计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(12):75−82 [doi: 10.15888/j.cnki.csa.006686]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

实时演进数据序列集的内在模式提取与行为预测①

艾锐峰[1], 欧阳军[1], 程　杰[2], 周　凯[2], 孙云鹏[2]

[1](解放军 63850 部队 总体所, 白城 137001)
[2](解放军 63850 部队 水文装备试验站, 烟台 264100)
通讯作者: 艾锐峰, E-mail: arfnavy@163.com

摘　要: 复杂系统数据序列集未来行为的预测是一个难点, 利用数据挖掘实现预测是有潜力的技术途径. 针对包含多元时间序列和非时间序列的实时演进数据集, 整合序列分割、聚类、模式在线匹配等处理流程, 提出了一种主题发现与联合决策相结合的预测方法. 在整个方法构建中, 将拟构造的主题发现式预测和联合决策预测融合进前期的序列分割与聚类中, 采用多时间粒度、多跨度对序列进行对应分层与分割, 聚合形成各层的标准模式集. 再以标准模式集, 依照预测策略, 反向搜索具有高稳定性延展行为的复合模式作为主题模式集, 从而实现基于在线模式匹配的行为预测. 最后, 采用分布式并行计算的架构实现整个处理算法. 理论推导和实验数据分析证明, 相比传统的时间序列预测方法准确度得到提高.

关键词: 多元时间序列; 时间粒度; 聚类; 主题发现; 融合预测

引用格式: 艾锐峰,欧阳军,程杰,周凯,孙云鹏.实时演进数据序列集的内在模式提取与行为预测.计算机系统应用,2018,27(12):75–82. http://www.c-s-a.org.cn/1003-3254/6686.html

Intrinsic Mode Extraction and Behavior Prediction for Real-Time Evolution Data Set

AI Rui-Feng[1], OUYANG Jun[1], CHENG Jie[2], ZHOU Kai[2], SUN Yun-Peng[2]

[1](General Planning Institute, No. 63850 Troops of PLA, Baicheng 137001, China)
[2](Hydrologic Equipment Testing Station, No. 63850 Troops of PLA, Yantai 264100, China)

Abstract: Prediction of future behavior of complex set of data sets is a difficult task. Data mining is a potential technical way. For the real-time evolutionary data sets containing multiple time series and non time sequence, a method of integrating the sequence segmentation, clustering, and pattern matching is proposed, which combines the theme discovery and joint decision. In the whole method construction, the topic discovery prediction and joint decision prediction are fused into the early sequence segmentation and clustering. The sequences are stratified and segmented for forming standard pattern sets of each layer, using multi time granularity and multi span. Then, according to the standard pattern set, with the prediction strategy, the compound pattern with high stability extension behavior is used as the theme pattern. This can predict with online pattern matching. Finally, a distributed parallel computing architecture is used to implement the whole processing algorithm. Theoretical deduction and experimental data analysis show that the accuracy of the method is improved compared with the traditional time series prediction method.

Key words: multivariate time series; time granularities; clustering; theme discovery; fusion prediction

借由数据集合内在模式的提取、内涵知识的挖掘　　形成有价值的信息, 以用于分析、评估、预测和控制,

① 基金项目: 原总装备部试验技术研究项目 (SYJS98170342)
Foundation item: Technological Research Project of Former General Equipment Department (SYJS98170342)
收稿时间: 2018-05-17; 修改时间: 2018-06-15; 采用时间: 2018-06-19; csa 在线出版时间: 2018-12-03

是目前大数据和人工智能领域的主要研究内容之一[1,2]. 当根据应用场景和数据特点对数据进行处理时, 若从时间纬度进行考察, 可分为: ① 无时序要求, 即数据本身是一种事物性逻辑关联关系而非时间序列, 或者数据为时间序列其处理应用无时序要求[3,4]; ② 严时序要求, 即数据本身为时间序列, 对其处理是利用历史时间序列分析实现对紧随其后的预测、控制等[5,6]; ③ 介于二者之间, 数据本身包含时间序列集和非时间序列集, 通过对其处理以用于未来[7,8]. 第三种数据集具有普遍性, 应用于各种领域[9,10]. 第一种数据集也可按照数据收集的时间戳构建成时序序列结构[11], 以用于描述所代表事物的演进过程.

通过实时演进的数据序列集的分析处理实现对事物未来行为的预测是数据分析的主要目的之一. 基于数据挖掘的行为预测, 从整个处理流程来看, 要实现从序列的建模、分割、相似性度量与搜索、聚类与分类、在线模式匹配, 到最终的预测决策. 目前的研究主要集中在序列的分割[12,13], 相似性搜索[14], 相似性度量[15], 序列的建模、聚类和分类[16–20]等方面, 侧重于单一方法的性能提升, 对融合整个流程以提升预测性能需要深入研究. 文献[21,22]介绍了一类模式挖掘方法, 主要用于从数据库中提取频繁出现的特定模式以找出数据的某种特性, 为静态分析, 对实时演化的时间序列集的行为预测缺乏论述. 文献[23]试图通过序列的模糊比对实现预测, 但参入比对的序列为现时子序列, 现时子序列如何延展到未来时刻没有分析. 文献 [24,25] 给出了多尺度融合的数据挖掘方法, 但对挖掘后的预测没有做进一步的研究. 文献[26]给出了对复杂系统数据挖掘分层建模的方法, 其所构建的模型对历史数据的拟合很好, 但是其预测效果并没有定量给出. 目前实用的时间序列预测方法为传统的 ARIMA 类方法, 但在非平稳条件及混沌情况下, 性能下降.

综上所述, 通过数据挖掘的方法可以对实时演进数据序列集在特定情况下的未来行为作预测, 但应当在模式提取时加入预测的考量. 若仅基于在数据中找相似点、聚类, 然后比对预测, 缺乏指向性. 再则, 要实现预测, 未来数据不可获取, 只有当下数据和历史数据, 而复杂事物的非平稳性、突变性, 使得当下子序列与模式的匹配, 并不能够说明未来的情形, 需要在序列分割、模式提取和在线匹配识别时向前延展. 鉴于此, 本文以实时演进数据集为对象, 通过融合处理, 提出了一种基于多时间粒度分层分割、模式提取、主题发现与联合决策的预测方法.

1 序列建模与模式提取

现实世界中所观测录取的数据是客观事物行为的记录和关联因子的描述. 构建数据序列集 $\Phi=\langle \boldsymbol{X},\boldsymbol{U},\boldsymbol{Y},\boldsymbol{V}\rangle(\boldsymbol{X}=\{\boldsymbol{x}\},\boldsymbol{U}=\{\boldsymbol{u}\},\boldsymbol{Y}=\{\boldsymbol{y}\},\boldsymbol{V}=\{\boldsymbol{v}\})$ 以刻画随时间而不断向前演进的客观事物 R. R 的主要行为由多元时间序列 $\boldsymbol{x}=[x_i(t)](i=1,2,\cdots,m_1)$和$m_2$个非时间序列 $\boldsymbol{u}=[u_i](i=1,2,\cdots,m_2)$记录, 关联的影响因素由多元时间序列 $\boldsymbol{y}=[y_i(t)](i=1,2,\cdots,m_3)$和$m_4$个非时间序列 $\boldsymbol{v}=[y_i](i=1,2,\cdots,m_4)$描述. 以 R 在 t 时刻之前的数据集$\Phi(t)$的分析实现对 R 在$t+\Delta t$时刻的行为预测即为要解决的问题.

R 受到各种因素的作用, 其数据随机性、确定性并存, 如金融经济数据、海洋气象数据、战场数据等. 可以认为 R 受到宏观基本规律的约束、当下现实因素的作用、微观层次的扰动以及外部稀疏的偶然性冲击. 根据以上推论, R 在某一时刻的最终行为可以认为是由以上四方面共同作用决定, 则如果由数据序列集 $\Phi=\langle \boldsymbol{X},\boldsymbol{U},\boldsymbol{Y},\boldsymbol{V}\rangle$导出表征以上四个方面的数据序列集: $\boldsymbol{A}$, 表征宏观规律; $\boldsymbol{B}$, 当下作用; $\boldsymbol{C}$, 微观层面; $\boldsymbol{E}$, 外部冲击, 则借由$\Psi=\langle \boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{E}\rangle$上的内在模式提取, 再进行融合预测, 将更符合事物逻辑, 有望提高预测的准确度. 由Φ导出Ψ可根据多时间粒度的概念[3,17], 借由多时间粒度的分层与分割实现.

1.1 基于时间粒度的序列分层与分割

以多元时间序列 $\boldsymbol{x}=[x_i(t)](i=1,2,\cdots,m_1)$为例. 若 $x_i(t)$可获得不同时间采样间隔的序列$x_i(nT_1),x_i(nT_2),\cdots,x_i(nT_Z)$, 则以待预测的时间粒度为中间层 B, 将$x_i(t)$分成 A、B、C 三层:

$$\begin{cases} x_i(nT_A)=\{x_i(nT_z)\}(T_z>T_B) \\ x_i(nT_B)=\{x_i(nT_z)\}(T_z\approx T_B) \\ x_i(nT_C)=\{x_i(nT_z)\}(T_z<T_B) \end{cases} \tag{1}$$

若记录数据只有一种固定采样率的序列$x_i(nT_0)$, 采用平均的方式, 将$x_i(nT_0)$整合出三层序列$x_i(nT_A)$、$x_i(nT_B)$、$x_i(nT_C)$, 记为 $\boldsymbol{A}$、$\boldsymbol{B}$、$\boldsymbol{C}$. 对 $\boldsymbol{Y}$ 的操作按照与 $\boldsymbol{X}$ 对齐的方式同步处理.

从序列$x_i(nT_A)$、$x_i(nT_B)$、$x_i(nT_C)$(简记为$x_i(n)$) 中提取模式, 需要对其进行分割. 对序列$x_i(n)$的分割即是将$x_i(n)$按照等时间长度或者变时间长度划分为一族子

序列, 通过子序列的聚类分析提取内在模式.

设$\boldsymbol{x}=[x_i(n)]$为m维长度为N的多元时间序列, 虚拟一个维度m长度W的窗. 令$W=\xi W_0$, W_0根据应用场景给定, ξ为调整系数. 跨度L为窗W向前滑动截取的步长, $T_z\leqslant L\leqslant W$. 窗$W$自$\boldsymbol{x}$的起点, 滑动到尾点, 截取一系列子序列$\boldsymbol{s}_k$, 得到子序列集合$\boldsymbol{S}=(\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_K)$. 当$L=T_z$时, 则一步一截取, 前后子序列有重叠部分, 计算量较大; 当$L=W$时, $\boldsymbol{S}$成为$\boldsymbol{x}$的一个首尾相衔接的子序列分割, 截取效率高, 但当出现跨子序列的模式时, 可能遗漏. 针对具体应用, 合理选取L值 (或者根据子序列聚类分析结果与L值的对照关系, 通过试验比较, 确定L值). 具体算法如下:

算法 1. 序列分割算法

1) 从集合 $\boldsymbol{X}$ 中输入待分割序列样本$\boldsymbol{x}$, 指定初值W_0, 调整系数$\xi\in[0.5,1.5]$, $i=0$, $j=1$.
2) 令$\xi=0.5+0.1i$, $W=\xi W_0$.
3) 根据W, 由$T_z\leqslant L\leqslant W$, 给定跨度值$L\in[L_1,L_2,\cdots,L_J]$.
4) 令$L=L_j$, 由$\boldsymbol{x}$起始位置向前截取长度为W的子序列, 赋给$\boldsymbol{s}_1$.
5) 滑动截取窗向前步进 L, 截取$\boldsymbol{s}_2$, 循环操作直到序列尾点, 得到一个截取集$\boldsymbol{S}_{i,j}=(\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_K)$.
6) 令$j=j+1$, 返回第 4) 步, 直到 L 遍历$[L_1,L_2,\cdots,L_J]$.
7) 令$i=i+1$, 返回第 2) 步, 直到ξ遍历[0.5,1.5].
8) 合并截取集$\boldsymbol{S}_{i,j}$为最终集合$\boldsymbol{S}$, $\boldsymbol{S}$即为序列$\boldsymbol{x}$分割后的全体子序列集.
9) 返回第 1) 步, 输入下一个待分割序列样本.

$\boldsymbol{S}$为$\boldsymbol{x}$的一个分割, 由不等长的一系列子序列$\boldsymbol{s}_k$组成, 代表了在时间粒度T_z上、在一定时间区间内, 序列可能呈现出的各种表现形式. 通过$\boldsymbol{S}$的聚类分析, 提取其中蕴含的内在模式, 可用于对$\boldsymbol{x}$未来时刻行为的预测.

以海洋数据集为例, 不同海区的水温序列总集可看做$\boldsymbol{X}$, 特定海区的水温序列可以看做$\boldsymbol{x}$, 则既可以进行总体特征分析也可以进行特定区域特征分析.

1.2 模式提取

序列集合$\boldsymbol{S}=(\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_K)$是$\boldsymbol{x}$的子序列集, 假定存在$\boldsymbol{x}$的内含模式集$\boldsymbol{\Gamma}=(\boldsymbol{\Gamma}_1,\boldsymbol{\Gamma}_2,\cdots,\boldsymbol{\Gamma}_P)$, 则$\forall\boldsymbol{s}_k\in\boldsymbol{S}$, $\exists\boldsymbol{\Gamma}_p\in\boldsymbol{\Gamma}$, 使得:

$$\boldsymbol{s}_k=\boldsymbol{\Gamma}_p+\varepsilon\quad(p\in[1,P]) \tag{2}$$

其中, ε是子序列$\boldsymbol{s}_k$与它所分属模式$\boldsymbol{\Gamma}_p$之间的差异, $\boldsymbol{s}_k$与$\boldsymbol{\Gamma}_p$越相似, ε越小. 对于度量相似性的处理方法, 有闵可夫斯基距离法、动态时间弯曲距离法 (Dynamic Time Warping, DTW)[15,27]、扩展 Frobenius 范数法 (Extended Frobenius Norm, Eros) 等. 闵式距离简单直观, 其特例欧式距离是常用的距离计算方法, 但它对波动、噪声非常敏感, 且需要序列等长. Eros 不满足距离三角不等式, 对于本文后续预测处理不适用, 因而下面采取 DTW 进行相似性度量. DTW 通过时间序列弯曲部分的自我复制, 实现序列相似波形的对齐匹配, 不要求序列等长.

设$\boldsymbol{s}_i=(\boldsymbol{s}_{i,1},\boldsymbol{s}_{i,2},\cdots,\boldsymbol{s}_{i,N_i})$, $\boldsymbol{s}_j=(\boldsymbol{s}_{j,1},\boldsymbol{s}_{j,2},\cdots,\boldsymbol{s}_{j,N_j})$是维度为$m$, 时间点长度分别为$N_i$、$N_j$的两个多元子序列, 其 DTW 距离[15]:

$$d(\boldsymbol{s}_i,\boldsymbol{s}_j)=d_0(\boldsymbol{s}_{i,1},\boldsymbol{s}_{j,1})+\min\begin{cases}d(\boldsymbol{s}_i,\boldsymbol{s}_j[2:N_j])\\ d(\boldsymbol{s}_i[2:N_i],\boldsymbol{s}_j)\\ d(\boldsymbol{s}_i[2:N_i],\boldsymbol{s}_j[2:N_j])\end{cases} \tag{3}$$

其中, $d_0(\boldsymbol{s}_{i,1},\boldsymbol{s}_{j,1})$为$\boldsymbol{s}_{i,1}$, $\boldsymbol{s}_{j,1}$的基距离, 用欧式距离计算.

对分割得到的子序列集合$\boldsymbol{S}$根据 DTW 距离进行相似性度量, 利用 K-mean 法进行聚类. 设在 A, B, C 层上分别聚合为P_A、P_B、P_C簇, 以各簇质心所对应的子序列及各簇内复现频数靠前的若干子序列作为标准模式, 得到$\boldsymbol{\Gamma}_A=(\boldsymbol{\Gamma}_A^1,\boldsymbol{\Gamma}_A^2,\cdots,\boldsymbol{\Gamma}_A^{P_A})$, $\boldsymbol{\Gamma}_B=(\boldsymbol{\Gamma}_B^1,\boldsymbol{\Gamma}_B^2,\cdots,\boldsymbol{\Gamma}_B^{P_B})$, $\boldsymbol{\Gamma}_C=(\boldsymbol{\Gamma}_C^1,\boldsymbol{\Gamma}_C^2,\cdots,\boldsymbol{\Gamma}_C^{P_C})$.

对$\boldsymbol{Y}$的操作按照与$\boldsymbol{X}$对齐的方式同步处理. 于是由 R 的原始数据序列集$\Phi=\langle\boldsymbol{X},\boldsymbol{U},\boldsymbol{Y},\boldsymbol{V}\rangle$, 经过前述处理, 得到$\boldsymbol{X},\boldsymbol{Y}\to\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}\to\boldsymbol{\Gamma}_A,\boldsymbol{\Gamma}_B,\boldsymbol{\Gamma}_C$. 非时间序列集$\boldsymbol{U}$、$\boldsymbol{V}$根据时间戳对应归类为孤立事件集$\boldsymbol{E}=(\boldsymbol{e}_U,\boldsymbol{e}_V)$. 于是完成了$\Phi\to\Psi\to\Omega_0$, $\Psi=\langle\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{E}\rangle$为$\Phi$经多时间粒度分层整合后的数据序列集, $\Omega_0=\langle\boldsymbol{\Gamma}_A,\boldsymbol{\Gamma}_B,\boldsymbol{\Gamma}_C,\boldsymbol{E}\rangle$为$\Psi$经过时间分割、相似度量、聚类分析后的内在模式表征集.

2 主题发现与预测策略

根据 R 记录数据所提取的内在模式表征集$\Omega_0=\langle\boldsymbol{\Gamma}_A,\boldsymbol{\Gamma}_B,\boldsymbol{\Gamma}_C,\boldsymbol{E}\rangle$, 对它未来行为进行预测, 可以有多种策略, 应进行融合处理.

2.1 主题发现

对于 R 而言, 可知的是t_0及$t\leqslant t_0$前的数据序列集. 在t_0时刻附近的表现受到宏观、中观、微观层及外部冲击的影响, 呈现的序列对很多事物而言不一定具有连续性和稳定性, 但是其呈现的模式具有近似意义上的可复现性. 当特定模式出现时, R 的后续行为表现出相对稳定性. 即总体上不一定可以准确预测, 但是当序列开始呈现这种特定的模式时, 利用这种稳定的表现, 向未来进行延展, 即可以用于此时刻 R 未来行为的预测. 这些特定模式定义为主题模式. $\Omega_0=\langle\boldsymbol{\Gamma}_A,\boldsymbol{\Gamma}_B,\boldsymbol{\Gamma}_C,\boldsymbol{E}\rangle$

囊括了 R 的行为特征. 主题模式集$\boldsymbol{M}=(\boldsymbol{m}_m)(m=1,2,\cdots,M)$可以由提取的标准模式集合$\boldsymbol{\Gamma}_A$, $\boldsymbol{\Gamma}_B$, $\boldsymbol{\Gamma}_C$, $\boldsymbol{E}$, 中的模式组合得到, 如公式 (4) 所示.

$$\boldsymbol{m}_m=<\boldsymbol{\Gamma}_A^{p_A},\boldsymbol{\Gamma}_B^{p_B},\boldsymbol{\Gamma}_C^{pC},\boldsymbol{e}_U^{p_U},\boldsymbol{e}_V^{pV}> \tag{4}$$

其中, $p_A\in[1,P_A]$, $p_B\in[1,P_B]$, $p_C\in[1,P_C]$, $p_U\in[1,P_U]$, $p_V\in[1,P_V]$. 组合方式可以基于专家经验或者对全集合进行遍历. 具体如下:

第一步: 对于$\boldsymbol{m}_m$, 基于相似性度量, 从一定时长L的历史数据序列中进行匹配, 统计其出现频数$f(\boldsymbol{m}_m)$.

第二步: 根据 R 预测要求, 以二元决策 (H_0, H_1) 为例 (天气预报的下雨、不下雨, 证券价格的涨跌等; 对于非二元决策, 可以进行预测区间离散化处理, 形成一个多元决策问题, 处理方式一致), 统计当出现$\boldsymbol{m}_m$时 R 后续行为为 H_0、H_1 的出现频数$f'_0(\boldsymbol{m}_m)$、$f'_1(\boldsymbol{m}_m)$.

第三步: 计算决策H_0、H_1的正确率$\eta(\mathrm{H}_0/\boldsymbol{m}_m)$、$\eta(\mathrm{H}_1/\boldsymbol{m}_m)$, 如公式 (5) 所示:

$$\begin{cases}\eta(\mathrm{H}_0/\boldsymbol{m}_m)=\dfrac{f'_0(\boldsymbol{m}_m)}{f(\boldsymbol{m}_m)}\times100\%\\ \eta(\mathrm{H}_1/\boldsymbol{m}_m)=\dfrac{f'_1(\boldsymbol{m}_m)}{f(\boldsymbol{m}_m)}\times100\%\end{cases} \tag{5}$$

设定正确率门限$\delta(\delta\in(0.5,1.0])$, 对于$\eta(\mathrm{H}_0/\boldsymbol{m}_m)\geqslant\delta$的$\boldsymbol{m}_m$归于$\mathrm{H}_0$主题模式$\boldsymbol{M}_{\mathrm{H}0}$, $\eta(\mathrm{H}_1/\boldsymbol{m}_m)\geqslant\delta$的$\boldsymbol{m}_m$归于$\mathrm{H}_1$主题模式$\boldsymbol{M}_{\mathrm{H}_1}$. 再根据出现频数$f(\boldsymbol{m}_m)$对$\boldsymbol{M}_{\mathrm{H}_0}$、$\boldsymbol{M}_{\mathrm{H}_1}$中$\boldsymbol{m}_m$由高到低排序, 设定频数门限$\omega$, 剔除$f(\boldsymbol{m}_m)<\omega$的低频度模式. 至此, 得到可资利用的主题模式$\boldsymbol{M}_{\mathrm{H}_0}$、$\boldsymbol{M}_{\mathrm{H}_1}$.

2.2 预测策略

对于实时演进的系列集 R 而言, 现时刻为t_0, 则可获得的即为$t\leqslant t_0$之前的数据和相关联的孤立事件$\boldsymbol{u},\boldsymbol{v}$. 需要以其为基础对$t_0+\Delta t$时的行为进行预测. 在$t_0$时刻, 以 1.1 节的时间粒度处理方法, 实时在线截取 A 层的待匹配子序列$\boldsymbol{x}_A(nT_A)$, 记为$\boldsymbol{x}_A(n)$, 同样处理得到$\boldsymbol{x}_B(n)$,$\boldsymbol{x}_C(n)$. $\boldsymbol{x}_A(n)$,$\boldsymbol{x}_B(n)$,$\boldsymbol{x}_C(n)$的时间点数分别为N_A,N_B,N_C, 其值取对应层标准序列长度的平均值 (为了描述简单, 假设 A, B, C 层都只有一种时间粒度, 实际处理中可以在每一层尝试多种时间粒度). 在t_0时刻附近, R 的行为由$\boldsymbol{R}_{t_0}=<\boldsymbol{x}_A(n),\boldsymbol{x}_B(n),\boldsymbol{x}_C(n),\boldsymbol{u},\boldsymbol{v}>$表示.

根据前述处理$\boldsymbol{\Gamma}_A=(\boldsymbol{\Gamma}_A^1,\boldsymbol{\Gamma}_A^2,\cdots,\boldsymbol{\Gamma}_A^{P_A})$是$\boldsymbol{S}_A=(\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_K)$聚合后的标准模式. 待匹配子序列$\boldsymbol{x}_A(n)$, 需要计算它与$\boldsymbol{\Gamma}_A$中何种标准模式最为接近, 抽取此模式用做预测. 给定距离度量门限d_0^A, 以$\boldsymbol{\Gamma}_A^{p_A}$为中心, d_0^A为半径, 将$\boldsymbol{S}_A$中成员分为P_A个不重叠的簇. 当$d(\boldsymbol{\Gamma}_A^{p_A},\boldsymbol{s}_i)\leqslant d_0^A$时, $\boldsymbol{s}_i$属于p_A簇, 圆外的$\boldsymbol{s}_i$全部剔除, 剔除后的余集记为$\boldsymbol{S}'_A$, 同理得到$\boldsymbol{S}'_B$、$\boldsymbol{S}'_C$. 如图 1 所示.

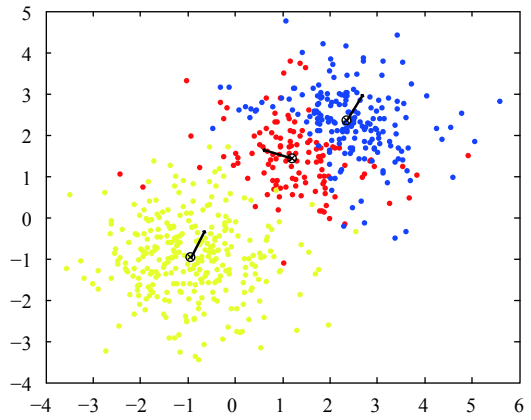

图 1　余集获取示意图

以余集$\boldsymbol{S}'_A$、$\boldsymbol{S}'_B$、$\boldsymbol{S}'_C$为基础, 构建下面两种预测方法.

(1) 主题发现预测

先不考虑孤立事件影响, 根据所挖掘的主题模式$\boldsymbol{M}_{\mathrm{H}0}=<\boldsymbol{\Gamma}_{A0}^{p_A},\boldsymbol{\Gamma}_{B0}^{p_B},\boldsymbol{\Gamma}_{C0}^{pC}>$, $\boldsymbol{M}_{\mathrm{H}1}=<\boldsymbol{\Gamma}_{A1}^{p_A},\boldsymbol{\Gamma}_{B1}^{p_B},\boldsymbol{\Gamma}_{C1}^{pC}>$, 对$t_0$时刻的复合序列$\boldsymbol{x}_t=<\boldsymbol{x}_A(n),\boldsymbol{x}_B(n),\boldsymbol{x}_C(n)>$进行匹配. 匹配过程以 DTW 距离进行度量, 采取 K 最邻近法 (k-Nearest Neighbor, KNN) 对$\boldsymbol{x}_A(n)$、$\boldsymbol{x}_B(n)$、$\boldsymbol{x}_C(n)$在余集$\boldsymbol{S}'_A$、$\boldsymbol{S}'_B$、$\boldsymbol{S}'_C$中进行分类处理. 若经过分类处理$\boldsymbol{x}_A,\boldsymbol{x}_B,\boldsymbol{x}_C$同时匹配上$\boldsymbol{\Gamma}_{A0}^{p_A}$、$\boldsymbol{\Gamma}_{B0}^{p_B}$、$\boldsymbol{\Gamma}_{C0}^{pC}$, 则认为复合序列$\boldsymbol{x}_t$匹配上$\boldsymbol{M}_{\mathrm{H}0}$中的某特定主题模式$\boldsymbol{m}_{\mathrm{H}0}$, 抽取$\boldsymbol{m}_{\mathrm{H}0}$利用其后续延展的行为进行$t_0+\Delta t$时$\boldsymbol{x}_t$的预测; 若$\boldsymbol{x}_A,\boldsymbol{x}_B,\boldsymbol{x}_C$同时匹配上$\boldsymbol{\Gamma}_{A1}^{p_A}$、$\boldsymbol{\Gamma}_{B1}^{p_B}$、$\boldsymbol{\Gamma}_{C1}^{pC}$, 则认为$\boldsymbol{x}_t$匹配上$\boldsymbol{M}_{\mathrm{H}1}$中的某特定主题模式$\boldsymbol{m}_{\mathrm{H}1}$, 抽取$\boldsymbol{m}_{\mathrm{H}1}$进行预测; 匹配不上则放弃$t_0$时刻的预测, 序列向前推进, 进行下一个处理.

分析 2.1 节利用历史数据挖掘主题模式的过程, 及这种在线匹配、主题发现预测的策略, 可知其为低频度模式. 鉴于此, 制定主题发现预测方法的补充策略, 即联合决策预测.

(2) 联合决策预测

联合决策预测策略为对$\boldsymbol{x}_A(n)$、$\boldsymbol{x}_B(n)$、$\boldsymbol{x}_C(n)$在余集$\boldsymbol{S}'_A$、$\boldsymbol{S}'_B$、$\boldsymbol{S}'_C$中分别匹配, 只要在各层匹配上标准模式, 抽取标准模式进行联合推断. 方法如下:

第一步: 根据 DTW 距离度量, 根据 KNN 法对$\boldsymbol{x}_A(n)$在余集$\boldsymbol{S}'_A$中进行分类处理. 设定参数$\rho\in[70\%,90\%]$,

截取$\boldsymbol{x}_A$的后ρ部分与$\boldsymbol{s}_i$的前ρ部分, 分别记为$\boldsymbol{x}_{A,\rho}$、$\boldsymbol{s}_{i,\rho}$, 计算 DTW 距离$d(\boldsymbol{x}_{A,\rho},\boldsymbol{s}_{i,\rho})$. 若经过分类处理$\boldsymbol{x}_A$属于$\boldsymbol{\Gamma}_A^{p_A}$簇, 则将$\boldsymbol{\Gamma}_A^{p_A}$簇标准模式$\boldsymbol{\Gamma}_A^{p_A}$赋给$\boldsymbol{x}_A$, 并认为$\boldsymbol{\Gamma}_A^{p_A}$的后$1-\rho$序列即为$\boldsymbol{x}_A$向前延展的预测值. 记此预测为$D_A$. 对$\boldsymbol{x}_B$、$\boldsymbol{x}_C$进行同样处理, 得到$D_B$、$D_C$.

第二步: 以二元决策$(\mathrm{H}_0,\mathrm{H}_1)$为例, 制定规则 : 只有当$D_A$、$D_B$、$D_C$同时指示$\boldsymbol{x}(t_0+\Delta t)$的行为为$\mathrm{H}_0$时, 推断为$\mathrm{H}_0$, 同理处理$\mathrm{H}_1$(也可以根据宏观、中观、微观的先验知识, 对$D_A$、$D_B$、$D_C$进行加权处理, 本文采取“同时指示”这种强准则). 则由D_A、D_B、D_C进行联合预测的正确概率如公式 (6) 所示:

$$P_f(\mathrm{H}_0/A,B,C)=\frac{P_{f1}}{P_{f1}+P_{f2}} \tag{6}$$

其中, $P_{f1}=P_f(\mathrm{H}_0)P_f(A)P_f(B)P_f(C)$, $P_{f2}=(1-P_f(\mathrm{H}_0))(1-P_f(A))(1-P_f(B))(1-P_f(C))$, $P_f(\mathrm{H}_0)$为H_0出现的先验概率, $P_f(A)$、$P_f(B)$、$P_f(C)$为根据D_A、D_B、D_C决策的正确概率 (在此假设D_A、D_B、D_C决策相互独立, 若完全相关则退化为单层模式), 与模式复现的稳定性相关.

实际的预测要求是在本层 (B 层) 时间粒度上对$\boldsymbol{x}(t_0+\Delta t)$的行为作出判断. 由公式 (6) 推导可得到:

$$\begin{cases}\forall P_f(\mathrm{H}_0)\geq 0.5, P_f(\mathrm{A})+P_f(\mathrm{C})\geq 1\\ \exists P_f(\mathrm{H}_0/\mathrm{A},\mathrm{B},\mathrm{C})\geq P_f(\mathrm{B})\end{cases} \tag{7}$$

考察公式 (7), 假设序列总体上呈现随机漫步, 毫无偏向, 则$P_f(\mathrm{H}_0)=0.5$, 此时可以认为模式识别无意义, $P_f(A)=0.5$、$P_f(C)=0.5$; 若序列具有偏向, 则或者$P_f(\mathrm{H}_0)>0.5$或者$P_f(\mathrm{H}_1)>0.5$, 考虑到D_A、D_C是根据提取的模式进行匹配识别作出的二元判断, 其准确度应$P_f(A)\geqslant 0.5$、$P_f(C)\geqslant 0.5$. 综上所述, $P_f(\mathrm{H}_0/A,B,C)\geqslant P_f(B)$的条件可以认为满足, 即在退化条件下, “同时指示”这种强准则下的联合决策正确率也至少等于基于本层的决策$P_f(B)$, 若序列展现偏向性, 则联合决策的正确率将会提升.

综合两种预测策略, 设计下面的整体预测方案:

算法 2. 整体预测方案

1) 经历史数据处理得到标准模式库$\Omega_0=\langle\boldsymbol{\Gamma}_A,\boldsymbol{\Gamma}_B,\boldsymbol{\Gamma}_C,\boldsymbol{E}\rangle$, 经主题挖掘得到主题模式集合$\boldsymbol{M}_{\mathrm{H}_0}=(\boldsymbol{m}_{\mathrm{H}_0})$、$\boldsymbol{M}_{\mathrm{H}_1}=(\boldsymbol{m}_{\mathrm{H}_1})$.

2) 获取聚类分析后的余集$\boldsymbol{S}'_A$、$\boldsymbol{S}'_B$、$\boldsymbol{S}'_C$.

3) 以当前时刻t_0为基准向后截取并整合出待匹配子序列$\boldsymbol{x}_A,\boldsymbol{x}_B,\boldsymbol{x}_C$.

4) 采用 DTW 距离度量, 根据 KNN 法对$\boldsymbol{x}_A,\boldsymbol{x}_B,\boldsymbol{x}_C$在余集$\boldsymbol{S}'_A$、$\boldsymbol{S}'_B$、$\boldsymbol{S}'_C$中进行分类、匹配. 若匹配不上, t_0不做预测, 转入 3), 等待序列向前演进后的下一时刻的处理. 若匹配上, 即$\boldsymbol{x}_A\rightarrow\boldsymbol{\Gamma}_A^{p_A}$、$\boldsymbol{x}_B\rightarrow\boldsymbol{\Gamma}_B^{p_B}$、$\boldsymbol{x}_C\rightarrow\boldsymbol{\Gamma}_C^{p_C}$, 转入 5).

5) 若$\langle\boldsymbol{\Gamma}_A^{p_A},\boldsymbol{\Gamma}_B^{p_B},\boldsymbol{\Gamma}_C^{p_C}\rangle\in\boldsymbol{M}_{\mathrm{H}_0}$则以主题模式$\boldsymbol{m}_{\mathrm{H}_0}=\langle\boldsymbol{\Gamma}_A^{p_A},\boldsymbol{\Gamma}_B^{p_B},\boldsymbol{\Gamma}_C^{p_C}\rangle$的行为进行$\mathrm{R}(t_0+\Delta t)$预测; 若$\langle\boldsymbol{\Gamma}_A^{p_A},\boldsymbol{\Gamma}_B^{p_B},\boldsymbol{\Gamma}_C^{p_C}\rangle\in\boldsymbol{M}_{\mathrm{H}_1}$则以主题模式$\boldsymbol{m}_{\mathrm{H}_1}=\langle\boldsymbol{\Gamma}_A^{p_A},\boldsymbol{\Gamma}_B^{p_B},\boldsymbol{\Gamma}_C^{p_C}\rangle$的行为进行$\mathrm{R}(t_0+\Delta t)$预测; 若$\langle\boldsymbol{\Gamma}_A^{p_A},\boldsymbol{\Gamma}_B^{p_B},\boldsymbol{\Gamma}_C^{p_C}\rangle\notin\{\boldsymbol{M}_{\mathrm{H}_0},\boldsymbol{M}_{\mathrm{H}_1}\}$, 则转入 6).

6) 联合决策预测: 设定匹配百分比参数ρ, 转入 4) 重新计算; 当$\boldsymbol{x}_A\rightarrow\boldsymbol{\Gamma}_A^{p_A}$、$\boldsymbol{x}_B\rightarrow\boldsymbol{\Gamma}_B^{p_B}$、$\boldsymbol{x}_C\rightarrow\boldsymbol{\Gamma}_C^{p_C}$, 获取延展预测值$D_A$、$D_B$、$D_C$, 若$D_A$、$D_B$、$D_C$同时指示$\mathrm{R}(t_0+\Delta t)$的行为为$\mathrm{H}_0$时, 推断为$\mathrm{H}_0$; 同理处理$\mathrm{H}_1$否则转入 2), 等待序列向前演进后的下一时刻的处理.

注: 若存在孤立事件的冲击, 则通过历史对照的方法, 加入模式匹配中.

3 系统实现与实例分析

基于上述模型构建与算法设计, 在计算机系统上予以实现并选取实例进行效果分析.

3.1 系统实现

前述模型与算法中, 变跨度滑窗子序列截取、DTW 距离计算、相似性搜索、K-mean 法聚类分析和 KNN 分类等, 计算量都较大, 为了更好的从历史数据序列中提取模式, 需尽可能的采用较长的时间序列, 从而造成计算量急剧上升. 在线匹配预测, 其计算量要小于模式提取的过程. 鉴于此, 采用分布式并行处理架构, 如图 2 所示.

整个系统由 A、B 两个子系统组成. A 系统采取外部云计算托管; B 系统在线监控实时处理, 由 N 个并行计算节点组成. 软件设计采用 Python 语言粘合 MPI 并行编程环境的方式. 以 Python 编制数据端口, 将数据导入分发, 一份为模式提取全时长数据库, 一份为在线数据片集. 将模式提取并行计算程序布置在外部云系统上, 在全时长数据库上进行提取操作, 维持一个标准模式库并进行主题模式的挖掘, 所得到的模式库发往 B 系统. 在本地并行计算机系统上布置在线匹配预测的并行计算程序, 将模式库与在线数据片集结合, 根据数据序列的驱动, 实时更新处理. 累积一定时间, 在 A 系统上重新启动模式提取处理, 监测 R 是否会演化, 出现新的标准模式或者主题模式则更新模式库, 并发往 B 系统.

3.2 实例分析

本文目的是构建一种通用的处理架构, 主要面向气象海洋数据、战场数据以及经济金融数据. 出于数据获取便利性的考虑, 下面以石油期货相关数据为例

进行算法验证.

试验数据: NYMEX 原油期货主力合约数据 (2002.1.1 至 2016.1.1, 取其年月周日分的价格序列的开盘价、收盘价、最高价、最低价、相关的宏观经济数据以及关联国际事件), 2002.1.1–2012.1.1 为模式提取数据区间, 2012.1.2–2016.1.1 为模拟预测处理数据区间.

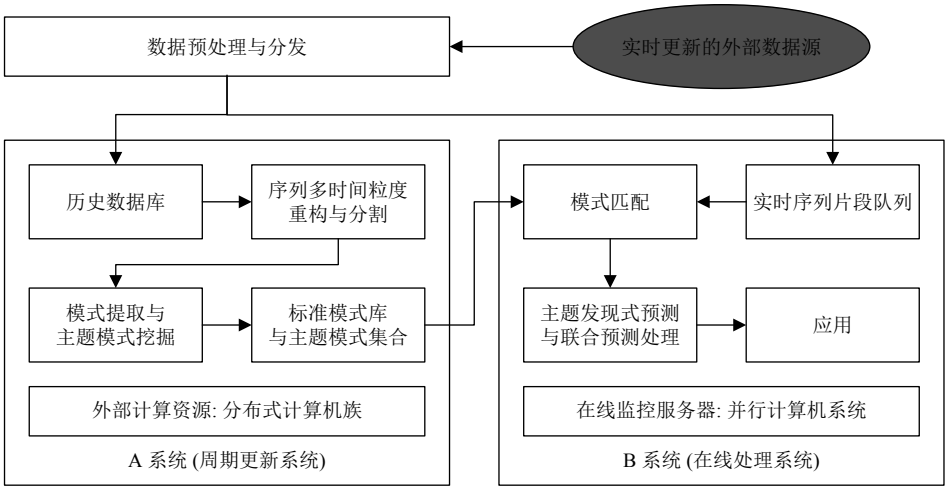

图 2 系统框图

经处理, A 层时间粒度取为 $6T$(T 代表一日)、B 层为 T、C 层为 2 个小时, 以 B 层日预测为目标, 预测日线上 $T+n$ 日价格行为 (本文取 T+2 日的预测). 匹配百分比$\rho = 0.8$. K-mean 聚类时, 设置 A、B、C 层初始分类数目均为 6. 根据程序数据结果, A 层额外抽取频繁子序列 2 个, B 层 4 个, C 层 1 个. 最终提取结果为: A 层标准模式数$P_A = 8$, B 层标准模式数$P_B = 10$、C 层标准模式数$P_C = 7$.

主题模式挖掘中组合模式$<\boldsymbol{\Gamma}_A^{p_A},\boldsymbol{\Gamma}_B^{p_B},\boldsymbol{\Gamma}_C^{p_C}>$在历史数据集中 (总长 10 年) 的出现频数 (从高到低) 如图 3 所示.

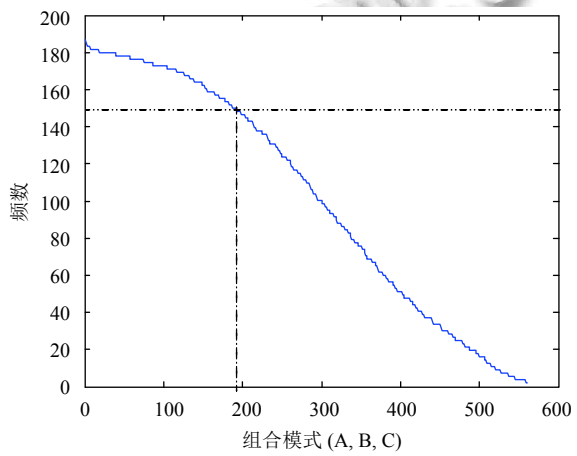

图 3 组合模式在历史数据集中的出现频数

图 4 为, 组合模式$<\boldsymbol{\Gamma}_A^{p_A},\boldsymbol{\Gamma}_B^{p_B},\boldsymbol{\Gamma}_C^{p_C}>$对 T+2 日预测的正确概率统计: H_0 表示预测 T+2 日涨, H_1 表示预测 T+2 日跌.

根据图 3、图 4 的统计, 取出现频数较高、对 H0 预测正确概率较高的组合模式作为主题模式 $\boldsymbol{M}_{H_0} = (\boldsymbol{m}_{H_0})$; 对 H_1 预测正确概率较高的组合模式作为主题模式$\boldsymbol{M}_{H_1} = (\boldsymbol{m}_{H_1})$.

根据上述结果, 定义: F1, 主题发现式预测; F2, 联合决策式预测; F3, 传统的基于日线的 ARIMA 预测; F4, 日线子序列模式匹配预测; F5, 分层小波分解预测, 对 5 种方法的预测性能进行比较. 其中, F5 为将日线通过小波变换, 分解为表示宏观的和表示细节的部分, 在每层上分别用 ARIMA 递推, 再相加的方式进行. 模拟预测数据区间为 2012.1.2-2016.1.1 共计 4 年.

F1: 抽取$\boldsymbol{M}_{H_0}$中 H_0 决策正确率最高, 且在 2002.1.1–2012.1.1 中出现频数排序在前 30% 的主题模式进行匹配预测; 同理抽取$\boldsymbol{M}_{H_1}$中主题模式. 在 F1 预测做出时, 同步记录 F3、F4、F5 的预测结果. 其决策的统计结果如表 1 所示.

由表 1 可发现, F1 相比其他方法有较高的正确率, 说明经过主题模式挖掘, 某些特定的复合模式出现时, 其后续的行为十分稳定, 用之预测有较高的准确率. 但是其中预测最准确且复现频率排序前 30% 的主题模式出现的频率也只有年平均 18.3 次, 十分稀疏.

F2: 通过在线分层匹配处理, 当均匹配上时, 启动决策. 若“同时指示”H_0 或 H_1, 则取此指示为决策, 统计正误; 否则放弃. 同步记录 F3、F4、F5 的预测结果. 其决策的统计结果如表 2 所示.

根据表 2 结果, F2 也比 F3、F4、F5 准确率要高, 但是其复现频率也不高, 年平均出现 49.5 次, 且联合决策的放弃数也较高. 将 F1、F2 结合, 按照前述算法 2 的流程进行在线监测, 可以提高可预测的频数.

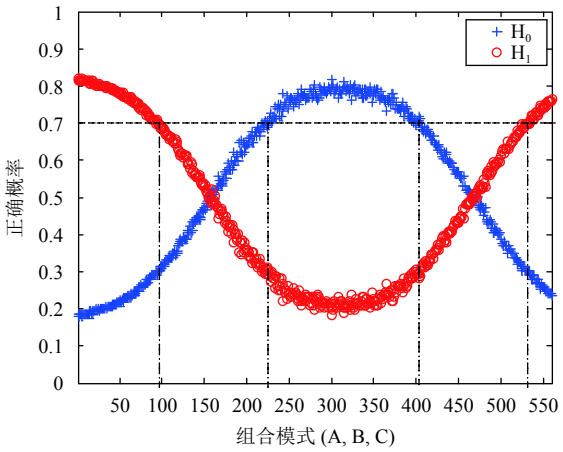

图 4 组合模式在历史数据集中预测的正确率统计

表 1 F1 方法与其他方法预测性能比较

决策		方法			
		F1	F3	F4	F5
H_0	决策总频数	73	73	73	73
	正确数	61	45	49	48
	错误数	12	28	24	25
	正确率 (%)	83.6	61.6	67.1	65.8
H_1	决策总频数	59	59	59	59
	正确数	43	29	34	36
	错误数	16	30	25	23
	正确率 (%)	72.9	49.2	57.6	61.0

表 2 F2 方法与其他方法预测性能比较

决策	方法			
	F2	F3	F4	F5
决策总频数	237	237	237	237
正确数	94	135	155	151
错误数	35	102	82	86
放弃数	108	0	0	0
正确率 (%)	72.9	57.0	65.4	63.7

综合言之, 准确性的提高得益于特定模式的挖掘和联合判断, 但这种处理同时注定了在线处理时, 只能等待序列在实时演进过程中呈现出此模式近似态时才可进行决策. 考察实际应用场景, 这种新的预测方法具有意义 (如在投资决策中, 当机会出现时再投入显然比贸然介入更有利; 在海洋水文参数与作战场景呈现某种特定态势时, 作出未来态势推断并付诸行动比较适宜), 而常规的时时刻刻做未来预测的准确性值得警惕.

4 结束语

复杂事物行为的数据序列集, 变化复杂、序列前后时刻存在逻辑上的不确定性、且概率分布未知、具有混沌突变性. 在实时演进过程中, 其平稳运行与突然变化相互杂交, 无法实时推断下一时刻会发生什么. 但是某些模式或会反复出现. 当在监测过程中, 这些特殊形态显现大部的时候, 其后续有较大概率按照此模式运行. 文中根据事物影响因子的宏微观特性, 将序列集通过多时间粒度和跨度的分层分割, 提取代表各层特性的标准模式集, 再挖掘具有稳定延展表现的主体模式, 构建出主题模式在线匹配和联合决策的预测方法. 此方法与传统的几种序列预测方法相比, 具有较高的预测准确性, 但是在线复现率不高. 如果对算法中的部分门限和参数进行放宽处理, 则可以提高频数, 但是预测准确性可能降低. 准确率、复现率与门限和参数的对应关系、折中处理等, 需要结合具体应用场景作进一步研究.

参考文献

1 Lynch C, Goldston D, Howe D, *et al*. Big data: Science in the petabyte era. Nature, 2008, 455(7209): 1–136. [doi: 10.1038/455001a]

2 Los W, Wood J. Dealing with data: Upgrading Infrastructure. Science, 2011, 331(6024): 1515–1516.

3 梁吉业, 钱宇华, 李德玉, 等. 大数据挖掘的粒计算理论与方法. 中国科学: 信息科学, 2015, 45(11): 1355–1369.

4 Aghabozorgi S, Shirkhorshidi AS, Wah TY. Time-series clustering—a decade review. Information Systems, 2015, 53: 16–38. [doi: 10.1016/j.is.2015.04.007]

5 陈宁, 薛禹胜, 丁杰, 等. 风速时间序列的符号化描述. 电力系统自动化, 2017, 41(11): 33–38. [doi: 10.7500/AEPS20170109003]

6 李建林, 籍天明, 孔令达, 等. 光伏发电数据挖掘中的跨度选取. 电工技术学报, 2015, 30(14): 450–456. [doi: 10.3969/j.issn.1000-6753.2015.14.060]

7 Nunes SA, Romani LAS, Avila AMH, *et al*. Finding spatio-temporal Patterns in multidimensional data streams. Journal

of Information and Data Management, 2013, 4(3): 327–340.
8 李德仁, 张良培, 夏桂松. 遥感大数据自动分析与数据挖掘. 测绘学报, 2014, 43(12): 1211–1216.
9 李清泉. 从 Geomatics 到 Urban Informatics. 武汉大学学报•信息科学版, 2017, 42(1): 1–6.
10 Chang XM, Chen BY, Li QQ, *et al*. Estimating real-time traffic carbon dioxide emissions based on intelligent transportation system technologies. IEEE Transactions on Intelligent Transportation System, 2013, 14(1): 469–479. [doi: 10.1109/TITS.2012.2219529]
11 牟乃夏, 张恒才, 陈洁, 等. 轨迹数据挖掘城市应用研究综述. 地球信息科学学报, 2015, 17(10): 1136–1142.
12 倪志伟, 王超, 胡汤磊, 等. 面向数据流的多粒度时变分形维数计算. 软件学报, 2015, 26(10): 2614–2630.
13 李爱国, 覃征. 在线分割时间序列数据. 软件学报, 2004, 15(11): 1671–1679.
14 李正欣, 张凤鸣, 张晓丰, 等. 多元时间序列相似性搜索研究综述. 控制与决策, 2017, 32(4): 577–583.
15 李正欣, 郭建胜, 毛红保, 等. 多元时间序列相似性度量方法. 控制与决策, 2017, 32(2): 368–372.
16 Huo YK, Wang T, Maunder RG, *et al*. Motion-aware mesh-structured trellis for correlation modelling aided distributed multi-view video coding. IEEE Transactions on Image Processing,2014,23(1):319–331.[doi:10.1109/TIP.2013.2288913]
17 徐健锋, 张远健, Zhou DN, 等. 基于粒计算的不确定性时间序列建模及其聚类. 南京大学学报 (自然科学), 2014, 50(1): 86–94.
18 McMahon C, Soe B, Loeb A, *et al*. Boundary identification in EBSD data with a generalization of fast multiscale clustering. Ultramicroscopy, 2013, 133: 16–25. [doi: 10.1016/j.ultramic.2013.04.009]
19 Soheily-Khah S, Douzal-Chouakria A, Gaussier E. Generalized K-means-based clustering for temporal data under weighted and kernel time warp. Pattern Recognition Letters, 2016, 75: 63–69. [doi: 10.1016/j.patrec.2016.03.007]
20 原继东, 王志海, 孙艳歌, 等. 面向复杂时间序列的 K 近邻分类器. 软件学报, 2017, 28(11): 3002–3017.
21 方刚, 吴跃. 基于复合粒度计算的频繁模式挖掘研究. 计算机应用研究, 2016, 33(6): 1620–1624. [doi: 10.3969/j.issn.1001-3695.2016.06.005]
22 dos Alex JA, Gosselin PH, Philipp-Foliguet S, *et al*. Interactive multiscale classification of high-resolution remote sensing images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2013, 6(4): 2020–2034. [doi: 10.1109/JSTARS.2012.2237013]
23 徐信喆. 基于模糊 K 线序列比对的股市技术分析模型. 计算机应用与软件, 2010, 27(9): 28–32, 48. [doi: 10.3969/j.issn.1000-386X.2010.09.011]
24 柳萌萌, 赵书良, 韩玉辉, 等. 多尺度数据挖掘方法. 软件学报, 2016, 27(12): 3030–3050.
25 舒平达, 陈华辉. 支持多时间粒度的数据流上最频繁 K 项挖掘. 宁波大学学报 (理工版), 2009, 22(4): 500–505. [doi: 10.3969/j.issn.1001-5132.2009.04.011]
26 康卓, 黄竞伟, 李艳, 等. 复杂系统数据挖掘的多尺度混合算法. 软件学报, 2003, 14(7): 1229–1237.
27 Bankó Z, Abonyi J. Correlation based dynamic time warping of multivariate time series. Expert Systems with Applications, 2012, 39(17): 12814–12823. [doi: 10.1016/j.eswa.2012.05.012]