散化序列后,可以用 *k*-means 和 *k*-medoids 算法进行聚类[23].I-*k*Means 算法[61]利用多分辨率的小波特性,是一种任意时间的递增的 *k*-means 扩展聚类算法.通过多次、逐步地提高分辨率的聚类,每次结果也为下一次聚类提供初始质心位置,从而避免了 *k*-means 算法可能陷入局部最小的缺陷.任意时间(anytime)算法在任意时点都可以获得当时的最好回答,人为地决定是否继续运行算法,适合于交互、大数据集的处理.I-*k*Means 算法还可以扩展到迭代 EM 算法和 DFT 分解方法.

对于事件序列问题,CLUSEQ 聚类算法[62]根据序列的统计特性来实施聚类.该算法使用给定前导分段的下一符号的条件概率分布来描述序列的行为特征,也作为相似性度量,并将其存储于概率后缀树.而面向模式的凝聚层次算法[63]着重于解决算法可伸缩性,寻找变长的子序列事件,仅关注它们的模式频繁集.

对于交易序列问题,基于大项集的算法[64]适合于大范畴数据集的聚类,还可以通过应用增加聚类内交易项的重叠的全局准则函数来提高聚类速度[65].

层次方法无须提供任何参数就能产生数据对象组的嵌套层次,但其时间开销大(大于 n^2).虽然 *k*-means 有许多缺点(已知类数、球状组、局部最小值等),但仍然是最流行的.I-*k*Means 算法通过 DWT 获得的序列多种分辨率来逐步逼近所需要的聚类效果,形成结果的层次结构,效果较好.但是,对于大数据量、高维度的序列,还应寻找逼近最终结果的捷径.聚类关注序列的结构特征,序列模式发现仍然是基本的寻找特征方法.事件序列一直缺少有效的相似性度量,编辑距离仅捕捉序列间的全局特征,CLUSEQ 算法使用统计特征作为相似性度量可以获得序列的局部特征.在现实应用中,聚类方法常与其他方法一起使用,或作为其他方法的预处理,如先用聚类概括序列,再用关联分析找出其中的最强关系.

4.4.2 基于模型的方法

当类个数 *K* 为已知时,聚类被看作是按混合模型形式为序列建立 *K* 组模型,可以用 EM 算法估计混合系数和模型参数[35];当 *K* 为未知时,可以基于评分函数比较不同类数时的效果后再确定,也可以使用 Monte-Carlo 交叉验证方法[28]或贝叶斯信息标准(Bayesian information criterion,简称 BIC)[25]学习 *K* 值.

BCD(Bayesian clustering by dynamics)方法[27]聚类以 Markov 模型表示的序列.这是一种模型和距离的综合方法,首先将序列表示为 Markov 链和转换概率矩阵,然后使用 Kullback-Leibler 距离度量转换概率矩阵的不相似性,自底向上地迭代搜索合并最接近的序列.

文献[29]提出先使用 DTW 度量来聚类时间序列,获得初始序列类划分,再对每个初始类训练对应的 HMM,并基于序列与各个 HMM 的似然反复调整序列所属类及 HMM 参数,最后为每个类形成一个 HMM.对于回归混合模型,通过联合使用线性随机影响模型,采用基于 MAP(maximum a posteriori)的 EM 算法,避免使用 Monte-Carlo 技术做参数推导,从而降低程序复杂性[24].对于混合的 ARMA 模型,可以使用 EM 算法学习混合系数和各种模型参数,并用 BIC 确定聚类数[25].对于 ARIMA 模型,先移去非平稳性后获得 ARMA 模型,然后用 PAM(partitioning around medoids)划分方法将 ARMA 模型分成若干个组(预先规定组数)[26].

与基于距离的方法相比,基于模型的方法能够以更为自然的方式综合使用先验知识,找出合适的聚类数,还在处理不同长度序列时具有优势.在此,混合模型是常用的方法.直观地,各模型对应于指定的类,由 EM 算法估计聚类参数.模型方法通常需要假定模型形式,EM 算法可能导致局部最小值,还有聚类数的选择等是此类方法的主要问题.目前的解决方法多数是综合使用模型和距离方法,并多次执行算法后确定最佳参数.

4.5 增量挖掘与高阶挖掘

在现实世界,数据集随时间的变化而变化.相应地,有趣模式在多次挖掘期间(session)也会随时间呈现出某种发展变化.已有的规则可能不再有效,而新的有趣模式还有待进一步的发现.通常有两种维护规则的方式:① 强更新,重新计算所有数据,用新规则替换所有老规则;② 弱更新,仅重新计算与增量有关的数据,替换不适用的老规则.考虑到 TDM 的复杂性,挖掘操作更倾向于采用弱更新的方式.增量挖掘关注当数据持续增加时,如何维护已有规则,即给定数据库 *DB*,当新数据集 *db* 添加到 *DB* 时,如何维护 $DB \cup db$ 中规则的问题.

Cheung 首先考虑了关联规则的更新问题[66].对于序列模式的增量挖掘,应考虑时间戳和序列变化,因而比关联规则的维护问题复杂得多.为了找出增量后数据集的所有序列模式,必须确定 *DB* 中的非频繁模式,在 $DB \cup db$ 中是否已变成频繁的.同时,必须检查原有频繁序列是否由于增量而变成非频繁的.ISM(incremental sequence mining)算法[67]利用 SPADE 算法[51]建立增量序列格,使用垂直数据存储方式,保存维护“最大频繁”和“最小非频繁”序列信息.ISE(incremental sequence extraction)算法[68]通过重用以前频繁序列的支持度信息来降低计算频繁序列的开销.IncSpan 算法[69]引入“近似频繁”序列集、逆向匹配和共享投影等新思路进行增量挖掘.

挖掘结果的规则常带有若干参数(如支持度),多次挖掘后即构成规则参数序列.高阶挖掘(higher order mining)[70]试图从规则参数序列中发现有趣的、更易于理解的高层次语义规则,即高阶规则.高阶规则发现需要至少对各原数据集执行一次一阶规则归纳,可按分布方式进行,因而适合于海量挖掘.从规则参数序列中也可发现规则结构的变化,如规则中前件变量的增加.文献[70]研究了对增量数据集合的高阶规则建模和挖掘问题,提出高阶规则的一般建模方法以及获得新知识的框架.在这个框架中,可以集成现有的规则进化、增量挖掘等技术.文献[71]基于一阶时态逻辑给出时间序列挖掘的形式化定义以及归纳高阶规则的方法.

现有的序列模式增量算法大多存在保存的频繁和非频繁信息数量庞大、容易导致内存溢出、需要多次扫描数据库等问题.高阶数据挖掘实际上采用了分段挖掘的思路,但对分段挖掘后如何合并获得整体规则尚待研究.另外,规则结构的演化规律、规则的泛化或特化关系等都是值得深入研究的问题.目前,还很少有对时态数据挖掘的形式化研究,Cotofrei 等人[71]也仅对时间序列规则给出部分形式化定义,且其形式化工作是基于特殊的时序离散化方法,缺少一般性.对于高阶数据挖掘和时态数据挖掘的形式化问题,我们将另文加以讨论.

5 总结与发展方向

近几年来,虽然 TDM 取得了长足进步,但也存在着不足之处.Keogh 等人[72]从 340 篇已发表的论文中选择了最常引用的 56 篇,并用各行业(涉及金融、医学、生物、化学、网络等)的 50 种时间序列数据集进行测试后发现:对多数已发布的改进算法,若改变少许实现细节或用多种多样的实际数据集进行测试,则会发现其性能明显下降,所以,这些改进算法只是对某些特定数据集和某种实现方式是有效的,而缺少解决现实数据问题的普遍有效性.Keogh 等人的工作说明了定量比较多种算法的难度,也表明该领域还拥有广阔的研究空间.开发更健壮、更高效和更准确的时态数据相似性发现算法是学术界面临的主要挑战.

TDM 的相似性发现技术关注现实数据的发展变化,试图从时态数据中发现事物动态演化的相似性规律.由于时态数据的独特结构,传统技术难以对其进行有效处理,亟需研究新的方法加以解决.对于时间序列,通过维度约简等方法变换其表示形式,实现各种相似性发现任务.对于事件序列和交易序列,其表示方法较为简单、直接,需要利用序列的统计特性和约束条件来缩小搜索空间,实现规则发现和聚类.时态数据的增量挖掘与高阶挖掘则主要处理规则维护和规则变化规律的发现.

我们认为,以下几个方面是未来相似性发现技术研究发展的方向:① 序列分类方法,应研究能够克服高特征相关和噪音的特征提取方法;② 多维序列的发现技术,需要提出一套新理论,能够有效地对序列多种属性值相关性进行统一建模;③ 将先验知识、用户反馈与计算智能方法相结合,以缩小搜索空间,提高算法效率和规则有趣性;④ 深入研究增量挖掘和高阶挖掘,包括不同粒度数据集的规则变化、已发现规则的时态特性、规则结构的演化等;⑤ 相似性发现技术的应用,增强算法的实用性和健壮性,将现有成果向应用领域推广.

References:

[1] Agrawal R, Faloutsos C, Swami A. Efficient similarity search in sequence databases. In: David BL, ed. Proc. of the 4th Int'l Conf. on Foundations of Data Organization and Algorithms, FODO'93. Chicago: Springer-Verlag, 1993. 69−84.

[2] Zeng HQ. Research on mining and similarity searching in time series database [Ph.D. Thesis]. Shanghai: Fudan University, 2003 (in Chinese with English abstract).

[3] Roddick J, Spiliopoulou M. A survey of temporal knowledge discovery paradigms and methods. IEEE Trans. on Knowledge and Data Engineering, 2002,14(4):750−768.

[4] Keogh EJ, Pazzani MJ. Scaling up dynamic time warping to massive dataset. In: Zytkow JM, Rauch J, eds. Proc. of the 3rd

European Conf. on Principles of Data Mining and Knowledge Discovery, PKDD'99. Prague: Springer-Verlag, 1999. 1−11.

[5] Yi B, Jagadish H, Faloutsos C. Efficient retrieval of similar time sequences under time warping. In: Sipple RS, ed. Proc. of the 4th Int'l Conf. on Data Engineering, ICDE'98. Orlando: IEEE Computer Society, 1998. 201−208.

[6] Kim SW, Park S, Chu WW. Efficient processing of similarity search under time warping in sequence databases: An index-based approach. Information Systems, 2004,29(5):405−420.

[7] Keogh EJ, Ratanamahatana CA. Exact indexing of dynamic time warping. Knowledge and Information Systems, 2005,7(3): 358−386.

[8] Vlachos M, Hadjieleftheriou M, Gunopulos D, Keogh EJ. Indexing multi-dimensional time-series with support for multiple distance measures. In: Getoor L, Senator TE, eds. Proc. of the 9th ACM SIGKDD 2003. Washington: ACM Press, 2003. 216−225.

[9] Chen L, Ng RT. On the marriage of Lp-norms and edit distance. In: Nascimento MA, Özsu MT, Kossmann D, Miller RJ, Blakeley JA, Schiefer KB, eds. Proc. of the 30th Int'l Conf. on Very Large Data Bases, VLDB 2004. Toronto: Morgan Kaufmann Publishers, 2004. 792−804.

[10] Keogh EJ, Smyth P. A probabilistic approach to fast pattern matching in time series databases. In: Heckerman D, Mannila H, Pregibon D, eds. Proc. of the 3rd Int'l Conf. on Knowledge Discovery and Data Mining. Newport Beach: AAAI Press, 1997. 24−30.

[11] Faloutsos C, Ranganathan M, Manolopoulos Y. Fast subsequence matching in time-series databases. In: Snodgrass RT, Winslett M, eds. Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. Minneapolis: ACM Press, 1994. 419−429.

[12] Chan K, Fu AW. Efficient time series matching by wavelets. In: Kitsuregawa M, Maciaszek L, Papazoglou M, Pu C, eds. Proc. of the 15th Int'l Conf. on Data Engineering, ICDE'99. Sydney: IEEE Computer Society, 1999. 126−133.

[13] Popivanov I, Miller RJ. Similarity search over time series data using wavelets. In: Agrawal R, Dittrich K, Ngu AH, eds. Proc. of the 18th Int'l Conf. on Data Engineering, ICDE 2002. San Jose: IEEE Computer Society, 2002. 212−221.

[14] Korn F, Jagadish H, Faloutsos C. Efficiently supporting ad hoc queries in large datasets of time sequences. In: Peckham J, ed. Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. Tucson: ACM Press, 1997. 289−300.

[15] Yi B, Faloutsos C. Fast time sequence indexing for arbitrary Lp norms. In: Abbadi AE, Brodie ML, Chakravarthy S, Dayal U, Kamel N, Schlageter G, Whang KY, eds. Proc. of the 26th Int'l Conf. on Very Large Data Bases, VLDB 2000. Cairo: Morgan Kaufmann Publishers, 2000. 385−394.

[16] Keogh EJ, Chakrabarti K, Pazzani M, Mehrotra S. Dimensionality reduction for fast similarity search in large time series databases. Knowledge and Information Systems, 2000,3(3):263−286.

[17] Keogh EJ, Chakrabarti K, Mehrotra S, Pazzani MJ. Locally adaptive dimensionality reduction for indexing large time series databases. In: Aref WG, ed. Proc. of the SIGMOD Int'l Conf. on Management of Data. Santa Barbara: ACM Press, 2001. 151−162.

[18] Perng CS, Wang H, Zhang S, Parker DS. Landmark: A new model for similarity-based pattern querying in time series database. In: Young DC, ed. Proc. of the 16th Int'l Conf. on Data Engineering, ICDE 2000. San Diego: IEEE Computer Society, 2000. 33−42.

[19] Das G, Lin K, Mannila H, Renganathan G, Smyth P. Rule discovery from time series. In: Agrawal R, Stolorz PE, Piatetsky-Shapiro G, eds. Proc of the 4th Int'l Conf. on Knowledge Discovery and Data Mining, KDD'98. New York: AAAI Press, 1998. 16−22.

[20] Huang Y, Yu PS. Adaptive query processing for time-series data. In: Proc. of the 5th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. San Diego: ACM Press, 1999. 282−286.

[21] Giles CL, Lawrence S, Tsoi A. Noisy time series prediction using recurrent neural networks and grammatical inference. Machine Learning, 2001,44(1):161−184.

[22] Lin J, Keogh EJ, Lonardi S, Chiu BY. A symbolic representation of time series, with implications for streaming algorithms. In: Zaki MJ, Aggarwal CC, eds. Proc. of the 8th SIGMOD Workshop on DMKD 2003. San Diego: ACM Press, 2003. 2−11.

[23] Bagnall AJ, Janacek GJ. Clustering time series from ARMA models with clipped data. In: Kim W, Kohavi R, Gehrke J, DuMouchel W, eds. Proc. of the 10th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Seattle: ACM Press, 2004. 49−58.

[24] Gaffney SJ, Smyth P. Curve clustering with random effects regression mixtures. In: Bishop CM, Frey BJ, eds. Proc. of the Workshop on Artificial Intelligence and Statistics. Florida: Society for Artificial Intelligence and Statistics, 2003.

[25] Xiong Y, Yeung DY. Time series clustering with ARMA mixtures. Pattern Recognition, 2004,37(8):1675−1689.

[26] Kalpakis K, Gada D, Puttagunta V. Distance measures for effective clustering of ARIMA time-series. In: Cercone N, Lin TY, Wu X, eds. Proc. of the IEEE Int'l Conf. on Data Mining, ICDM 2001. San Jose: IEEE Computer Society, 2001. 273−280.

[27] Ramoni M, Sebastiani P, Cohen P. Bayesian clustering by dynamics. Machine Learning, 2002,47(1):91−121.

[28] Smyth P. Clustering sequences with hidden Markov models. In: Mozer M, Jordan MI, Petsche T, eds. Proc. of the Advances in Neural Information Processing Systems 9, NIPS'96. Cambridge: MIT Press, 1997. 648−654.

[29] Oates T, Firoiu L, Cohen PR. Using dynamic time warping to bootstrap HMM-based clustering of time series. In: Sun R, Giles CL,

eds. Sequence Learning-Paradigms, Algorithms, and Applications. LNAI 1828, Heidelberg: Springer-Verlag, 2001. 35−52.

[30] Ge X, Smyth P. Deformable Markov model templates for time-series pattern matching. In: Proc. of the 6th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Boston: ACM Press, 2000. 81−90.

[31] Mannila H, Toivonen H, Verkamo AI. Discovery of frequent episode in event sequences. Data Mining and Knowledge Discovery, 1997,1(3):259−289.

[32] Mannila H, Meek C. Global partial orders from sequential data. In: Proc. of the 6th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Boston: ACM Press, 2000. 161−168.

[33] Oliveira AL, Silva JPM. Efficient algorithms for the inference of minimum size DFAs. Machine Learning, 2001,44(7):93−119.

[34] Nakamura K, Matsumoto M. Incremental learning of context free grammars. In: Adriaans P, Fernau H, van Zaanen M, eds. Proc. of the 6th Int'l Colloquium Grammatical Inference. ICGI 2002, Amsterdam: Springer-Verlag, 2002. 174−184.

[35] Smyth P. Probabilistic model-based clustering of multivariate and sequential data. In: Heckerman D, Whittaker J, eds. Proc. of the 7th Int'l Workshop on AI and Statistics. Los Gatos: Morgan Kaufmann Publishers, 1999. 299−304.

[36] Park S, Chu WW, Yoon J, Won J. Similarity search of time-warped subsequences via a suffix tree. Information Systems, 2003, 28(7):867−883.

[37] Keogh EJ, Pazzani MJ. An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. In: Agrawal R, Stolorz PE, eds. Proc. of the 4th Int'l Conf. on KDD. New York: AAAI Press, 1998. 239−241.

[38] Keogh EJ, Pazzani MJ. An indexing scheme for fast similarity search in large time series databases. In: Özsoyoglu ZM, ed. Proc. of the 11th Int'l Conf. on Scientific and Statistical Database Management. Cleveland: IEEE Computer Society, 1999. 56−67.

[39] Morinaka Y, Yoshikawa M, Amagasa T, Uemura S. The L-index: An indexing structure for efficient subsequence matching in time sequence databases. In: Industrial Track and Workshops Proc. of the 5th PAKDD 2001. Hong Kong, 2001. 51−60.

[40] Povinelli RJ, Feng X. A new temporal pattern identification method for characterization and prediction of complex time series events. IEEE Trans. on Knowledge and Data Engineering, 2003,15(2):339−352.

[41] Agrawal R, Srikant R. Mining sequential patterns. In: Yu PS, Chen ALP, eds. Proc. of the 11th Int'l Conf. on Data Engineering, ICDE'95. Taipei: IEEE Computer Society, 1995. 3−14.

[42] Srikant R, Agrawal R. Mining sequential patterns: Generalizations and performance improvements. In: Apers PMG, Bouzeghoub M, Gardarin G, eds. Proc. of the 5th Int'l Conf. on Extending Database Technology, EDBT'96. Avignon: Springer-Verlag, 1996. 3−17.

[43] Han J, Pei J, Yin Y, Mao R. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. Data Mining and Knowledge Discovery, 2004,8(1):53−87.

[44] Rainsford CP, Roddick JF. Adding temporal semantics to association rules. In: Zytkow JM, Rauch J, eds. Proc. of the 3rd European Conf. on Principles of Data Mining and Knowledge Discovery, PKDD'99. Prague: Springer-Verlag, 1999. 504−509.

[45] Chen X, Petrounias I. Mining temporal features in association rules. In: Zytkow JM, Rauch J, eds. Proc. of the 3rd European Conf. on Principles of Data Mining and Knowledge Discovery, PKDD'99. Prague: Springer-Verlag, 1999. 295−300.

[46] Ale JM, Rossi GH. An approach to discovering temporal association rules. In: Carroll J, Damiani E, Haddad H, Oppenheim D, eds. Proc. of the 2000 ACM Symp. on Applied Computing. New York: ACM Press, 2000. 294−300.

[47] Özden B, Ramaswamy S, Silberschatz A. Cyclic association rules. In: Sipple RS, ed. Proc. of the 14th Int'l Conf. on Data Engineering, ICDE'98. Orlando: IEEE Computer Society, 1998. 412−421.

[48] Ramaswamy S, Mahajan S, Silberschatz A. On the discovery of interesting patterns in association rules. In: Gupta A, Shmueli O, Widom J, eds. Proc. of the 24th Int'l Conf. on Very Large Data Bases, VLDB'98. New York: Morgan Kaufmann Publishers, 1998. 368−379.

[49] Li Y, Ning P, Wang XS, Jajodia S. Discovering calendar-based temporal association rules. Data and Knowledge Engineering, 2003, 44(2):193−218.

[50] Han J, Pei J, Yan X. From sequential pattern mining to structured pattern mining: A pattern-growth approach. Journal of Computer Science and Technology, 2004,19(3):257−279.

[51] Zaki MJ. SPADE: An efficient algorithm for mining frequent sequences. Machine Learning, 2001,42(1-2):31−60.

[52] Harms SK, Deogun JS. Sequential association rule mining with time lags. Journal of Intelligent Information Systems, 2004,22(1): 7−22.

[53] Han JW, Pei J, Mortazavi-Asl B, Chen QM, Dayal U, Hsu MC. FreeSpan: Frequent pattern-projected sequential pattern mining. In: Proc. of the 6th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Boston: ACM Press, 2000. 355−359.

[54] Pei J, Han JW, Mortazavi-Asl B,Pinto H. PrefixSpan: Mining sequential patterns efficiently by prefix projected pattern growth. In: Young DC, ed. Proc. of the 17th Int'l Conf. on Data Engineering, ICDE 2001. Heidelberg: IEEE Computer Society, 2001. 215−226.

[55] Garofalakis M, Rastogi R, Shim K. Mining sequential patterns with regular expression constraints. IEEE Trans. on Knowledge and Data Engineering, 2002,14(3):530−552.

[56] Antunes CM, Oliveira AL. Inference of sequential association rules guided by context-free grammars. In: Adriaans P, Fernau H, van Zaanen M, eds. Proc. of the 6th Int'l Colloquium Grammatical Inference. ICGI 2002, Amsterdam: Springer-Verlag, 2002. 1−13.

[57] Pei J, Han J, Wang W. Mining sequential patterns with constraints in large databases. In: Proc. of the 2002 Int'l Conf. on Information and Knowledge Management, CIKM 2002. McLean: ACM Press, 2002. 18−25.

[58] Geurts P. Pattern extraction for time series classification. In: Raedt L, Siebes A, eds. Proc. of the 5th European Conf. on Principles of Data Mining and Knowledge Discovery, PKDD 2001. Freiburg: Springer-Verlag, 2001. 115−127.

[59] Povinelli RJ, Johnson MT, Lindgren AC, Ye J. Time series classification using Gaussian mixture models of reconstructed phase spaces. IEEE Trans. on Knowledge and Data Engineering, 2004,16(6):779−783.

[60] Lesh N, Zaki MJ, Ogihara M. Mining features for sequence classification. In: Proc. of the 5th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. San Diego: ACM Press, 1999. 342−346.

[61] Lin J, Vlachos M, Keogh EJ, Gunopulos D. Iterative incremental clustering of time series. In: Bertino E, Christodoulakis S, eds. Proc. of the 9th Int'l Conf. on Extending Database Technology, EDBT 2004. Crete: Springer-Verlag, 2004. 106−122.

[62] Yang J, Wang W. CLUSEQ: Efficient and effective sequence clustering. In: Dayal U, Ramamritham K, Vijayaraman TM, eds. Proc. of the 19th Int'l Conf. on Data Engineering, ICDE 2003. Bangalore: IEEE Computer Society, 2003. 101−112.

[63] Morzy T, Wojciechowski M, Zakrzewicz M. Scalable hierarchical clustering method for sequences of categorical values. In: Cheung DW, ed. Proc. of the 5th Pacific-Asia Conf. on Knowledge Discovery and Data Mining. Hong Kong: Springer-Verlag, 2001. 282−293.

[64] Wang K, Xu C, Liu B. Clustering transactions using large items. In: Proc. of the 1999 Int'l Conf. on Information and Knowledge Management, CIKM'99. Kansas: ACM Press, 1999. 483−490.

[65] Yang Y, Guan X, You J. CLOPE: A fast and effective clustering algorithm for transactional data. In: Hand D, Keim D, Ng R, Zaïane OR, Goebel R, eds. Proc. of the 8th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Edmonton: ACM Press, 2002. 682−687.

[66] Cheung DW, Han J, Ng VT, Wong CY. Maintenance of discovered association rules in large databases. In: Su SYW, ed. Proc. of the 12th Int'l Conf. on Data Engineering, ICDE'96. New Orleans: IEEE Computer Society, 1996. 106−114.

[67] Parthasarathy S, Zaki MJ, Ogihara M, Dwarkadas S. Incremental and interactive sequence mining. In: Proc. of the 1999 Int'l Conf. on Information and Knowledge Management, CIKM'99. Kansas: ACM Press, 1999. 251−258.

[68] Masseglia F, Poncelet P, Teisseire M. Incremental mining of sequential patterns in large databases. Data and Knowledge Engineering, 2003,46(1):97−121.

[69] Cheng H, Yan X, Han J. IncSpan: Incremental mining of sequential patterns in large database. In: Kim W, Kohavi R, Gehrke J, eds. Proc. of the 10th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Seattle: ACM Press, 2004. 527−532.

[70] Spiliopoulou M, Roddick JF. Higher order mining: Modeling and mining the results of knowledge discovery. In: Ebecken N, Brebbia CA, eds. Proc. of the 2nd Int'l Conf. on Data Mining Methods and Databases. Cambridge: WIT Press, 2000. 309−320.

[71] Cotofrei P, Stoffel K. From temporal rules to temporal meta-rules. In: Kambayashi Y, Mohania MK, Wöß B, eds. Proc. of the 6th Int'l Conf. Data Warehousing and Knowledge Discovery, DaWaK 2004. Zaragoza: Springer-Verlag, 2004. 169−178.

[72] Keogh EJ, Kasetty S. On the need for time series data mining benchmarks: A survey and empirical demonstration. Data Mining and Knowledge Discovery, 2003,7(4):349−371.

附中文参考文献:

[2] 曾海权.时间序列挖掘与相似性查找技术研究[博士学位论文].上海:复旦大学,2003.

潘定(1963－),男,江苏宝应人,博士,高级工程师,主要研究领域为数据挖掘,数据仓库.

沈钧毅(1939－),男,教授,博士生导师,CCF高级会员,主要研究领域为数据库理论,数据挖掘.