基于多维时态关联规则的演化模糊推理预测算法

王 玲 [1,2] 孟建瑶 [1,2] 李俊飞 [1,2] 彭开香 [1,2]

摘 要 挖掘时态关联规则的目的是为了发现带有时态信息的项集之间有趣的关系. 由于数据库经常动态更新, 时态关联规则的挖掘也应该适应数据库的更新. 然而, 现有的大多数算法不仅需要重新挖掘更新的数据库, 浪费了大量的时间和效率, 而且不能利用已存在的规则定量地预测某些项的变化趋势. 本文提出了一个基于多维时态关联规则的演化模糊推理预测建模算法 (Evolving fuzzy inference model based on multidimensional temporal association rules, EFI-MTAR), 主要优势是构建了一种基于多维时态关联规则的模糊推理建模算法 (Fuzzy inference modeling algorithm based on multidimensional temporal association rules, FI-MTAR), 实现了对时间序列的定量预测. 此外, 为了降低规则更新的代价和加快规则预测的速度, 提出了概念漂移检测策略来处理时间序列数据以适应数据库的动态更新. 实验结果表明了本文提出算法的有效性和准确性.

关键词 多维时态关联规则, 模糊推理, 演化, 概念漂移

引用格式 王玲, 孟建瑶, 李俊飞, 彭开香. 基于多维时态关联规则的演化模糊推理预测算法. 自动化学报, 2018, **44**(8): 1446−1459

DOI 10.16383/j.aas.2018.c170222

An Evolving Fuzzy Inference Algorithm With Multi-dimensional Temporal Association Rules

WANG Ling[1,2] MENG Jian-Yao[1,2] LI Jun-Fei[1,2] PENG Kai-Xiang[1,2]

Abstract The purpose of mining temporal association rules is to find interesting relationships between item sets with temporal information. Due to the dynamic update of the database, the mining of temporal association rules should adapt to the updates. However, most of the existing algorithms not only need to remine the updated database but also are unable to quantitatively predict the tendency of certainitem. In this paper, an evolving fuzzy inference model based on multidimensional temporal association rules (EFI-MTAR) is proposed to predict the time series quantitatively, In addition, in order to reduce the cost and accelerate the efficiency for prediction, a concept drift detection method is put forward to deal with time series data to adapt to the updates dynamically. Experimental results show the effectiveness and accuracy of the proposed algorithm.

Key words Multi-dimensional temporal association rules, fuzzy inference, evolving, concept drift

Citation Wang Ling, Meng Jian-Yao, Li Jun-Fei, Peng Kai-Xiang. An evolving fuzzy inference algorithm with multi-dimensional temporal association rules. *Acta Automatica Sinica*, 2018, **44**(8): 1446−1459

多维实时序列数据是现实世界中一种普遍存在且具有重要意义的数据类型, 例如工业现场的监控数据、互联网节点的通信流量数据、气象数据、医疗监测数据以及语音视频数据等. 这些数据通常以时间点或时间区间的形式存在, 且呈现出多属性、非平稳、动态性以及信息的非线性等特征, 因此, 如何有效地从这些海量实时序列数据中挖掘有用的知识, 并进行实时预测已经成为各个领域的研究热点.

近年来, 时间序列定量预测技术已经开展了许多方法的研究, 例如文献 [1−2] 通过组合模型的方式来预测时间序列. 文献 [3] 采用时间搜索模型将连续时间序列数据转换为离散空间来挖掘股票数据中的频繁模式, 在此基础上对股票价格进行短期趋势预测. 文献 [4] 通过应用多元自适应回归曲线和逐步回归建立基于指标选择的时间序列模型. 此外, 还通过支持向量回归 (Support vector regression, SVR) 建立预测模型, 并采用遗传算法 (Genetic algorithm, GA) 进行优化. 文献 [5] 针对时序动态关联规则挖掘中, 支持度向量在时间特性上不宜观察其整体变化趋势与预测问题, 提出将小波变换应用到动态关

收稿日期 2017-04-28 录用日期 2017-11-22
Manuscript received April 28, 2017; accepted November 22, 2017
国家自然基科学金 (61572073), 北京科技大学研究生教育发展基金资助
Supported by National Natural Science Foundation of China (61572073) and the Graduate Education Development Funds for University of Science and Technology Beijing
本文责任编委 张敏灵
Recommended by Associate Editor ZHANG Min-Ling
1. 北京科技大学自动化学院 北京 100083 2. 北京科技大学工业过程知识自动化教育部重点实验室 北京 100083
1. School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083 2. Key Laboratory of Knowledge Automation for Industrial Processes, Ministry of Education, University of Science and Technology Beijing, Beijing 100083

联规则挖掘中, 并建立自回归模型预测整体趋势变化. 但这些方法主要还是将时间序列数据作为一个整体来分析和构建全局模型.

值得注意的是, 能够观察到的多维实时序列数据往往仅是系统的部分演化数据, 无法通过历史数据建立一个全局模型覆盖整个数据空间. 为了及时有效地分析多维实时序列数据, 往往更为合理的方式是对时间序列的局部可用数据进行在线分析, 而不是全局序列数据. 关联规则挖掘[6] 可以发现存在于数据中属性之间有趣的关联关系, 以语言和规则相结合的产物呈现最终的挖掘结果, 更易于操作人员的理解和管理者的决策支持. 考虑到时间序列的特性, 衍生出了时态关联规则挖掘方法[7−12], 实际是加了时间约束的关联规则, 目的是找出时态事务集中同一维属性与时间之间的关联, 以及基于时域的不同维属性之间的关系等, 更好地挖掘隐藏在数据中的与时间关联的知识. 因此, 我们的重点是发现多维实时序列数据中的局部模式, 当某一属性发生变化后或导致多个属性紧接着发生变化, 它们之间可能是按时间的顺序关系或因果关系. 针对多维实时序列数据的时态关联规则挖掘吸引了一些学者进行研究. 文献 [13] 提出了基于两个项的时态关系来挖掘时态关联规则, 但在挖掘的过程中对参数的依赖性较强, 且并不适合获取多个项之间的时态关系. 文献 [14] 基于密度的子空间聚类将集群定义为存在于多维数据集的子空间中的高密度区域, 提出一种将多元时间序列转换为符号序列的算法, 能够捕获时间序列变量组之间的相互依赖性和共同变化的顺序模式. 文献 [15] 提出一种新型的多属性时间序列模式, 通过修剪单时间序列中的冗余模式以及避免过度计数的关联, 捕获时间序列变量组之间的相互依赖和共同变化的顺序模式. 文献 [16] 为了避免一段时间出现的频繁项目对结果的影响, 将时间转化为粒度, 并考虑不同级别粒度的时态数据挖掘, 设计了一个三阶段的挖掘框架.

虽然这种时态关联规则已经保留了多维时间序列隐含的过程变化信息, 定性地描述了多维时间序列数据间的时态关系, 但这种知识并没有提供定量预测多维时间序列的未来运动趋势的能力. 通过重构时态关联规则的前件和后件, 并借鉴模糊 Takagi-Sugeno (T-S) 模型的规则形式构造局部模型可以估计多维时间序列的未来值. 此外, 上述方法主要针对离线数据进行挖掘, 但时态数据库通常会动态更新, 时态关联规则挖掘要与数据的变化保持同步, 避免对数据集中所有的数据重新进行挖掘而增加计算复杂度. 因此, 本文提出了基于多维时态关联规则的演化模糊推理算法, 主要包含初始时态关联规则挖掘、基于多维时态关联规则的模糊推理预测、概念漂移检测与系统演化更新三个部分: 1) 采集部分时间序列数据集完成时态关联规则的初始化挖掘; 2) 借鉴区间 TS 模糊推理方法重构时态关联规则, 利用优化算法辨识规则后件参数, 进而实现对时态数据的模糊推理预测; 3) 随着时间序列流数据不断添加到数据库, 需要增量地挖掘时态关联规则. 为了适应时间序列数据发生概念的漂移, 利用滑动窗技术划分时间序列数据. 若当前滑动窗中的数据发生概念漂移, 则认为已有规则库中的时态关联规则已不再适用当前窗口中的数据, 需要对当前窗口中的数据重新进行时态关联规则的挖掘, 并将所得规则按时间顺序存储到规则库中; 若当前滑动窗中的数据没有发生概念漂移, 则需对时态关联规则对应的数据点进行更新, 继续等待下一个时间序列滑动窗口的到来, 直到时间序列数据采集结束.

1 时态关联规则挖掘算法

目前, 针对时态关联规则提出了很多研究方法, 通常我们处理的时间序列对象是数值属性, 离散化后的项所处的时间区间往往不同, 因此, 发现不同时间序列对象不同时间区间的多维数值时态关联规则更加符合现实, 对时态数据的变化趋势具有一定的预测作用. 本文感兴趣的是发现多维时间序列的片段模式之间的先后关联性, 因此必须找到这些片段模式. 我们采用了一种基于多维时间序列形态特征的相似性的动态聚类算法[17] 提取时间序列的片段模式, 基本思想是通过时间序列降维压缩获取时间序列片段, 根据这些时间序列片段的形态特征进行相似性度量, 进而通过动态聚类的方法发现片段模式. 本文以时间区间中的多元时间序列数据集 $S = \{s_1, s_2, \cdots, s_i, \cdots, s_m\}$ $(1 \le i \le m)$ 为例展开说明, 其中, m 表示时间序列属性个数. 令 $\tilde{s}_i = \{I_{i1}, I_{i2}, \cdots, I_{ij}, \cdots, I_{iq_i}\}$ 表示时间序列 s_i 的模式序列项集, I_{ij} 对应片段模式的第 j 个离散化项, q_i 为时间序列 s_i 的片段模式的个数.

本文的多维时态关联规则的挖掘与其他关联规则一样, 就是寻找满足某种时态约束的频繁发生的模式序列. 考虑树结构在规则挖掘中无需产生候选项集等优势, 采用一种基于频繁项集树的时态关联规则挖掘算法[18], 构建树结构与频繁项集挖掘同时进行, 只需要扫描时态数据库一次, 以时间区间为单位表示数据的有效时间区间, 能够有效地计算频繁项集的时态置信度, 提高了规则挖掘效率. 整个算法的核心仍然是寻找频集, 基本思想是: 1) 在多元时间序列数据集降维离散化的基础上, 将所得离散时态事务集转换为布尔离散时态矩阵; 2) 根据布尔离散时态矩阵及向量运算得到时态频繁 1 – 项集和频繁 2 – 项集; 3) 由所得时态频繁项集 (考虑了项集之

间的时态关系) 构建初始频繁项集树, 包含任意两个频繁 1 – 项集间的关联关系, 用于频繁 k ($k \geq 3$) – 项集的生成; 4) 由初始频繁项集树得到完整频繁项集树; 5) 遍历所得完整频繁项集树, 得到所有时态频繁项集; 6) 由所得频繁项集生成强时态关联规则.

2 基于多维时态关联规则的模糊推理预测模型 (FI-MTAR)

时态关联规则挖掘的最终目是通过所得时态关联规则反映数据集中隐藏的时态关联信息, 用于新数据的分类预测等应用. 但目前时态关联规则的形式为

$$\begin{aligned} Rule^{k}_{[T_1,T_2]}: & (I^{k}_{1j_1},[T^{n}_{b},T^{n}_{e}]) \cap \cdots \cap \\ & (I^{k}_{ij_i},[T^{v}_{b},T^{v}_{e}]) \ \cap \cdots \cap \\ & (I^{k}_{mj_m},[T^{1}_{b},T^{1}_{e}])^{[T_1,T_2]}\longrightarrow \\ & (I^{k}_{(m+1)j_{(m+1)}},[T^{0}_{b},T^{0}_{e}]) \end{aligned} \tag{1}$$

其中, $[T^{n}_{b},T^{n}_{e}] \cup \cdots \cup [T^{v}_{b},T^{v}_{e}] \cup \cdots \cup [T^{1}_{b},T^{1}_{e}] \cup [T^{0}_{b},T^{0}_{e}] \in [T_1,T_2]$, 规则的前件中的 $(I^{k}_{1j_i},[T^{n}_{b},T^{n}_{e}])$ 表示第 1 维时间序列在 $t-n$ 时刻对应的时间区间 $[T^{n}_{b},T^{n}_{e}]$ 的第 j_1 类的模式项集, 规则后件中的 $(I^{k}_{(m+1)j_{(m+1)}},[T^{0}_{b},T^{0}_{e}])$ 表示第 $m+1$ 维时间序列在当前 t 时刻对应的时间区间 $[T^{0}_{b},T^{0}_{e}]$ 的第 $j_{(m+1)}$ 类的模式项集. 这里的模式项集对应的时间序列片段模式, 只是定性描述了多维时间序列数据之间的时态区间的变化模式, 因此可以认为目前时态关联规则的形式并不适用对新数据进行定量预测, 为了能够定量地刻画未来时间区间内的实时运动轨迹, 需要对时态关联规则的模式项集定量表征. 考虑到模糊推理模型在复杂系统预测建模中的优势, 本课题拟借鉴模糊 T-S 模型的规则形式, 重构模糊推理时态关联规则以实现定量预测.

2.1 多维区间 T-S 模糊模型的构建

在钢铁、冶金、建材、化工等流程工业中存在一类耗能大、排污大和工艺复杂的大型生产设备, 例如炼铁的高炉、炼钢的转炉、球团竖炉、烧结机以及水泥回转窑等. 它们工艺流程极其复杂, 描述工况和产品质量的参数繁多、工况的自由度难以确定. 由于不确定性的存在, 很难获得精确的输入输出数据, 还没有很好的数据物理方程实现精确描述, 很多情况下, 能够获得的是变量或参数的某一变化范围. 因此, 可以采用区间数建立对象的动力学模型.

近年来, 区间回归[19−20] 和区间时间序列预测[21] 逐渐成为一个新的研究领域. 从本质上看, 现有的区间回归和区间时间序列预测仍是在欧氏空间中建立模型. 文献 [22−24] 提出基于运动模式的建模和控制, 将实际的多维工况模式经主成分分析 (Principal component analysis, PCA) 压缩至一维后, 研究一维运动模式的建模和控制方法. 但实际的工况模式一般都是多维的, 为了更好地描述系统的动力学特性, 需要研究多维运动模式的建模问题. 文献 [25] 以烧结实际生产的数据为例, 利用原始数据构建二维模式运动空间, 然后在模式运动空间中建立二维带输入的区间自回归模型, 描述了烧结终点的动力学特性. 文献 [26] 定义一种多维区间 T-S 模糊模型, 并以此构建多维运动模式的预测模型, 以烧结生产过程的实际数据为例, 验证了所提出的多维运动模式预测模型的有效性.

为了便于描述多维时间序列变化趋势的预测模型, 根据片段模式定义了多维时间序列的区间 T-S 模糊模型.

定义 1. 给定多维区间时间序列 $\tilde{S}(t)$, 其中

$$\begin{aligned} \tilde{S}(t) = \begin{bmatrix} \tilde{s}_1(t) \\ \tilde{s}_2(t) \\ \vdots \\ \tilde{s}_i(t) \\ \vdots \\ \tilde{s}_m(t) \end{bmatrix} = \begin{bmatrix} I_{1j_1}(t) \\ I_{2j_2}(t) \\ \vdots \\ I_{ij_i}(t) \\ \vdots \\ I_{mj_m}(t) \end{bmatrix} = \\ \begin{bmatrix} (y_{c1j_1}(t),\Delta y_{1j_1}(t)) \\ (y_{c2j_2}(t),\Delta y_{2j_2}(t)) \\ \vdots \\ (y_{cij_i}(t),\Delta y_{ij_i}(t)) \\ \vdots \\ (y_{cmj_m}(t),\Delta y_{mj_m}(t)) \end{bmatrix} \end{aligned} \tag{2}$$

其中, $I_{ij_i}(t)$ 为第 i 个时间序列在第 t 个时间区间 $[T^{t}_{b},T^{t}_{e}]$ 的第 j_i 类的模式项集, $y_{cij_i}(t)$ 和 $\Delta y_{ij_i}(t)$ 分别为该模式项集对应区间的数值的中心和区间的数值的幅度变化, $y_{ij_i}(t) \geq 0$, $\Delta y_{ij_i}(t) \geq 0$, $i = 1, 2, \cdots, m$, $j_i \in \{1, 2, \cdots, q_i\}$, 其中 q_i 为时间序列 s_i 经过离散化后得到的模式类别的个数, 则多维区间 T-S 模糊模型的定义为

$$\begin{aligned} R^k: & \textbf{If } y_{c1j_1}(t-1) \textbf{ is } f^{k}_{11} \textbf{ and} \\ & y_{c2j_2}(t-1) \textbf{ is } f^{k}_{21},\cdots, \\ & y_{cmj_m}(t-1) \textbf{ is } f^{k}_{m1},\cdots, \\ & y_{c1j_1}(t-n) \textbf{ is } f^{k}_{1n} \textbf{ and} \\ & y_{c2j_2}(t-n) \textbf{ is } f^{k}_{2n},\cdots, \\ & y_{cmj_m}(t-n) \textbf{ is } f^{k}_{mn} \end{aligned}$$

$$\textbf{Then}\ \tilde{S}(t)=\begin{bmatrix}\tilde{s}_1^k(t)\\ \tilde{s}_2^k(t)\\ \vdots\\ \tilde{s}_i^k(t)\\ \vdots\\ \tilde{s}_m^k(t)\end{bmatrix}=$$

$$\begin{bmatrix}(y_{c1j_1}^k(t),\Delta y_{1j_1}^k(t))\\ (y_{c2j_2}^k(t),\Delta y_{2j_2}^k(t))\\ \vdots\\ (y_{cij_i}^k(t),\Delta y_{ij_i}^k(t))\\ \vdots\\ (y_{cmj_n}^k(t),\Delta y_{mj_m}^k(t))\end{bmatrix}=\begin{bmatrix}\tilde{\theta}_{10}^k\\ \tilde{\theta}_{20}^k\\ \vdots\\ \tilde{\theta}_{i0}^k\\ \vdots\\ \tilde{\theta}_{m0}^k\end{bmatrix}+$$

$$\begin{bmatrix}\tilde{\theta}_{11}^k & \tilde{\theta}_{12}^k & \cdots & \tilde{\theta}_{1i}^k & \cdots & \tilde{\theta}_{1m}^k\\ \tilde{\theta}_{21}^k & \tilde{\theta}_{22}^k & \cdots & \tilde{\theta}_{2i}^k & \cdots & \tilde{\theta}_{2m}^k\\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots\\ \tilde{\theta}_{i1}^k & \tilde{\theta}_{i2}^k & \cdots & \tilde{\theta}_{ii}^k & \cdots & \tilde{\theta}_{mm}^k\\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots\\ \tilde{\theta}_{m1}^k & \tilde{\theta}_{m2}^k & \cdots & \tilde{\theta}_{mi}^k & \cdots & \tilde{\theta}_{mm}^k\end{bmatrix}\times$$

$$\begin{bmatrix}y_{c1j_1}^k(t-1)\\ y_{c2j_2}^k(t-1)\\ \vdots\\ y_{cij_i}^k(t-1)\\ \vdots\\ y_{cmj_m}^k(t-1)\end{bmatrix}+\cdots+$$

$$\begin{bmatrix}\tilde{\theta}_{1(nn-n+1)}^k & \tilde{\theta}_{1(nn-n+2)}^k & \cdots & \tilde{\theta}_{1(nn)}^k\\ \tilde{\theta}_{2(nn-n+1)}^k & \tilde{\theta}_{2(nn-n+2)}^k & \cdots & \vdots\\ \vdots & \vdots & & \vdots\\ \tilde{\theta}_{i(nn-n+1)}^k & \tilde{\theta}_{i(nn-n+2)}^k & \ddots & \vdots\\ \vdots & \vdots & & \vdots\\ \tilde{\theta}_{m(nn-n+1)}^k & \tilde{\theta}_{m(nn-n+2)}^k & \cdots & \tilde{\theta}_{m(nn)}^k\end{bmatrix}\times$$

$$\begin{bmatrix}y_{c1j_1}^k(t-n)\\ y_{c2j_2}^k(t-n)\\ \vdots\\ y_{cij_i}^k(t-n)\\ \vdots\\ y_{cmj_m}^k(t-n)\end{bmatrix}\tag{3}$$

其中, $\tilde{\theta}_{ip}^k$ 为区间参数, f_{ig}^k 为模糊集, $i=1,2,\cdots,m$, $p=0,1,\cdots,nn$, $g=1,2,\cdots,n$, $k=1,2,\cdots,l$, n 为模型的阶次, l 为模型规则数.

多维区间 T-S 模糊模型的最终输出为

$$\tilde{S}(t)=\frac{\sum_{k=1}^{l}w_k\tilde{s}^k(t)}{\sum_{k=1}^{l}w_k}\tag{4}$$

其中,

$$w_k=\prod_{i=1}^{m}\prod_{v=1}^{n}f_{iv}^k(y_{cij_i}(t-v))\tag{5}$$

$f_{iv}^k(y_{cij_i}(t-v))$ 代表 $y_{cij_i}(t-v)$ 隶属于模糊集 f_{iv}^k 的隶属度.

2.2 基于多维时态关联规则的模糊推理预测

目前其他研究者所提算法[19−26] 主要是利用区间模糊逻辑获取确定的模糊推理规则用于预测, 而本文所提算法是基于数据挖掘的结果迅速地实现预测. 通过在多维时态关联规则挖掘算法中最小支持度和最小置信度的分别设置, 所获得的时态关联规则可以明确地表示数据集中的潜在信息.

根据式 (1) 所示的时态关联规则, 当有新的离散时态事务样本 TID_{new} 待预测时, 首先实现规则的匹配, 若满足如下两个条件, 则认为离散时态事务 TID_{new} 与时态关联规则 $Rule_{[T_1,T_2]}^k$ 相匹配.

1) $T_b^{\text{Rule}_{[T_1,T_2]}^k}\le T_b^{TID_{\text{new}}}\le T_e^{TID_{\text{new}}}\le T_e^{\text{Rule}_{[T_1,T_2]}^k}$, 即离散时态事务在时态关联规则的有效时间内出现.

2) $I_{\text{Rule}_{[T_1,T_2]}^k}\subseteq I_{TID_{\text{new}}}$, 即前后件项集 $I_{\text{Rule}_{[T_1,T_2]}^k}$ 包含于项集 $I_{TID_{\text{new}}}$ 中.

对于与规则匹配的新事务, 为了定量地实现对新样本事务输出的精确预测, 重构了基于多维区间时态关联规则的模糊推理预测模型.

$$\begin{aligned}R^k:\ &\textbf{If}\ dx_{c1j_1}(t-1)\ \textbf{is}\ 1\ \textbf{and}\\ &dx_{c2j_2}(t-1)\ \textbf{is}\ 1,\cdots,\\ &dx_{c(m-1)j_{(m-1)}}(t-1)\ \textbf{is}\ f_{(m-1)1}^k,\\ &dx_{cmj_m}(t-1)\ \textbf{is}\ 1,\cdots,\\ &dx_{cij_i}(t-v)\ \textbf{is}\ 1\ \textbf{and}\\ &dx_{c2j_2}(t-1)\ \textbf{is}\ 1,\cdots,\\ &dx_{cij_i}(t-v)\ \textbf{is}\ f_{iv}^k,\cdots,\\ &dx_{cmj_m}(t-1)\ \textbf{is}\ 1,\cdots,\\ &dx_{c1j_1}(t-n)\ \textbf{is}\ f_{1n}^k\ \textbf{and}\\ &dx_{c2j_2}(t-n)\ \textbf{is}\ 1,\cdots,\end{aligned}$$

$dx_{cmj_m}(t-n)$ **is** 1

Then $dx^k_{c(m+1)j_{(m+1)}}(t)=\tilde{\theta}^k_{m0}+$

$$\left[\begin{array}{cccccc}\tilde{\theta}^k_{m1} & \tilde{\theta}^k_{m2} & \cdots & \tilde{\theta}^k_{mi} & \cdots & \tilde{\theta}^k_{mm}\end{array}\right]\times$$

$$\left[\begin{array}{c}dx^k_{c1j_1}(t-1)\\ dx^k_{c2j_2}(t-1)\\ \vdots\\ dx^k_{cij_i}(t-1)\\ \vdots\\ dx^k_{cmj_m}(t-1)\end{array}\right]+\cdots+$$

$$\left[\tilde{\theta}^k_{m(nn-n+1)}\ \tilde{\theta}^k_{m(nn-n+2)}\cdots\tilde{\theta}^k_{m(nn-n+i)}\cdots\tilde{\theta}^k_{m(nn)}\right]\times$$

$$\left[\begin{array}{c}dx^k_{c1j_1}(t-n)\\ dx^k_{c2j_2}(t-n)\\ \vdots\\ dx^k_{cij_i}(t-n)\\ \vdots\\ dx^k_{cmj_m}(t-n)\end{array}\right] \tag{6}$$

其中, $\tilde{\theta}^k_{mp}=(\tilde{\theta}^k_{cmp},\tilde{\theta}^k_{rmp})$ 为区间参数, $\tilde{\theta}^k_{cmp}$ 和 $\tilde{\theta}^k_{rmp}$ 分别为区间参数的中心和半径, $p=0,1,\cdots,nn$, $k=1,2,\cdots,l$, k 为模糊规则数; $[dx^k_{c1j_1}(t-v)\ dx^k_{c2j_2}(t-v)\ \cdots\ dx^k_{cij_i}(t-v)\ \cdots\ dx^k_{cmj_m}(t-v)]^{\mathrm{T}}$ $(v=1,2,\cdots,n)$ 为片段模式 $[dx^k_{1j_1}(t-v)\ dx^k_{2j_2}(t-v)\ \cdots\ dx^k_{ij_i}(t-v)\ \cdots\ dx^k_{mj_m}(t-v)]^{\mathrm{T}}$ 的中心值; 如果某片段模式的隶属函数为 1, 表明该时刻对应的某类项集在此阶段没有变化; $dx^k_{c(m+1)j_{(m+1)}}(t)$ 为第 k 条规则中第 $m+1$ 个时间序列在 t 时刻对应的时间区间 $[T^0_b,T^0_e]$ 内项集 j_{m+1} 所对应样本的模糊推理预测模型输出.

基于多维区间时态关联规则的模糊推理模型的最终输出为

$$dx_{c(m+1)j_{m+1}}(t)=\frac{\sum\limits_{k=1}^{l}w_k dx^k_{cmj_m}(t)}{\sum\limits_{k=1}^{l}w_k} \tag{7}$$

其中,

$$w_k=\prod_{i=1}^{m}\prod_{v=1}^{n}f^k_{iv}(dx_{cij_i}(t-v)) \tag{8}$$

$f^k_{iv}(dx_{cij_i}(t-v))$ 代表 $dx_{cij_i}(t-v)$ 隶属于模糊集 f^k_{iv} 的隶属度, 这里采用区间隶属函数

$$f^k_{iv}(dx_{cij_i}(t-v))=\frac{1}{1+E_1+E_2+E_3} \tag{9}$$

其中,

$$E_1=\exp\left(-\frac{1}{\sigma^k_{ij_i}(t-v)}\times\left(dx_{cij_i}(t-v)-x^k_{ij_i\min}(t-v)\right)\right)$$

$$E_2=\exp\left(\frac{1}{\sigma^k_{ij_i}(t-v)}\times\left(dx_{cij_i}(t-v)-x^k_{ij_i\max}(t-v)\right)\right)$$

$$E_3=\exp\left(\frac{1}{\sigma^k_{ij_i}(t-v)}\times\left(x^k_{ij_i\min}(t-v)-x^k_{ij_i\max}(t-v)\right)\right)$$

$dx_{cij_i}(t-v)$ 表示第 i 个时间序列的 $t-v$ 时刻对应项集中的中心值; $x^k_{ij_i\min}(t-v)$, $x^k_{ij_i\max}(t-v)$ 和 $\sigma^k_{ij_i}(t-v)$ 分别表示第 k 个规则中第 i 个时间序列在 $t-v$ 时刻对应的项集 j_i 包含的样本的最小值、最大值和标准差.

令 $\tilde{w}_k=w_k/\sum_{k=1}^{l}w_k$, 式 (7) 可以进一步整理, 得

$$dx_{c(m+1)j_{m+1}}(t)=\sum_{k=1}^{l}\tilde{w}_k dx^k_{cmj_m}(t)=$$

$$\sum_{k=1}^{l}\tilde{w}_k\left\{\tilde{\theta}^k_{m0}+\left[\tilde{\theta}^k_{m1}\ \tilde{\theta}^k_{m2}\cdots\tilde{\theta}^k_{mi}\cdots\tilde{\theta}^k_{mm}\right]\times\right.$$

$$\left[\begin{array}{c}dx^k_{c1j_1}(t-1)\\ dx^k_{c2j_2}(t-1)\\ \vdots\\ dx^k_{cij_i}(t-1)\\ \vdots\\ dx^k_{cmj_m}(t-1)\end{array}\right]+\cdots+$$

$$\left[\tilde{\theta}^k_{m(nn-n+1)}\ \tilde{\theta}^k_{m(nn-n+2)}\cdots\tilde{\theta}^k_{m(nn-n+i)}\cdots\tilde{\theta}^k_{m(nn)}\right]\times$$

$$\left.\left[\begin{array}{c}dx^k_{c1j_1}(t-n)\\ dx^k_{c2j_2}(t-n)\\ \vdots\\ dx^k_{cij_i}(t-n)\\ \vdots\\ dx^k_{cmj_m}(t-n)\end{array}\right]\right\}=$$

$$\sum_{k=1}^{l}\tilde{w}_k\left[(\theta_{cm0},\theta_{rm0})\right]+\sum_{k=1}^{l}\tilde{w}_k\times$$

$$\Big[(\tilde{\theta}^k_{cm1},\tilde{\theta}^k_{rm1})(\tilde{\theta}^k_{cm2},\tilde{\theta}^k_{rm2})\cdots(\tilde{\theta}^k_{cmi},\tilde{\theta}^k_{rmi})\cdots$$
$$(\tilde{\theta}^k_{cmm},\tilde{\theta}^k_{rm1})\Big]\left[dx^k_{c1j_1}(t-1)dx^k_{c2j_2}(t-1)\cdots\right.$$
$$\left.dx^k_{cij_i}(t-1)\cdots dx^k_{cmj_m}(t-1)\right]^{\mathrm{T}}+\cdots+$$
$$\sum_{k=1}^{l}\tilde{w}_k\Big[(\tilde{\theta}^k_{cm(nn-n+1)},\tilde{\theta}^k_{rm(nn-n+1)})\times$$
$$(\tilde{\theta}^k_{cm(nn-n+2)},\tilde{\theta}^k_{rm(nn-n+2)})\cdots$$
$$(\tilde{\theta}^k_{cm(nn-n+i)},\tilde{\theta}^k_{rm(nn-n+i)})\cdots(\tilde{\theta}^k_{cm(nn)},\tilde{\theta}^k_{rm(nn)})\Big]\times$$
$$\left[dx^k_{c1j_1}(t-n)dx^k_{c2j_2}(t-n)\cdots dx^k_{cij_i}(t-n)\cdots\right.$$
$$\left.dx^k_{cmj_m}(t-n)\right]^{\mathrm{T}}=$$
$$\sum_{k=1}^{l}\tilde{w}_k\Big[((\tilde{\theta}^k_{mc})^{\mathrm{T}}X(t),(\tilde{\theta}^k_{mr})^{\mathrm{T}}|X(t)|)\Big]=$$
$$\Big[((\tilde{\theta}^{\mathrm{T}}_{mc}\tilde{X}(t),(\tilde{\theta}^{\mathrm{T}}_{mr}|\tilde{X}(t)|)\Big] \tag{10}$$

其中,

$$(\tilde{\theta}^k_{mc})^{\mathrm{T}}=\left[\theta^k_{cm0},\theta^k_{cm1},\theta^k_{cm2},\cdots,\theta^k_{cm(nn-1)},\theta^k_{cm(nn)}\right]$$
$$(\tilde{\theta}^k_{rc})^{\mathrm{T}}=\left[\theta^k_{rm0},\theta^k_{rm1},\theta^k_{rm2},\cdots,\theta^k_{rm(nn-1)},\theta^k_{rm(nn)}\right]$$
$$\tilde{\theta}^{\mathrm{T}}_{mc}=\left[(\tilde{\theta}^1_{mc})^{\mathrm{T}},(\tilde{\theta}^2_{mc})^{\mathrm{T}},\cdots,(\tilde{\theta}^l_{mc})^{\mathrm{T}}\right]$$
$$\tilde{\theta}^{\mathrm{T}}_{mr}=\left[(\tilde{\theta}^1_{mr})^{\mathrm{T}},(\tilde{\theta}^2_{mr})^{\mathrm{T}},\cdots,(\tilde{\theta}^l_{mr})^{\mathrm{T}}\right]$$
$$\tilde{X}(t)=\sum_{k=1}^{l}w_k\left[1,dx_{c1j_1}(t-1),dx_{c2j_2}(t-1),\cdots,\right.$$
$$dx_{cmj_m}(t-1),\cdots,dx^k_{c1j_1}(t-n),$$
$$\left.dx^k_{c2j_2}(t-n),\cdots,dx^k_{cmj_m}(t-n)\right]$$
$$\left|\tilde{X}(t)\right|=\sum_{k=1}^{l}w_k\left[1,\left|dx_{c1j_1}(t-1)\right|,\right.$$
$$\left|dx_{c2j_2}(t-1)\right|,\cdots,$$
$$\left|dx_{cmj_m}(t-1)\right|,\cdots,\left|dx^k_{c1j_1}(t-n)\right|,$$
$$\left.\left|dx^k_{c2j_2}(t-n)\right|,\cdots,\left|dx^k_{cmj_m}(t-n)\right|\right]$$

2.3　模型参数的辨识

基于多维时态关联规则的模糊推理预测模型中的区间参数可通过使下式中的目标函数 J 在约束条件 $\theta^k_{rmp}\geq 0$ 取得极小求得, 其中, $p=0,1,2,\cdots,nn$, 且有

$$\min_{\theta_{mc},\theta_{rc}}J=\sum_{t=1}^{n}\left[dx^L_{cmj_m}(t)-d\tilde{x}^L_{cmj_m}(t)\right]^2+$$
$$\sum_{t=1}^{n}\left[dx^U_{cmj_m}(t)-d\tilde{x}^U_{cmj_m}(t)\right]^2$$
$$\text{s.t.}\quad\theta^k_{rmp}\geq 0,\quad k=1,2,\cdots,l \tag{11}$$

其中, $d\tilde{x}^U_{cmj_m}(t)$ 和 $d\tilde{x}^L_{cmj_m}(t)$ 分别为预测模型输出 $d\tilde{x}_{cmj_m}(t)$ 的上界和下界, $dx^U_{cmj_m}(t)$ 和 $dx^L_{cmj_m}(t)$ 分别为真实模式类别变量 $dx_{cmj_m}(t)$ 的上界和下界.

为了求解式 (11) 中的约束优化问题, 对该式进行整理, 根据文献 [26] 进一步得到与式 (11) 同解的二次优化问题.

$$\min_{\theta_{mc},\theta_{rc}}J=\left[\begin{array}{cc}\theta^{\mathrm{T}}_{mc} & \theta^{\mathrm{T}}_{mr}\end{array}\right]\times$$
$$\left[\begin{array}{cc}\sum\limits_{t=1}^{n}2X(t)X^{\mathrm{T}}(t) & \\ & \sum\limits_{t=1}^{n}2\left|X(t)\right|\left|X(t)\right|^{\mathrm{T}}\end{array}\right]\times$$
$$\left[\begin{array}{c}\theta_{mc}\\ \theta_{mr}\end{array}\right]-\left[\begin{array}{cc}\theta^{\mathrm{T}}_{mc} & \theta^{\mathrm{T}}_{mr}\end{array}\right]\times$$
$$\left[2\sum_{t=1}^{n}X(t)(dx^U_{c(m+)j_{m+1}}(t)+dx^L_{c(m+1)j_{m+1}}(t)),\right.$$
$$\left.2\sum_{t=1}^{n}\left|X(t)\right|(dx^U_{c(m+1)j_{m+1}}(t)-dx^L_{c(m+1)j_{m+1}}(t))\right]^{\mathrm{T}} \tag{12}$$

3　概念漂移检测和系统演化更新

3.1　概念漂移的检测

概念漂移[27] 是指随着时间推移或时序数据的更新, 时间序列数据自身分布及结构等发生变化. 通过对时序数据的概念漂移检测, 发现当前数据与原先状态的异同, 进而判断规则库中的已有关联规则是否适用于当前时序数据, 从而作出相应处理. 判断当前滑动窗是否发生概念漂移, 主要考虑以下两种情况, 只要满足任意一种情况, 都认为发生了概念漂移:

1) 规则覆盖率

如前所述, 如果满足规则匹配的条件, 则在此基础上统计没有规则可匹配的离散时态事务数 m. 若 $(m/N)>\varepsilon_{th}$, 即如果当前读取的时间序列滑动窗对应的离散时态数据集中无规则可匹配的样本事务所占比例大于给定阈值, 则发生了概念漂移. 其中, N 表示离散时态事务集中所含离散时态事务数, m 是离散时态事务集中无规则匹配的离散时态事务数, ε_{th} 是规则不匹配比例阈值, 本文取值范围设置为 [0.3, 0.5].

2) 时间序列的相似度

假定当前时间序列滑动窗的时间区间为 $[T_1, T_2]$, 则在时间 $t=T_1-1$ 处, 表示上一个滑动窗演化更新处理完成, 系统已有时态关联规则的所有

候选 1−项集及其支持度对 $\langle C_1, \sup(C_1)\rangle$ 构成的集合为 $C_1^{T_1-1}$, 所有频繁 1−项集及其支持度对 $\langle F_1, \sup(F_1)\rangle$ 构成的集合为 $F_1^{T_1-1}$. 若集合 $F_1^{T_1-1}$ 与集合 $F_1^{T_2}$ (表示当前时间 T_2 获得的所有频繁 1−项集) 所含项集元素不同, 则需对集合 $C_1^{T_1-1}$ 及集合 $C_1^{T_2}$ 中所含项集进行如下处理: 对两集合所含项集取并集 (时间也合并), 所得项集构成集合 $C^{[T_1-1,T_2]}$, 对比集合 $C^{[T_1-1,T_2]}$ 中所含项集, 将集合 $C_1^{T_1-1}$ 及集合 $C_1^{T_2}$ 中不含有的项集补全, 对应支持度记为零, 这样可以将集合 $C_1^{T_1-1}$ 及 $C_1^{T_2}$ 中所含项集元素变为相同. 将集合 $C_1^{T_1-1}$ 及 $C_1^{T_2}$ 中所含项集支持度看成是两等长序列, 分别记为 $s_{C_1^{T_1-1}}$ 和 $s_{C_1^{T_2}}$, 利用式 (13) 计算两序列间的相似性 $sim(s_{C_1^{T_1-1}}, s_{C_1^{T_2}})$, 若 $sim(s_{C_1^{T_1-1}}, s_{C_1^{T_2}}) < \eta_{th}$, 则认为当前数据发生了概念漂移, 其中 η_{th} 是用户给定的最小相似性阈值, 本文取值范围设置为 $[0.7, 0.9]$.

$$sim(s_{C_1^{T_1-1}}, s_{C_1^{T_2}}) = \frac{\sum\limits_{i=1}^{p} s_{C_1^{T_1-1}}(i) \times s_{C_1^{T_2}}(i)}{\sqrt{\sum\limits_{i=1}^{p} s^2_{C_1^{T_1-1}}(i)}\sqrt{\sum\limits_{i=1}^{p} s^2_{C_1^{T_2}}(i)}} \tag{13}$$

其中, p 是序列长度, $s_{C_1^{T_1-1}}(i)$, $s_{C_1^{T_2}}(i)$ 是序列 $s_{C_1^{T_1-1}}$, $s_{C_1^{T_2}}$ 中第 i 个元素.

若时间序列滑动窗发生概念漂移, 则需利用时态关联规则挖掘算法对滑动窗重新进行基时态关联规则的挖掘, 并对规则库进行规则更新, 进而实现基于多维时态关联规则的模糊推理预测建模; 否则, 认为当前已有时态关联规则与滑动窗相匹配, 无需进行时态关联规则的挖掘更新, 继续等待下一个滑动窗, 实现系统演化更新.

3.2 系统演化更新

系统演化更新是指在充分利用已有规则库 R 的基础上, 对系统中新加入的时间序列滑动窗中的数据进行选择性处理, 从而提高数据挖掘的效率, 实现系统的自适应变化. 本文采用滑动窗技术分析系统的演化更新. 根据 Hoeffding[28] 边界检测数据的分布, 自动确定滑动窗口的大小. 具体的滑动窗口的实现过程见图 1 所示.

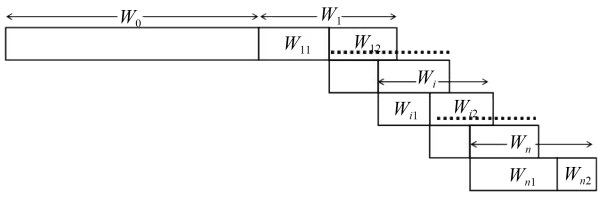


图 1 滑动窗的实现过程

Fig. 1 The implementation process of sliding window

图 1 中, w_0 表示系统进行初始关联规则挖掘的滑动时间窗的大小, w_i $(1 \le i \le n)$ 表示系统在第 i 次演化更新时滑动窗口的大小, 其中, w_{i1} $(0 \le |w_{i1}| \le |w_i|)$ 表示当前滑动窗口中与规则匹配的样本数据窗口大小, w_{i2} $(0 \le |w_{i2}| \le |w_i|)$ 表示当前滑动窗中无规则匹配的样本数据窗口大小, 且 $|w_{i1}| + |w_{i2}| = |w_i|$; 根据概念漂移的第一种情况, 如果当前窗口中无规则匹配的事务数所占比例大于给定阈值, 则保留窗口 w_{i2} 中的样本数据, 删除窗口 w_{i1} 中的样本数据, 同时接收新的数据填充窗口 w_{i1}, 在更新数据的基础上重新挖掘. 系统演化更新的具体流程如图 2 所示, 首先, 当时间序列数据库读入的时间序列数据达到设定时间长度, 则存储当前时间序列滑动窗的数据. 然后, 通过概念漂移模块对读取的时间序列滑动窗进行概念漂移检测, 判断当前时间序列数据是否发生改变或是否适用当前规则库中的关联规则.

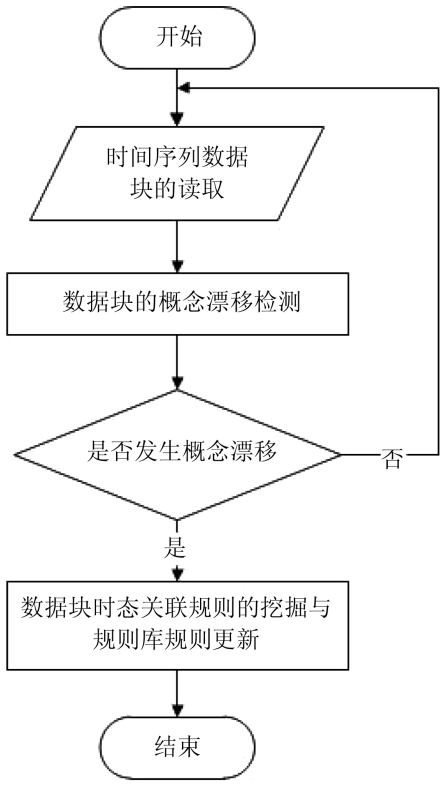


图 2 系统演化更新

Fig. 2 The update process of system evolving system evolving window

4 性能评估

为了对所提算法的性能及有效性进行验证, 实验使用 UCI 数据库[29] 中的多个时间序列数据集 (Air Quality 数据集, Istanbul 数据集及 Synthetic Control Chart 数据集) 分别设计并对比 3 种不同的

实验方案 (如表 1 所示). 首先, 考虑到本文提出的算法 (方案 2) 采用了一种频繁项集树结构挖掘时态关联规则, 为了对比类似挖掘算法的影响, 选择方案 1 进行对比研究, 其采用了 FP-growth 树结构来挖掘时态关联规则. 其次, 考虑到本文提出的算法 (方案 2) 采用了模糊推理进行预测, 而我们正是借鉴了 TS 模糊推理方法的思路, 为此, 采用方案 3 进行对比研究.

表 1　对比方案

Table 1　The comparison program

对比方案	方案简述
方案 1	利用 FP-growth[30] 算法进行时态关联规则挖掘, 然后利用 FI-MTAR 算法进行推理预测
方案 2	利用时态关联规则算法[18] 和 EFI-MTAR 算法演化模糊推理
方案 3	直接利用 TS 模糊推理[31] 进行推理预测

定义 2 (分类正确率). 对于包含 n 条离散事务的离散事务集, 将每条离散事务与已有规则进行匹配, 若存在 m 条离散事务与已有规则匹配, 则分类正确率为

$$CA = \frac{m}{n} \times 100\,\% \tag{14}$$

定义 3 (均方根误差 *RMSE*).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y} - y_i)^2}{n}} \tag{15}$$

其中, n 表示数据样本个数, $\hat{y}$ 表示样本预测输出, y_i 表示样本实际输出.

4.1　基于多维时态关联规则的模糊推理预测模型的验证

为了验证基于时态关联规则的模糊推理预测性能以及更好地理解规则的物理意义, 以 Air Quality 数据集中部分数据挖掘的时态关联规则为例进行仿真实验和说明.

$$(21\,|[35,36]) \cap (31\,|\ [37,\,40]) \xrightarrow{[1,49]} (61\,|[41,45])\,,$$
$$\sup = 0.17,\ \ \mathrm{conf} = 0.73,\ \ \mathrm{imp} = 0.70 \tag{16}$$

规则的前件包含项 {21}、{31} 以及它们发生的有效时间区间 [35, 36] 和 [37, 40], 规则的后件包含项 {61} 以及它发生的有效时间区间 [41, 45]. 根据时间序列变量离散化得到的这些项集, 可以对应到 Air Quality 数据集的时间序列的序列片段模式, 如表 2 所示.

表 2　离散化项集的序列片段模式

Table 2　The segment patterns of the time series for the discrete item

时间序列变量	离散化项	斜率均值	斜率对应角度	幅值变化均值	归一化变化幅值均值
PT08.S2(NMHC)	21	74.25	89.25	144.2	0.48
NOx(GT)	31	−31.6	−87.1	−92.4	−0.16
PT08.S5(O3)	61	80.8	89.3	911.1	0.45

因此, 规则可以进一步表示为

$$\begin{aligned}&(\mathrm{NOx\,(GT)}\,\{\bar{k} = -31.6; \Delta\bar{y} = -92.4\}\\ &\quad |[2004.03.11.01, 2004.03.11.02])\,,\\ &(\mathrm{PT08.S2\,(NMHC)}\,\{\bar{k} = 74.25; \Delta\bar{y} = 144.2\}\\ &\quad |[2004.03.11.03, 2004.03.11.05])\\ &\xrightarrow{[2004.03.10.18, 2004.03.11.12]}\\ &(\mathrm{PT08.S5(O3)}\{\bar{k} = 80.8; \Delta\bar{y} = 911.1\}\\ &\quad |[2004.03.11.06, 2004.03.11.10])\,,\\ &\sup = 0.17,\ \ \mathrm{conf} = 1,\ \ \mathrm{imp} = 0.60\end{aligned} \tag{17}$$

其中, $\bar{k}$ 表示斜率的均值; $\Delta\bar{y}$ 表示幅值变化的均值. 但形如式 (17) 的规则还是难以理解其时间序列的变化趋势, 为此, 进一步给出了序列片段模式的语义描述, 如表 3 所示.

表 3　序列片段模式的语义描述

Table 3　The semantic description of the segment patterns

斜率对应角度范围	语义描述	归一化变化幅值范围	语义描述
$[-90, -60]$	剧烈下降	$[-1, -0.6]$	大幅下降
$[-60, -30]$	快速下降	$[-0.6, -0.3]$	中幅下降
$[-30, 0]$	平稳下降	$[-0.3, 0]$	小幅下降
$[0, 30]$	平稳上升	$[0, 0.3]$	小幅上升
$[30, 60]$	快速上升	$[0.3, 0.6]$	中幅上升
$[60, 90]$	剧烈上升	$[0.6, 1]$	大幅上升

上述规则可以进一步描述为: 在 2004 年 3 月 10 日 18 时到 2004 年 3 月 11 日 12 时的时间区间中, 时间序列 NOx(GT) 如果在时间 2004 年 3 月

11 日 1 时到 2004 年 3 月 11 日 2 时内 (即时间区间 [35, 36]) 按照快速小幅下降的趋势变化, 而时间序列 PT08.S2(NMHC) 在时间 2004 年 3 月 11 日 3 时到 2004 年 3 月 11 日 5 时内 (即时间区间 [37, 39]) 按照剧烈中幅上升的趋势变化, 则可以得到时间序列 PT08.S5(O3) 在时间 2004 年 3 月 11 日 6 时到 2004 年 3 月 11 日 10 时内 (即时间区间 [41, 45]), 将会按照剧烈中幅上升的趋势变化.

在此基础上, 仿真实验采用本文提出的基于多维时态关联规则的模糊推理模型进行预测, 具体形式为

$$
\begin{aligned}
R:\ &\textbf{If}\ dx_{c1j_1}(t-1)\ \textbf{is}\ 1\ \textbf{and}\\
&dx_{c2j_2}(t-1)\ \textbf{is}\ f_{21},\ dx_{c3j_3}(t-1)\ \textbf{is}\ 1,\\
&dx_{c1j_1}(t-2)\ \textbf{is}\ f_{12}\ \textbf{and}\\
&dx_{c2j_2}(t-1)\ \textbf{is}\ 1,\ dx_{c3j_3}(t-2)\ \textbf{is}\ 1\\
&\textbf{Then}\ dx_{c3j_3}(t)=\\
&\tilde{\theta}_{30}+\begin{bmatrix}\tilde{\theta}_{31} & \tilde{\theta}_{32} & \tilde{\theta}_{33}\end{bmatrix}\begin{bmatrix}dx_{c1j_1}(t-1)\\ dx_{c2j_2}(t-1)\\ dx_{c3j_3}(t-1)\end{bmatrix}+\\
&\begin{bmatrix}\tilde{\theta}_{34} & \tilde{\theta}_{35} & \tilde{\theta}_{36}\end{bmatrix}\begin{bmatrix}dx_{c1j_1}(t-2)\\ dx_{c2j_2}(t-2)\\ dx_{c3j_3}(t-2)\end{bmatrix} \qquad (18)
\end{aligned}
$$

进一步整理得到:

$$
\begin{aligned}
R:\ &\textbf{If}\ dx_{c2j_2}(t-1)\ \textbf{is}\ f_{21}\ \textbf{and}\\
&dx_{c1j_1}(t-2)\ \textbf{is}\ f_{12}\\
&\textbf{Then}\ dx_{c3j_3}(t)=\\
&\tilde{\theta}_{30}+\tilde{\theta}_{32}dx_{c2j_2}(t-1)+\tilde{\theta}_{34}dx_{c1j_1}(t-2)\quad (19)
\end{aligned}
$$

其中, 模型的阶次为 2, $dx_{c1j_1}(t-2)$ 对应的是时态关联规则前件中的项 $(21|[35,36])$, $dx_{c2j_2}(t-1)$ 对应的是时态关联规则前件中的项 $(31|[37,40])$, $dx_{c3j_3}(t)$ 对应的是时态关联规则后件中的项 $(61|[41,45])$, 采用其中 30 条样本用于训练, 10 条样本用于测试, 由式 (10) ∼ (12), 辨识得到的该条规则的模糊推理模型的后件参数为

$$
\begin{aligned}
&\begin{bmatrix}\tilde{\theta}_{30} & \tilde{\theta}_{32} & \tilde{\theta}_{34}\end{bmatrix}=\\
&\begin{bmatrix}(0.092, 31.2) & (1.0181, 72.1) & (0.2182, 405)\end{bmatrix} \qquad (20)
\end{aligned}
$$

基于本文提出的基于多维时态关联规则的模糊推理模型进行预测, 表 4 给出了最终预测输出的上下界的均方根误差.

表 4 最终预测输出上下界均方根误差

Table 4 The RMSE of upper bound and lower bound for the prediction output

$dx^L_{c3j_3}$	$dx^U_{c3j_3}$
0.1656	0.1803

4.2 可扩展性的评估

为验证滑动窗口大小对系统演化更新效果的影响, 该部分利用 Air Quality 时间序列数据集, 在系统演化更新过程中, 分别人为定义滑动窗口大小为 5 %, 10 %, 15 %, 20 % 和 25 %, 利用自动确定滑动窗口大小的方法确定滑动窗口大小为 7 %, 对比不同滑动窗口大小情况下, 系统演化更新过程中产生的时态关联规则的规则数、规则分类正确率及预测均方根误差, 结果如表 5 所示.

表 5 不同滑动窗口的演化更新效果

Table 5 The evolution effect of different sliding window size

评价指标	演化次数	滑动窗口大小 5 %	10 %	15 %	20 %	本文方法 7 %
Rules	0	16	17	19	19	17
	1	20	19	22	24	18
	2	20	21	24	26	20
	3	21	21	25	29	21
	4	24	27	29	33	21
	5	25	24	30	34	24
	6	25	29	31	33	23
CA	0	0.909	0.899	0.876	0.866	0.911
	1	0.879	0.904	0.851	0.851	0.906
	2	0.886	0.909	0.846	0.839	0.916
	3	0.891	0.896	0.849	0.832	0.919
	4	0.876	0.887	0.851	0.842	0.924
	5	0.869	0.879	0.854	0.847	0.896
	6	0.881	0.896	0.862	0.859	0.909
RMSE	0	0.144	0.152	0.187	0.197	0.137
	1	0.179	0.145	0.221	0.224	0.145
	2	0.168	0.139	0.234	0.264	0.121
	3	0.159	0.157	0.227	0.279	0.112
	4	0.187	0.165	0.224	0.241	0.104
	5	0.194	0.183	0.209	0.236	0.154
	6	0.175	0.156	0.201	0.214	0.146

不同滑动窗口大小导致演化次数不同, 滑动窗口大小为 10 % 时, 演化次数为 12 次; 滑动窗口为 25 % 时, 演化次数为 6 次, 为了方便比较, 本文选择共有演化次数为 6 次. 如表 5 所示, 在不同大小的滑动窗口下系统演化更新时所得规则数不同, 且滑动窗口越大, 产生的规则数也相对较多; 自动确定滑动窗口中均具有最高的分类正确率; 针对不同演化次数, 系统所得规则的预测误差的波动变化不大. 同一演化次数中, 自动确定的滑动窗口中均具有最低的预测均方根误差.

4.3　有效性和准确性的验证

为验证本文提出的方案 2 在获取规则的个数、分类准确率 (如表 6 所示) 和预测性能 (如表 7 所示) 方面的有效性, 首先通过算法自动确定滑动窗口大小, 进而确定各个数据集的演化次数, 分别为 13 次、6 次和 7 次.

从表 6 可以看出, 与方案 1 重新挖掘规则相比, 由于采用了概念漂移处理机制, 方案 2 的演化更新系统不仅减少了挖掘规则的数量, 而且节省了存储空间, 提高了算法效率, 证明了方案 2 的有效性. 此外, 通过对比两种方案所得规则的分类正确率, 可以看到方案 2 的各演化更新阶段所得规则具有较高的分类正确率, 说明演化更新所得规则具有较好的性能. 最后, 评估三种方案对时间序列变化趋势的预测性能, 如表 7 所示, 与其他两个方案相比, 方案 2 的模型预测误差最小. 对比方案 3 与方案 1, 本文提出的方案 2 具有更好的准确性.

这里, 分别以三种时间序列数据集的最后一次演化更新时滑动窗口中数据的 70 % 作为训练样本, 30 % 作为测试样本. 图 3 ∼ 5 给出了测试样本的三种方案的输出预测值与实际值的拟合曲线. 可以看到, 对于三种数据集, 方案 1 的拟合值与实际值相差最大, 方案 3 次之, 本文所提的方案 2 的拟合值几乎与原有时间序列重合, 具有更好的预测准确性.

表 6　不同数据集的有效性和准确性对比

Table 6　Comparison of the validity and accuracy of different data sets

数据集	演化次数	方案 1		方案 2	
		Rules	CA	Rules	CA
Air Quality	0	35	0.902	17	**0.911**
	1	49	0.897	18	**0.906**
	2	54	0.879	20	**0.916**
	3	62	0.887	21	**0.919**
	4	66	0.874	21	**0.924**
	5	71	0.867	24	**0.896**
	6	78	0.874	23	**0.909**
	7	87	0.894	20	**0.921**
	8	94	0.886	23	**0.914**
	9	99	0.879	25	**0.900**
	10	101	0.875	27	**0.894**
	11	109	0.874	24	**0.898**
	12	112	0.895	24	**0.927**
	13	119	0.9004	26	**0.932**
Istanbul	0	32	0.806	19	**0.855**
	1	36	0.814	17	**0.881**
	2	39	0.743	21	**0.805**
	3	51	0.807	21	**0.801**
	4	57	0.764	21	**0.805**
	5	67	0.794	25	**0.801**
	6	74	0.7778	23	**0.889**
Synthetic Control Chart	0	65	0.904	44	**0.946**
	1	78	0.882	49	**0.891**
	2	86	0.856	49	**0.888**
	3		990.862	56	**0.875**
	4	108	0.843	48	**0.851**
	5	110	0.856	53	**0.956**
	6	114	0.854	59	**0.896**
	7	124	0.889	48	**0.926**

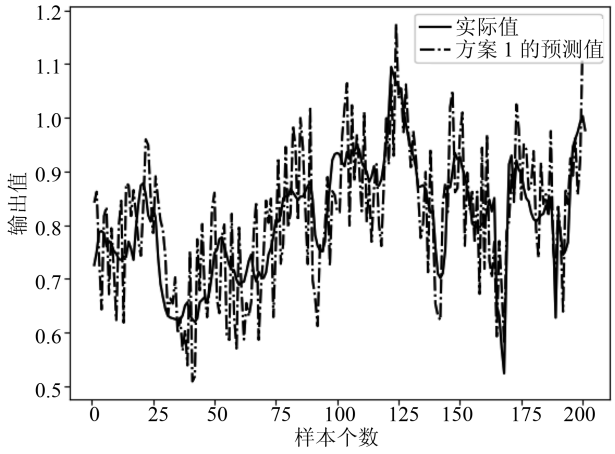
5　结论

为了处理时间序列数据, 时态关联规则挖掘广泛地用于各种实际应用. 通常时态数据库中的时间序列数据是动态变化的, 而现有的时态关联规则挖掘技术对于最近更新的数据进行挖掘时, 往往要依赖所有的时间序列数据进行重复挖掘, 造成了资源的浪费. 本文提出了一种新的算法, 不仅可以动态更新时态关联规则, 而且通过结合模糊推理可以实现时态关联规则的推理预测. 本文采用了滑动窗技术对时间序列分块处理, 通过概念漂移检测实现系统演化更新, 避免了重复性地对模型进行重建的工作.

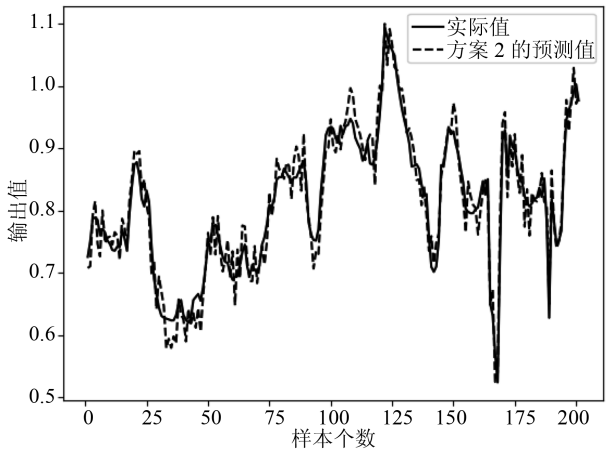
表 7 拟合误差
Table 7 Fitting error

数据集	演化次数	方案 1	方案 2	方案 3
	0	0.189	**0.137**	0.168
	1	0.195	**0.145**	0.172
	2	0.176	**0.121**	0.156
	3	0.169	**0.112**	0.144
	4	0.162	**0.104**	0.136
Air	5	0.201	**0.154**	0.159
Quality	6	0.194	**0.146**	0.174
	7	0.156	**0.108**	0.144
	8	0.186	**0.133**	0.176
	9	0.197	**0.148**	0.185
	10	0.211	**0.158**	0.195
	11	0.197	**0.155**	0.186
	12	0.154	**0.099**	0.129
	13	0.146	**0.091**	0.132
	0	0.144	**0.094**	0.129
	1	0.184	**0.139**	0.165
	2	0.226	**0.172**	0.208
Istanbul	3	0.198	**0.149**	0.184
	4	0.188	**0.139**	0.171
	5	0.209	**0.168**	0.196
	6	0.148	**0.103**	0.139
	0	0.132	**0.072**	0.095
	1	0.154	**0.094**	0.124
	2	0.169	**0.113**	0.146
Synthetic	3	0.187	**0.132**	0.167
Control	4	0.206	**0.149**	0.182
Chart	5	0.129	**0.076**	0.109
	6	0.138	**0.092**	0.116
	7	0.155	**0.087**	0.126

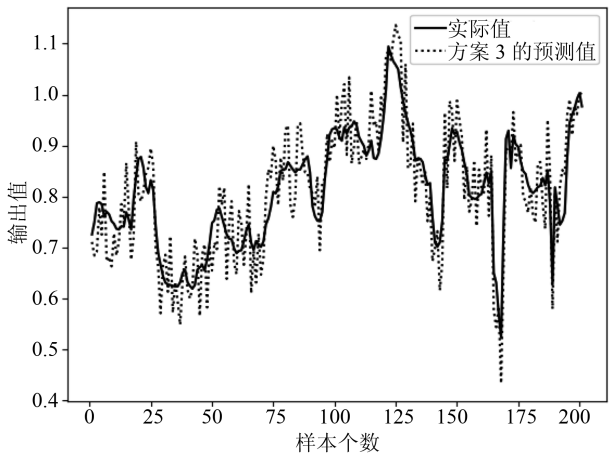
此外, 为了更有效地利用时态关联规则进行定量预测, 构建了基于多维时态关联规则的模糊推理预测方法, 通过实验对比研究, 表明了本文算法的有效性、可扩展性和准确性.



(a) 方案 1 拟合曲线对比
(a) Scheme 1 fitting curve comparison

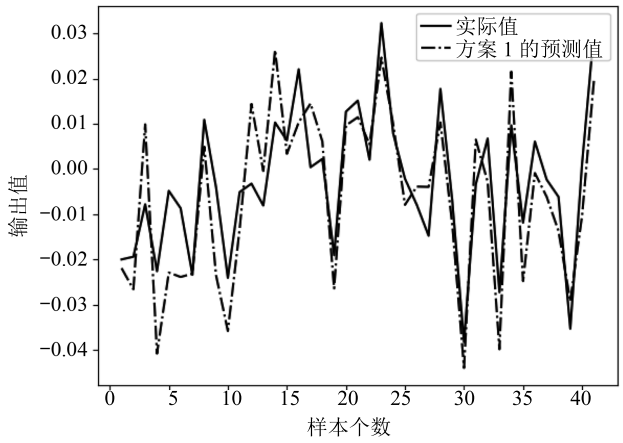


(b) 方案 2 拟合曲线对比
(b) Scheme 2 fitting curve comparison

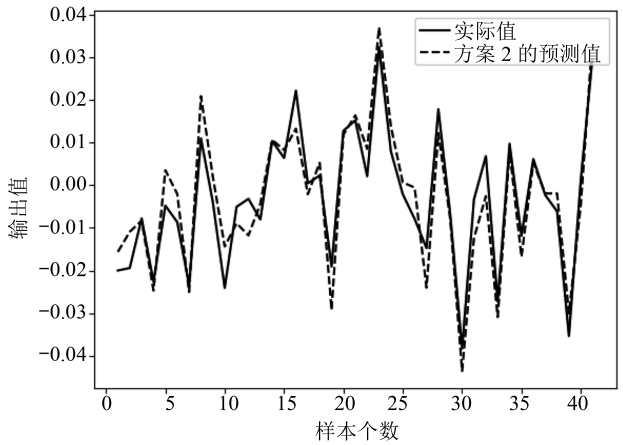


(c) 方案 3 拟合曲线对比
(c) Scheme 3 fitting curve comparison

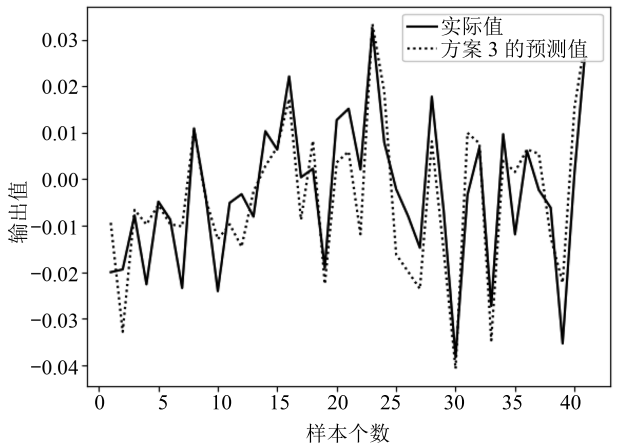
图 3 数据集 Air Quality 的拟合曲线
Fig. 3 The fitting curve of the data set Air Quality



(a) 方案 1 拟合曲线对比
(a) Scheme 1 fitting curve comparison

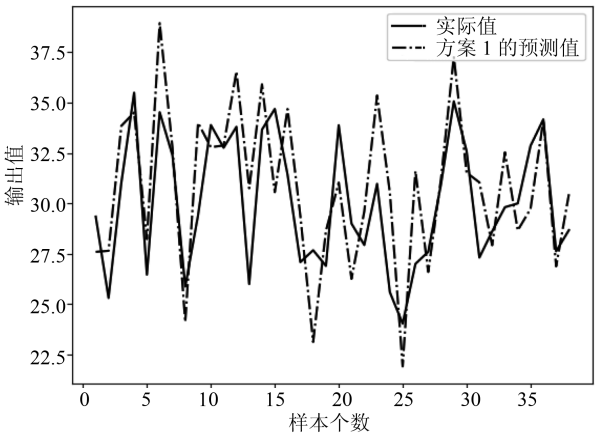


(b) 方案 2 拟合曲线对比
(b) Scheme 2 fitting curve comparison

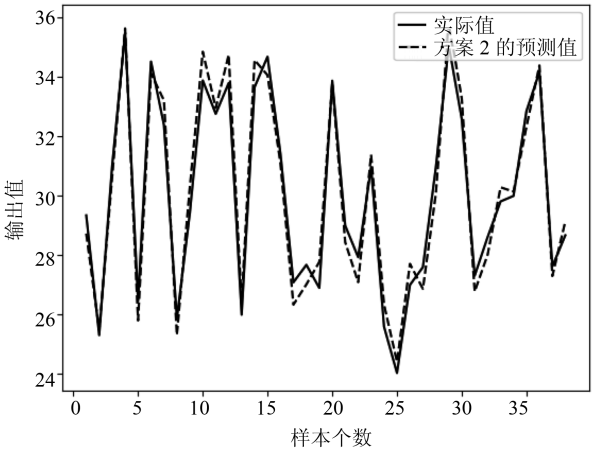


(c) 方案 3 拟合曲线对比
(c) Scheme 3 fitting curve comparison

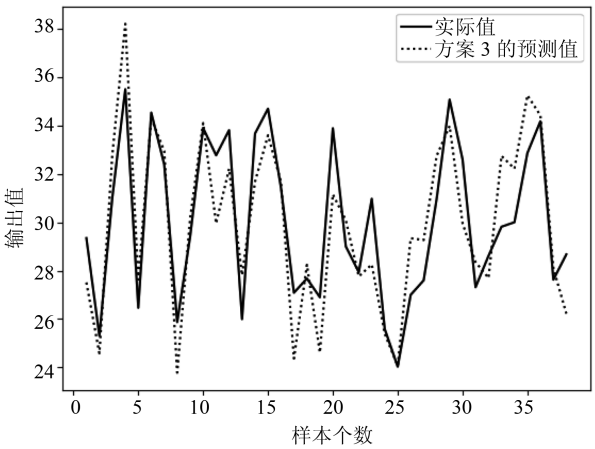
图 4　数据集 Istanbul 的拟合曲线
Fig. 4　The fitting curve of the data set Istanbul



(a) 方案 1 拟合曲线对比
(a) Scheme 1 fitting curve comparison



(b) 方案 2 拟合曲线对比
(b) Scheme 2 fitting curve comparison



(c) 方案 3 拟合曲线对比
(c) Scheme 3 fitting curve comparison

图 5　数据集 Synthetic Control Chart 的拟合曲线
Fig. 5　The fitting curve of the data set Synthetic Control Chart

References

1 Yolcu U, Aladag C H, Egrioglu E, Uslu V R. Time-series forecasting with a novel fuzzy time-series approach: an example for Istanbul stock market. *Journal of Statistical Computation and Simulation*, 2013, **83**(4): 599−612

2 Cheng C H, Yang J H. Rough-set rule induction to build fuzzy time series model in forecasting stock price. In: Proceedings of the 12th Conference on International Fuzzy Systems and Knowledge Discovery (FSKD). Zhangjiajie, China: IEEE, 2015. 278−284

3 Aslanargun A, Mammadov M, Yazici B, Yolacan S. Comparison of ARIMA, neural networks and hybrid models in time series: tourist arrival forecasting. *Journal of Statistical Computation and Simulation*, 2007, **77**(1): 29−53

4 Cheng C H, Shiu H Y. A novel GA-SVR time series model based on selected indicators method for forecasting stock price. In: Proceedings of the 2014 International Conference on Information Science, Electronics and Electrical Engineering (ISEEE). Sapporo, Japan: IEEE, 2014. 395−399

5 Zhao Hao, Wang Tao, Xu Fan, Fang Yan-Jun. Change tendency and forecast on time series in dynamic association rules mining. *Journal of Henan University of Science and Technology: Natural Science*, 2015, **36**(6): 40−45
(赵昊, 汪涛, 许凡, 方彦军. 时序动态关联规则挖掘中趋势变化与预测. 河南科技大学学报: 自然科学版, 2015, **36**(6): 40−45)

6 Zeng Y, Yin S Q, Liu J Y, Zhang M. Research of improved FP-Growth algorithm in association rules mining. *Scientific Programming*, 2015, **2015**: Article No. 910281

7 Xiao Y Y, Tian Y, Zhao Q H. Optimizing frequent time-window selection for association rules mining in a temporal database using a variable neighbourhood search. *Computers and Operations Research*, 2014, **52**: 241−250

8 Adhikari J, Rao P R. Identifying calendar-based periodic patterns. *Emerging Paradigms in Machine Learning*. Berlin, Heidelberg: Springer, 2013. 329−357

9 Ben Ahmed E, Nabli A, Gargouri F. On line mining of cyclic association rules from parallel dimension hierarchies. *Real World Data Mining Applications*. Cham: Springer International Publishing, 2015. 31−50

10 Matthews S G, Gongora M A, Hopgood A A, Ahmadi S. Web usage mining with evolutionary extraction of temporal fuzzy association rules. *Knowledge-Based Systems*, 2013, **54**: 66−72

11 Yang H D, Yang C C. Using health-consumer-contributed data to detect adverse drug reactions by association mining with temporal analysis. *ACM Transactions on Intelligent Systems and Technology*, 2015, **6**(4): Article No. 55

12 Nath B, Bhattacharyya D K, Ghosh A. Incremental association rule mining: a survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2013, **3**(3): 157−169

13 Zhuang D E H, Li G C L, Wong A K C. Discovery of temporal associations in multivariate time series. *IEEE Transactions on Knowledge and Data Engineering*, 2014, **26**(12): 2969−2982

14 Hong T P, Lan G C, Su J H, Wu P S, Wang S L. Discovery of temporal association rules with hierarchical granular framework. *Applied Computing and Informatics*, 2016, **12**(2): 134−141

15 Sirisha G N V G, Shashi M. A new multivariate time series transformation technique using closed interesting subspaces. In: Proceedings of the 2015 International Mining Intelligence and Knowledge Exploration. Cham: Springer, 2015. 392−405

16 Mohd K N, Mustapha A, Ahmad M H. Effect of temporal relationships in associative rule mining for web log data. *The Scientific World Journal*, 2014, **2014**: Article No. 813983

17 Wang Ling, Meng Jian-Yao, Xu Pei-Pei, Peng Kai-Xiang. Similarity dynamical clustering algorithm based on multidimensional shape features for time series. *Chinese Journal of Engineering*, 2017, **39**(7): 1114−1122
(王玲, 孟建瑶, 徐培培, 彭开香. 基于多维时间序列形态特征的相似性动态聚类算法. 工程科学学报, 2017, **39**(7): 1114−1122)

18 Pankaj G, Sagar B B. Discovering weighted calendar-based temporal relationship rules using frequent pattern tree. *Indian Journal of Science and Technology*, 2016, **9**(28). DOI: 10.17485/ijst/2016/v9i28/98455

19 Xu Zheng-Guang. Pattern recognition method of intelligent automation and its implementation in engineering [Ph. D. dissertation], College of Automation, University of Science and Technology, China, 2001.
(徐正光. 智能自动化的模式识别方法及其工程实现 [博士学位论文], 北京科技大学自动化学院, 中国, 2001.)

20 Sun C P, Gao Q, Yu H, Xu Z G. Study on moving pattern based faultdetection method. *Applied Mechanics and Materials*, 2013, **427−429**: 1463−1466

21 Xu Z G, Sun C P. Moving pattern-based forecasting model of a class of complex dynamical systems. In: Proceedings of the 50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC). Orlando, FL, USA: IEEE, 2011. 4967−4972

22 Xu Z G, Sun C P. Moving pattern-based approach to modeling of a class of complex production processes. In: Proceedings of the 2011 IEEE International Conference on Fuzzy Systems. Taipei, China: IEEE, 2011. 2282−2287

23 Xu Zheng-Guang, Sun Chang-Ping, Wu Jin-Xia. Moving pattern measured by interval number for modeling and control. *Control Theory and Application*, 2012, **29**(9): 1115−1124
(徐正光, 孙昌平, 吴金霞. 基于区间数度量的运动模式建模与控制. 控制理论与应用, 2012, **29**(9): 1115−1124)

24 Xu Zheng-Guang, Sun Chang-Ping. Moving pattern forecasting using interval T-S fuzzy model. *Control and Decision*, 2012, **27**(11): 1699−1705
(徐正光, 孙昌平. 基于区间 T-S 模糊模型的运动模式预测. 控制与决策, 2012, **27**(11): 1699−1705)

25 Ding Yuan, Wang Bin, Yan Jin-Chong, Pan Sheng. Prediction of burning through point based on two-dimensional interval autoregressive model. *Sintering and Pelletizing*, 2017, **42**(3): 1−6, 15
(丁园, 王斌, 鄢进冲, 潘昇. 基于二维区间自回归模型的烧结终点预测. 烧结球团, 2017, **42**(3): 1−6, 15)

26 Sun Chang-Ping, Xu Zheng-Guang. Multi-dimensional moving pattern prediction based on multi-dimensional interval T-S fuzzy model. *Control and Decision*, 2016, **31**(9): 1569−1576
(孙昌平, 徐正光. 基于多维区间 T-S 模糊模型的多维运动模式预测. 控制与决策, 2016, **31**(9): 1569−1576)

27 Dries A, Rückert U. Adaptive concept drift detection. *Statistical Analysis and Data Mining*, 2009, **2**(5−6): 311−327

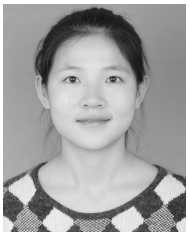
28 Buntine W. Learning classification trees. *Statistics and Computing*, 1992, **2**(2): 63−73

29 Aha D. UCI Machine learning repository: Center for machine learning and intelligent systems [Online], available: http://archive.ics.uci.edu/ml, August 25, 2017

30 Wang H B, Liu Y C, Wang C D. Research on association rule algorithm based on distributed and weighted FP-growth. *Advances in Multimedia, Software Engineering and Computing*. Berlin, Heidelberg: Springer, 2011, **1**: 133−138

31 Wang J S, Zhang Y, Sun S F. Multiple T-S fuzzy neural networks soft sensing modeling of flotation process based on fuzzy C-means clustering algorithm. *Advances in Neural Network Research and Applications*. Berlin, Heidelberg: Springer, 2010. 137−144

王 玲 北京科技大学自动化学院副教授. 主要研究方向为数据挖掘, 机器学习与复杂系统建模. 本文通信作者.
E-mail: lingwang@ustb.edu.cn
(**WANG Ling** Associate professor at the School of Automation and Electrical Engineering, University of Science and Technology Beijing. Her research interest covers data mining, machine learning, and complex system modeling. Corresponding author of this paper.)



孟建瑶 北京科技大学自动化学院硕士研究生. 主要研究方向为数据挖掘和机器学习. E-mail: 18810481455@163.com
(**MENG Jian-Yao** Master student at the School of Automation and Electrical Engineering, University of Science and Technology Beijing. Her research interest covers data mining and machine learning.)

李俊飞 北京科技大学自动化学院硕士研究生. 主要研究方向为数据挖掘和机器学习. E-mail: hpuljfei@163.com
(**LI Jun-Fei** Master student at the School of Automation and Electrical Engineering, University of Science and Technology Beijing. His research interest covers data mining and machine learning.)

彭开香 北京科技大学自动化学院教授. 主要研究方向为复杂工业系统的故障诊断与一体化控制.
E-mail: kaixiang@ustb.edu.cn
(**PENG Kai-Xiang** Professor at the School of Automation and Electrical Engineering, University of Science and Technology Beijing. His research interest covers fault diagnosis and integrated control for complex industrial system.)