计算机科学
COMPUTER SCIENCE

多元时序上状态转移模式的三支漂移检测

沈少朋，马洪江，张智恒，周相兵，朱春满，温佐承

引用本文

沈少朋，马洪江，张智恒，周相兵，朱春满，温佐承. 多元时序上状态转移模式的三支漂移检测[J]. 计算机科学, 2022, 49(4): 144-151. SHEN Shao-peng, MA Hong-jiang, ZHANG Zhi-heng, ZHOU Xiang-bing, ZHU Chun-man, WEN Zuo-cheng. Three-way Drift Detection for State Transition Pattern on Multivariate Time Series[J]. Computer Science, 2022, 49(4): 144-151.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210600045

多元时序上状态转移模式的三支漂移检测

沈少朋　马洪江　张智恒　周相兵　朱春满　温佐承
成都信息工程大学软件工程学院　成都 610225
(ssp8471@163.com)

摘　要　多元时序数据上的无监督模式漂移检测是机器学习领域的一个研究热点。然而,对模式及其漂移现象的定义十分灵活,使得该任务的难度较高。受"三分而治"思想启发,文中提出了一种基于FUP-STAP增量挖掘的、针对带通配符区间的状态转移模式的三支漂移检测算法(Three-Way Drift Detection Method for State Transition pAttern with Periodic Wildcard Gaps,3WDD-STAP),它由状态转移模式(STAP)的增量算法改进而来。在不使用额外参数的情况下,3WDD-STAP可同时获得频繁的以及发生漂移的STAP。根据增量前后的支持度变化情况,模式漂移被定义为3类:I类漂移表示本来频繁的STAP在增量后变得不频繁,需扫描增量数据集;II类漂移表示本来不频繁的STAP在增量后变得频繁,需扫描原始数据集;III类漂移表示STAP在增量后维持了频繁或者不频繁,视为正常,不扫描数据集。在空气质量与石油工程设备监控两个真实数据上的实验结果表明:1)α和β的值越大,两类漂移模式的数量越少,反之亦然;2)I类漂移的STAP在不同数据集上服从不同分布;3)所得STAP模式及其漂移现象均有很强的可读性。

关键词:多元时序;漂移监测;三分而治;序列模式发现;增量学习
中图法分类号　TP391

Three-way Drift Detection for State Transition Pattern on Multivariate Time Series

SHEN Shao-peng,MA Hong-jiang,ZHANG Zhi-heng,ZHOU Xiang-bing,ZHU Chun-man and WEN Zuo-cheng
School of Software Engineering,Chengdu University of Information Technology,Chengdu 610225,China

Abstract　Unsupervised drift detection for multivariate time series (MTSs) is an important task in machine learning. However, this issue is challenging because the definitions of sequential patterns and their drifts are very flexible. Inspired by the idea of "Think in Threes", this paper proposes a three-way drift detection method for state transition pattern with periodic wildcard gaps (3WDD-STAP), which is improved from the incremental mining algorithm of STAP. Without additional parameters, both frequent and drifted STAPs can be obtained simultaneously. Considering the support changes around the increments, we define three types of STAP drift. Type I drift indicates that STAPs change from frequent to infrequent. The incremental dataset needs to be rescanned. Type II drift indicates that STAPs change from infrequent to frequent. The original dataset needs to be rescanned. Type III drift indicates that STAPs retain frequent or infrequent, namely, these STAPs are normal. No dataset needs to be rescanned. Finally, experimental results on 2 real-world datasets show that: 1) we obtain less drifted STAPs with less α and β, and vice versa; 2) the two types of drifted STAPs obeys different distribution for various datasets; 3) the obtained STAPs and their drifts have strong readability.

Keywords　Multivariate time series, Anomaly detection, Think in Threes, Sequential pattern discovery, Incremental learning

1　引言

起源于粗糙集理论[1]的三支决策[2]已在多个研究和应用领域取得长足发展[3]。受到"Think in Threes"思想的启发,新理论如三支粒计算[4]、三支形式概念分析[5]、三支模糊集[6]、三支区间集[7]以及三支代价敏感学习[8-9]被相继提出。此外,诸多实际应用如三支推荐系统[10-11]、三支主动学习[12]、三支分类[13]、三支聚类[14]、三支模式发现[15]、三支目标

到稿日期:2021-06-04　返修日期:2021-09-24
基金项目:国家自然科学基金(41604114,62006200);教育部产学研协同育人项目(201902298010);四川省科技计划项目(2020YFG0307);成都市重点研发支撑计划(2021-YF05-00933-SN);四川旅游学院科研项目(2020SCTU14,19SCTUZY03)
This work was supported by the National Natural Science Foundation of China(41604114,62006200),Ministry of Education Industry-University-Research Collaborative Education Project(201902298010),Sichuan Science and Technology Department Project(2020YFG0307),Chengdu Key R&D Support Plan(2021-YF05-00933-SN) and Sichuan Tourism University Scientific Research Project(2020SCTU14,19SCTUZY03).
通信作者:张智恒(zhihengzhang406@163.com)

识别[16]、三支属性约简[17]以及三支邮件过滤[18]均被广泛研究。

多元时序上的模式有两种主流定义，分别是 Along-first[19]和 Cross-first[20]。由于 Cross-first 类型的模式有更好的可读性和可扩展性，近年来受到了较多的关注[21]。对于以上两种类型的序列模式，均有 3 种序列模式匹配规则，由一般到特殊依次是：一般性匹配[22]、无重叠区间匹配[23]和一次性匹配[24]。

模式漂移的检测任务广泛存在于网络安全[25]、设备控制[26]、交通出行[27]、医疗诊断[28]、文本分析[29-30]等领域。针对是否知晓具体漂移类型，漂移检测还可进一步分为无监督[31]、半监督[32]以及有监督[33]等类型。其中，因缺失标签信息或专家经验，无监督方法对模式漂移的定义多种多样。定义新的模式及其漂移检测技术是该领域的关键问题和重要挑战[34]。

本文提出了一种增量式的多元序列模式三支漂移检测算法。该算法在获得漂移序列的同时，还能获得正常的频繁序列，且不会增加任何额外的参数。本文的研究主要分为 4 个部分。

(1)漂移类型定义。根据序列模式在增量前后是否频繁将发生模式划分为 3 个类别。I 类：增量前频繁，而增量后不频繁。II 类：增量前不频繁，而增量后频繁。III 类：保持了频繁或者不频繁。

(2)构造增量补充数据，简称增补数据。与传统频繁项集增量挖掘问题不同，可能有新的模式匹配次数出现在新旧数据的连接部位。为了准确计算模式的支持度，需要将原始数据的后缀补充进增量数据，即将原始数据末尾的记录拷贝到增量数据的头部。虽然增补数据的长度与模式的长度和通配符区间约束有关，但对于所有长度一致的模式，仅需要构造一次增补数据。

(3)定义三支模式漂移检测策略。根据序列模式在原始数据集和增补数据上是否频繁，可将模式的漂移检测策略分为 4 种情况：1)在原始数据集和增补数据集上都频繁；2)在原始数据集上频繁，在增补数据集上不频繁；3)在原始数据集上不频繁，在增补数据集上频繁；4)在原始数据集和增补数据集上都不频繁。其中，情况 1)和情况 4)属于 III 类，而 I 类只可能出现在情况 2)中，II 类只可能出现在情况 3)中。至于是否真的发生了漂移，还需要将其支持度与事先指定的频繁阈值进行比较。当处于情况 2)中的模式的支持度小于阈值时，才会被认定为 I 类漂移。类似地，当处情况 3)中的模式的支持度大于阈值时，才会被认定为 II 类漂移。总之，其核心思想是将序列模式频繁性质在增量前后的变化作为漂移检测的标准。即保持了频繁或不频繁均认为没有发生漂移，否则就判定发生了漂移。

(4)算法性能分析。获得 I 类漂移的模式需要反复扫描增量数据。由于增量数据通常远小于原始数据，因此实际效率较高。获得 II 类漂移的模式需要反复扫描原始数据。原始数据的规模较大，因此该类漂移的模式越多，算法实际运行的时间越长。在最坏的情况下，算法的复杂度是指数级的。

本文的主要贡献包括以下 3 个方面。

(1)以高可读性的 Cross-first 状态转移模式增量挖掘算法为基础，在三支决策理论的指导下，定义了两类新的模式漂移现象，并设计、实现了检测算法。

(2)相比传统的功能单一的漂移检测算法，本文工作可以在获得频繁序列模式的同时检测出漂移模式，而不增加任何额外的参数。

(3)本文工作不需要依次对每个单变量时序都做出分布假设，避免了对监督信息的依赖。

本文第 2 节综述了模式漂移检测的相关工作；第 3 节从数据模型出发，依次定义了本文必需的形式化概念，以及问题的形式化描述；第 4 节提出了模式漂移检测算法的设计与实现；第 5 节在多个领域的多元时序上验证了算法性能和漂移检测结果的可解释性；最后总结全文，并对下一步研究工作进行展望。

2 相关工作

图 1 给出了现有主流的无监督模式漂移检测算法。它们均由两个阶段完成：阶段一，从已有数据中提取表征“正常”的数据轮廓/特征；阶段二，将新加入的数据和“正常”特征进行比对，差别越明显，新数据漂移的程度就越大。

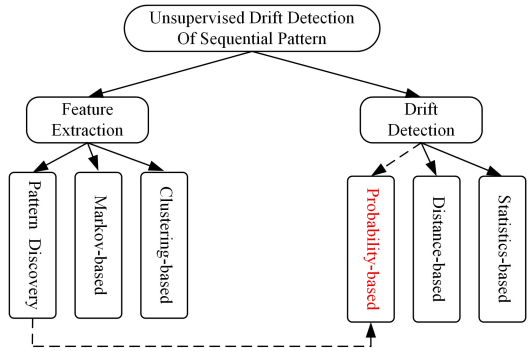


图 1 无监督模式漂移检测

Fig. 1 Unsupervised drift detection of sequential pattern

表 1 列出了第一阶段经常采用的 3 种技术及其优缺点。

表 1 特征提取技术的优缺点

Table 1 Strengths and weaknesses of feature extractions

技术	优点	缺点
模式发现[35]	尽可能少的超参数，免费的训练集	耗时，主观的频率阈值
马尔可夫[36]	精确度高	训练集较大，时间复杂度高
聚类[31]	维数缩减，检测速度非常快	距离难以设定，聚类非常耗时

在第一阶段中，人们提出了基于模式发现(Pattern Discovery)[35]、马尔可夫模型(Markov)[36]和聚类(Clustering)[31]的特征提取技术。将频繁序列作为数据特征具有两大明显优势：1)只需要考虑序列的频率，不需要额外引入辅助因子；2)不需要海量训练数据。因此，它受到了人们的广泛关注。

表 2 列出了第二阶段常用的技术及其优缺点。在第二阶段中，人们提出了基于距离(Distance)[37]、概率(Probability)[38]和统计模型(Statistics)[39]的漂移检测技术。

表 2 漂移检测技术的优缺点

Table 2 Strengths and weaknesses of drift detection

技术	优点	缺点
距离[37]	差异量化，顺序关系	时间和空间复杂度高，序列结构复杂
概率[38]	可信度高	耗时，经验概率阈值
统计模型[39]	外部事件敏感，精度高	高度依赖模型，经验性的阈值

总之，已有工作的核心思想是“与正常的数据分布特征有足够的差异”。

3 基本定义与问题描述

本节首先给出了状态转移模式及其漂移的形式化描述；其次将模式漂移程度的量化方式定义为支持度变化的绝对值；最后给出了模式漂移检测问题的形式化描述。

3.1 模式与漂移

定义 1(多元时序) 给定一个四元组

$$S=(T,A,V=\bigcup_{a\in A}V_a,f) \tag{1}$$

其中，$T=\{t_1,t_2,\cdots,t_m\}$ 表示时间点的有限集合；$A=\{a_1,a_2,\cdots,a_n\}$ 表示属性的有限集合；V_a 表示属性 $a\in A$ 的值域；$f:T\times A\to V$ 表示从关系 (t,a) 到某个元状态 $v_a\in V_a$ 的映射，即 $f(t,a)=v_a$。当 $|A|>1$ 时，S 被称作一个多元时序，否则被称作单变量时序。本文主要研究前者，若遇数值型数据，则将其离散化。

定义 2(状态转移模式) 给定一个符号型多元时序 $S=(T,A,V=\bigcup_{a\in A}V_a,f)$ 以及通配符区间 $[\underline{G},\overline{G}]$，模式 P 如下：

$$P=s_1[\underline{G},\overline{G}]s_2\cdots[\underline{G},\overline{G}]s_k \tag{2}$$

其被称为一个长度为 k 的状态转移模式，当且仅当：

(1) $\forall i\in[1,k],s_i\in 2^{\{(a,v_a)|a\in A,v_a\in V_a\}}-\varnothing$；

(2) $0\leqslant\underline{G}\leqslant\overline{G}\leqslant m$；

(3) $1\leqslant k\leqslant m$。

同一时刻的属性只能有一个取值。状态中包含的属性个数越多，则越特殊，反之则越一般。为了描述的简洁，当给定时刻 t_i 后，系统此时最特殊的状态被记为 $F(t_i)=\{(a,f(t_i,a))|a\in A\}$。给定一个状态 s，若它在时刻 t_i 发生，则可记为 $s\subseteq F(t_i)$。因此，状态 s 的支持度为：

$$sup(s)=\frac{\sum_{i=1}^{m}\text{index}(s\subseteq F(t_i))}{m} \tag{3}$$

其中，指示函数 index(·)表示其中条件为真时等于 1，否则等于 0。因此，给定一个阈值 α，若满足 $sup(s)\geqslant\alpha$，则称 s 是一个频繁状态。因此，所有频繁状态的集合可记为 $\mathbb{S}=\bigcup_{k=1}^{n}k\text{-}\mathbb{S}=\{s|sup(s)\geqslant\alpha\}$。特别地，令 $k\text{-}\mathbb{S}$ 表示所有大小为 k 的频繁状态集。

定义 3(状态转移模式匹配规则) 给定一个位置序列 $L=l_1,l_2,\cdots,l_k$，称模式 P 与序列 L 相匹配，当且仅当：

(1) $\forall i\in[1,k],l_{i+1}-l_i+1\in[\underline{G},\overline{G}]$；

(2) $\forall i\in[1,k],s_i\subseteq F(t_{l_i})$，即状态 s_i 在时刻 t_{l_i} 发生，简记为匹配函数 match$(P,L)=1$。

在长度为 k 的位置序列 L 中，给定任意的 $i\in[1,k-1]$，每个 l_i 后都可能有 $\overline{G}-\underline{G}+1$ 种 l_{i+1}。因此，总共有 $(\overline{G}-\underline{G}+1)^{k-1}$ 种位置序列。给定符号型多元时序 S 后，总共有 m 个时间点。因此，模式 $P=s_1[\underline{G},\overline{G}]s_2\cdots[\underline{G},\overline{G}]s_k$ 所有可能匹配的位置序列 L 的集合 $\mathbb{L}$ 及基数为：

$$|\mathbb{L}|=m(\overline{G}-\underline{G}+1)^{k-1} \tag{4}$$

进而，我们将真正匹配模式 P 的位置序列记为 $\mathscr{L}^P$，且它的基数为：

$$|\mathscr{L}^P|=\sum_{L\in\mathbb{L}}\text{match}(P,L) \tag{5}$$

因此，模式 P 的支持度可以记为：

$$sup(P)=\frac{|\mathscr{L}^P|}{|\mathbb{L}|} \tag{6}$$

当模式的长度 $k=1$ 时，式(6)则分化成式(3)。类似地，给定一个阈值 β，若满足 $\text{sup}(P)\geqslant\beta$，则称 P 是一个频繁状态转移模式。为了方便起见，所有长度为 k 的频繁模式所构成的集合为 $k\text{-}\mathbb{P}=\{P|sup(P)\geqslant\beta\}$，因此 $\mathbb{P}=\bigcup_{k=1}^{m}k\text{-}\mathbb{P}$ 表示所有频繁模式的集合。

新数据集被插入且使数据分布发生改变时，系统的行为模式会产生漂移。给定原始数据集 S 和增量数据集 ΔS 以及状态转移模式 P。若模式 P 在 S 上频繁，但在 $S^*=S\cup\Delta S$ 上不频繁，则认为发生了 I 类漂移现象。即：本来是频繁的，但在新数据上就不频繁了(支持度变得足够小)。若模式 P 在 S 上不频繁，但在 S^* 上频繁，则认为发生了 II 类漂移现象。即：本来是不频繁的，但在新数据上就频繁了(支持度变得足够大)。因此，可将状态转移模式漂移类型的形式化描述为：

$$\begin{cases}P\in\mathscr{D}_{\text{I}}, & \text{若 } sup(P,S)\geqslant\beta,sup(P,S^*)<\beta\\ P\in\mathscr{D}_{\text{II}}, & \text{若 } sup(P,S)<\beta,sup(P,S^*)\geqslant\beta\\ P, & \text{正常}\end{cases} \tag{7}$$

其中，$\mathscr{D}_{\text{I}}$ 表示发生 I 类漂移的模式集合，$\mathscr{D}_{\text{II}}$ 表示发生 II 类漂移的模式集合；剩下的则为保持了频繁/不频繁的正常模式。也就是说，式(7)根据漂移模式支持度的不同变化类型，将序列模式的漂移类型划分为了 3 类。

新的匹配会出现在新旧数据的连接处，这可能会使原本不频繁的模式变得频繁。例如，模式 A[0,2]B 在原始序列 AAAAA 上出现了 0 次，在增量数据 BBB 上出现了 0 次，但在新数据 AAAAABBB 上则出现了 6 次。因此，本文提出一种增补技术来获得这些新出现的次数。对于不同长度的模式而言，增补数据的大小也不同。对于给定长度为 k 的模式，用作增补的数据是原始数据的后缀 $S_{[n-(k-1)\times(\overline{G}-\underline{G}+1)+1,n]}$，其长度为 $(k-1)\times(\overline{G}-\underline{G}+1)$，则增补数据为：

$$\Delta S^+=S_{[n-(k-1)\times(\overline{G}-\underline{G}+1)+1,n]}\cup\Delta S \tag{8}$$

例如，原始数据 AAAAA 对模式 A[0,2]B 的后缀是 AAA。因此，增补数据为 AAABBB，由此则不会遗漏模式 A[0,2]B 新出现的次数。对任意模式 P 而言，根据其在原始数据和增补数据上是否频繁，可将模式 P 划分为 4 种情况之一，具体如表 3 所列。由表 3 可知，I 类漂移，即一个状态转移模式若从频繁变为不频繁，则只可能是策略 2)中的模式。类似地，II 类漂移，一个从不频繁变为频繁的模式，只可能来自策略 3)。而保持了原来的频繁信息的均视为正常(Normal)，即策略 1)和策略 4)。此外，将模式 P 的漂移程度定义为支持

度变化的绝对值,即:

$$d(P)=|sup(P,S)-sup(P,S^*)| \tag{9}$$

最后,当状态转移模式的长度为1时,算法会检测出发生两种漂移现象的状态。唯一不同的是,式(7)中的频繁阈值β需要被替换成α。

表3 状态转移模式三支漂移检测策略

Table 3 Three-way drift detection strategy of STAP

$S/\Delta S^+$	频繁	不频繁
频繁	1)正常:频繁(确定)	2)$\mathcal{D}_{\mathrm{I}}$:重新扫描 $S^+\cup\Delta S$(不确定)
不频繁	3)$\mathcal{D}_{\mathrm{II}}$:重新扫描 S(不确定)	4)正常:不频繁(确定)

3.2 问题描述与分析

问题1 增量挖掘的漂移模式检测

输入:$S^*=S\cup\Delta S,\alpha,\beta,\mathbb{P}$,以及$[\underline{G},\overline{G}]$

输出:$\{(P,d(P))|P\in\mathcal{D}_{\mathrm{I}}\cup\mathcal{D}_{\mathrm{II}}\}$

输出包含序列模式的两种漂移及其漂移程度。在最坏的情况下($\alpha=\beta=0$),频繁状态的数量为$\prod_{a\in A}|V_a|-1\geqslant 2^n-1$。由此可得长度为$k\in[1,m]$的状态转移模式的数量为$\sum_{k=1}^{m}(\prod_{a\in A}|V_a|-1)^k$。对于每个序列模式,最多可能在数据集上匹配$\sum_{k=1}^{m}(\prod_{a\in A}|V_a|-1)^k m\ (\overline{G}-\underline{G}+1)^{k-1}$次。时间复杂度是指数级的,这是频繁序列模式增量挖掘过程的复杂度,漂移检测任务与其是不可分的。适用于频繁序列模式挖掘的Apriori性质已经在文献[21]中得到证明。

4 基于增量学习的三支漂移序列检测算法

基于增量学习的三支漂移序列检测算法共有两个阶段:状态漂移挖掘和状态转移模式(STAP)漂移检测。最新的频繁状态即是长度为1的STAP,更长的候选模式基于此生成。通过对所有候选模式支持度的计算,来确定其漂移的类型并将其存入对应的集合中。即,由频繁变为不频繁的模式存入集合$\mathcal{D}_{\mathrm{I}}$,由不频繁变为频繁的模式存入集合$\mathcal{D}_{\mathrm{II}}$。第三类均被认定为正常,即保持了频繁或不频繁。此外,漂移检测与频繁STAP的增量挖掘同时进行,直到获得所有最新的频繁模式。具体过程如下。

阶段1 状态漂移检测

输入:$S,\Delta S,\alpha,1\text{-}\mathbb{P}$

输出:$1\text{-}\mathbb{P}^*$,$1\text{-}\mathcal{D}_{\mathrm{I}}$和$1\text{-}\mathcal{D}_{\mathrm{II}}$

(1)令$1\text{-}\mathbb{P}^*$为更新后的频繁STAP集合,初始化为空。令$1\text{-}\mathbb{S}^*$为更新后的频繁1状态集,初始化为空。在ΔS上挖掘频繁状态集合$1\text{-}\Delta\mathbb{S}$,将其与$1\text{-}\mathbb{P}$中的$1\text{-}\mathbb{S}$求交集。交集所含的状态在新数据上一定频繁,并将其存入$1\text{-}\mathbb{S}^*$,且令$1\text{-}\mathbb{P}^*=1\text{-}\mathbb{P}^*\cup 1\text{-}\mathbb{S}^*$。

(2)计算$1\text{-}\Delta\mathbb{S}$减去$1\text{-}\mathbb{S}$的差集。针对其中每个状态,扫描$S$并更新支持度,若$sup(s,S^*)\geqslant\alpha$,则判定其发生了II类漂移,并将其分别存入集合$1\text{-}\mathcal{D}_{\mathrm{II}}$和$1\text{-}\mathbb{S}^*$。若$sup(s,S^*)<\alpha$,则直接抛弃$s$。

(3)计算$1\text{-}\mathbb{S}$减去$1\text{-}\Delta\mathbb{S}$的差集。针对其中每个状态,扫描$\Delta S$并更新支持度。若$sup(s,S^*)\geqslant\alpha$,则并入$1\text{-}\mathbb{P}^*$。若$sup(s,S^*)<\alpha$,则判定其发生了I类漂移,并将其存入$1\text{-}\mathcal{D}_{\mathrm{I}}$。

(4)令$k\text{-}\mathbb{S}^*$($k\geqslant 2$)为更新后的频繁k状态集,初始化为空。若$(k-1)\text{-}\mathbb{S}^*$不为空,则将集合$(k-1)\text{-}\mathbb{S}^*$中的频繁状态两两组合以构建长度为$k$的候选状态$s$,并计算$\Delta S$上的支持度。若$s$在$k\text{-}\mathbb{P}$中,且在$\Delta S$上频繁,则更新其支持度并将其存入$k\text{-}\mathbb{S}^*$。

(5)若s存在于$k\text{-}\mathbb{P}$,但在ΔS上不频繁,则判定其发生了II类漂移,存入$1\text{-}\mathcal{D}_{\mathrm{II}}$中。若$s$不在$k\text{-}\mathbb{P}$中,但在$\Delta S$上频繁,则判定$s$发生了I类漂移,存入$1\text{-}\mathcal{D}_{\mathrm{I}}$中。令$1\text{-}\mathbb{P}^*=1\text{-}\mathbb{P}^*\cup k\text{-}\mathbb{S}^*$且$k=k+1$。

(6)重复步骤(4)和步骤(5),直到$k\text{-}\mathbb{S}^*$为空。检查$1\text{-}\mathbb{P}$剩下所有状态在ΔS上的支持度,直接判定其发生II类漂移,存入$1\text{-}\mathcal{D}_{\mathrm{II}}$中。

阶段2 状态转移模式漂移检测

输入:$S,\Delta S,\beta,[\underline{G},\overline{G}],\mathbb{P}$

输出:$\mathbb{P}^*$,$\mathcal{D}_{\mathrm{I}}$和$\mathcal{D}_{\mathrm{II}}$

(1)令$k\text{-}\mathbb{P}^*$($k\geqslant 2$)为更新后的频繁k-STAP,初始化为空。若$(k-1)\text{-}\mathbb{P}^*$不为空,则将集合$(k-1)\text{-}\mathbb{P}^*$中的模式与$1\text{-}\mathbb{P}^*$中的两两组合以构建长度为$k$的候选模式$P$。以参数$k$和$[\underline{G},\overline{G}]$构建增补数据$\Delta S^+$,并计算$P$在$\Delta S^+$上的支持度。

(2)若P在ΔS^+上频繁且存在于$k\text{-}\mathbb{P}$,则更新其支持度并存入$k\text{-}\mathbb{P}^*$中。若不存在于$k\text{-}\mathbb{P}$,则扫描S并更新$sup(P,S^*)$。若$sup(P,S^*)\geqslant\beta$,则判定发生II类漂移,将P存入$\mathcal{D}_{\mathrm{II}}$和$k\text{-}\mathbb{P}^*$中。若$sup(P,S^*)<\beta$,则抛弃$P$。

(3)若P在ΔS^+上不频繁且存在于$k\text{-}\mathbb{P}$中,则更新其支持度$sup(P,S^*)$,若$sup(P,S^*)<\beta$,则判定其发生I类漂移,将P存入$\mathcal{D}_{\mathrm{I}}$。若P在ΔS^+上不频繁且不存在于$k\text{-}\mathbb{P}$中,直接抛弃P。令$\mathbb{P}^*=\mathbb{P}^*\cup k-\mathbb{P}^*$且$k=k+1$。

(4)重复步骤(1)—步骤(3),直到$k\text{-}\mathbb{P}^*$为空。将$\mathbb{P}$中剩下的所有STAP判定为I类漂移,并存入$\mathcal{D}_{\mathrm{I}}$中。

5 实验和讨论

由于漂移检测和增量挖掘的过程密不可分,且增量挖掘的可扩展性已经在文献[21]中得到充分讨论。因此,本节通过实验讨论以下3个方面的主题:1)漂移模式的数量随阈值的变化规律;2)两类漂移模式的分布规律;3)所得漂移模式的可读性。

5.1 数据集和预处理

表4列出了本文采用的两个真实场景中的数据集,涉及空气质量(Dataset I)和石油工程(Dataset II)两个领域。

表4 数据集概述

Table 4 Outlines of datasets

数据集	名称	$\lvert T\rvert$	$\lvert A\rvert$
I	AirQuality	9358	15
II	ESP-Sanding-Diagnose	3893	22

表5列出了符号与数据波动程度之间的对应关系。其中,大写字母表示上升,小写字母表示下降。表达波动程度的不同符号是由相邻两个时刻之间的变化率决定的。本文工作与具体的离散化方法是低耦合的,采用不同的离散化只会得到相同的STAP释义,不会影响检测过程。

此外，增量数据的大小设置为 20%。

表 5 波动离散化符号的含义

Table 5 Meanings of fluctuation discretization symbols

符号	波动	符号	波动
A	[1%,2%)	a	(−2%,−1%]
B	[2%,4%)	b	(−4%,−2%]
C	[4%,8%)	c	(−8%,−4%]
D	[8%,16%)	d	(−16%,−8%]
E	[16%,32%)	e	(−32%,−16%]
F	[32%,64%)	f	(−64%,−32%]
G	[64%,128%)	g	(−128%,−64%]
H	[128%,+∞)	h	(−∞,−128%]
N	(−1%,1%)		

5.2 实验结果处理

由于被检测出的 STAP 漂移程度过低，取值范围是 $[10^{-4},10^{-6}]$。为了提高实验结果的可读性，本文引入了归一化技术来放大漂移程度。设发生 I 类漂移模式的 STAP 的数量为 Q，$\forall i\in[1,Q]$，P_i 是第 i 个 STAP，$d'(P)$ 表示归一化后的漂移程度，即：

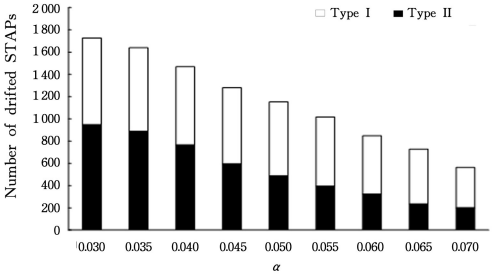
$$d'(P_i)=\frac{d(P_i)}{\sum_{i=1}^{n}d(P_i)} \tag{10}$$

II 类 STAP 漂移程度的归一化处理方式与此同理，且实验结果所展示的数据均经过了归一化处理。

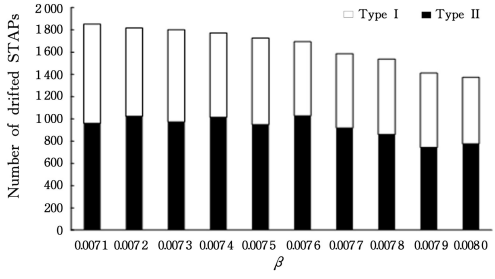
5.3 结果和分析

(1)变化规律

图 2 和图 3 分别给出了在 Dataset I 和 Dataset II 上随 α 和 β 变化发生漂移 $STAP$ 的数量变化规律。由图 2 和图 3 可知，随着阈值 α 和 β 的增加，发生漂移的模式总数以及两类漂移的模式数量均成下降的趋势。这是因为，当给定增量数据集的大小时，支持度越小的模式受到的影响越大。



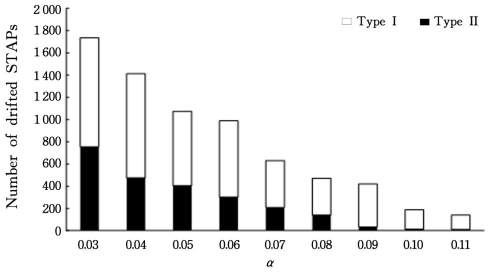
(a)$\beta=0.0075[\underline{G},\overline{G}]=[1,2]$)



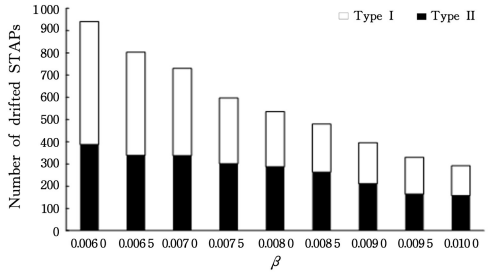
(b)$\alpha=0.03[\underline{G},\overline{G}]=[1,2]$)

图 2 Dataset I 上随 α，β 变化发生漂移 STAP 的数量分布

Fig. 2 Number distribution of STAP drifting with α and β changes on Dataset I



(a)$\beta=0.005[\underline{G},\overline{G}]=[1,2]$)

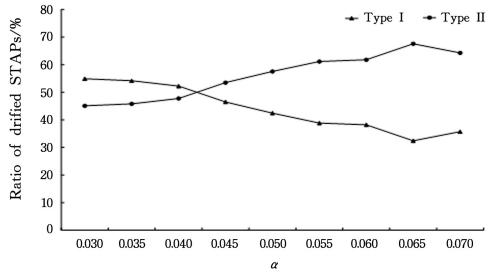


(b)$\alpha=0.05[\underline{G},\overline{G}]=[1,2]$)

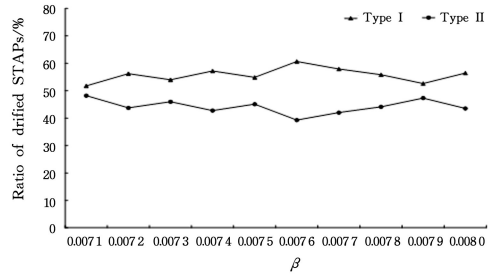
图 3 Dataset II 上随 α，β 变化发生漂移 STAP 的数量分布

Fig. 3 Number distribution of STAP drifting with α and β changes on Dataset II

图 4 给出了在 Dataset I 上两类漂移模式数量比例随 α 和 β 变化发生漂移 STAP 的数量比例的变化规律。



(a)$\beta=0.075[\underline{G},\overline{G}]=[1,2]$)



(b)$\alpha=0.03[\underline{G},\overline{G}]=[1,2]$)

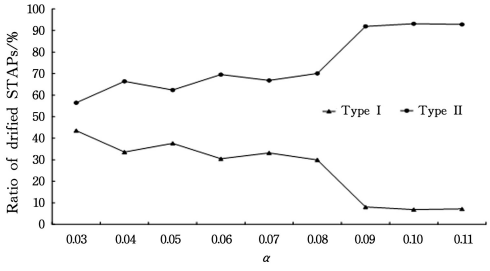
图 4 Dataset I 上随 α，β 变化发生漂移 STAP 的数量比例分布

Fig. 4 Number proportion distribution of STAP drifting with α and β changes on Dataset I

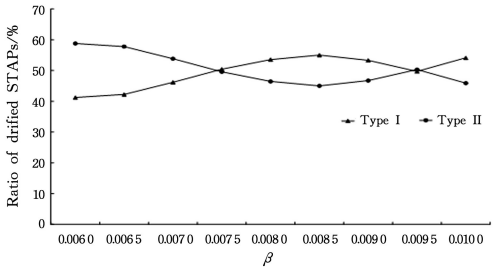
由图 4(a)可知，I 类模式的数量比例随着 α 的增加而降低，但 II 类模式的数量比例在升高。有趣的是，II 类模式

数量比例的变化趋势与数量相反，这说明增量数据对支持度在当前 α 左右，但不频繁的模式影响更大。换句话说，虽然随着 α 的增加，频繁模式总量在减少，但由不频繁变为频繁的模式越来越多。由图 4(b)可知，随着 β 的增加，I 类和 II 类的比例变化趋势稳定，I 类模式的数量比例始终比 II 类高 3.56%～21.37%。对于 Dataset I 而言，相比 β，α 的改变所造成的频繁模式漂移现象更明显。

图 5 给出了在 Dataset II 上两类漂移模式数量比例随 α 和 β 变化发生漂移 STAP 的数量比例的变化规律。由图 5(a)可知，I 类和 II 类模式的数量比例变化规律与 Dataset I 相同；由图 5(b)可知，对于 Dataset II 而言，相比 α，β 的改变所造成的频繁模式漂移现象更明显，I 类模式的数量比例在 β 取 0.0085 时最大，为 55%，而 II 类模式的数量比例在 β 取 0.006 时最大，为 58.77%。综上所述，不同数据集的模式漂移现象迥异。



(a)β=0.005[$\underline{G}$,$\overline{G}$]=[1,2])



(b)α=0.05[$\underline{G}$,$\overline{G}$]=[1,2])

图 5 Dataset I 上随 α、β 变化发生漂移 STAP 的数量比例分布

Fig. 5 Number proportion distribution of STAP drifting with α and β changes on Dataset II

(2)分布规律

表 6－表 9 列出了 Datasets I 和 Datasets II 上随 β 变化发生 I 类漂移和 II 类漂移模式数量的分布拟合误差。由表 6 可知，I 类漂移模式的拟合分布主要服从 Weibull 分布；由表 7 可知，II 类漂移模式的拟合分布主要服从 beta 分布；由表 8 可知，Datasets II 上的 I 类漂移模式的拟合分布主要服从 Logistic 分布，而 II 类漂移模式只服从 beta 分布。

表 6 Dataset I 上随 β 变化发生 I 类漂移模式的分布拟合误差

Table 6 Distribution fitting error of type I drift STAPs on Dataset I with β changes

分布\β	0.0071	0.0072	0.0073	0.0074	0.0075	0.0076	0.0077	0.0078	0.0079
Beta	0.00923	0.00727	**0.00521**	0.00754	0.00876	**0.00587**	0.00716	**0.00908**	0.01118
Logistic	0.01040	0.00942	0.00884	0.00927	0.01134	0.01104	0.01085	0.01334	0.01323
Normal	0.01040	0.00877	0.00827	0.00895	0.01013	0.01070	0.01019	0.01280	0.01304
Exponential	0.02021	0.02010	0.01969	0.01984	0.01828	0.01762	0.01681	0.02210	0.02412
Weibull	**0.00880**	**0.00666**	0.00559	**0.00668**	**0.00857**	0.00634	**0.00655**	0.00939	**0.01041**

表 7 Dataset I 上随 β 变化发生 II 类漂移模式的分布拟合误差

Table 7 Distribution fitting error of type II drift STAPs on Dataset I with β changes

分布\β	0.0071	0.0072	0.0073	0.0074	0.0075	0.0076	0.0077	0.0078	0.0079
Beta	0.01574	**0.01319**	**0.01478**	**0.01091**	**0.01214**	**0.01529**	**0.00817**	**0.00854**	**0.00828**
Logistic	0.01621	0.01662	0.01613	0.01377	0.01387	0.01867	0.01484	0.01496	0.01652
Normal	0.01874	0.02129	0.01912	0.02006	0.01847	0.02440	0.02068	0.02119	0.02168
Exponential	0.03000	0.02885	0.0306	0.03000	0.02963	0.03588	0.03076	0.02802	0.02638
Weibull	**0.01463**	0.01397	0.0151	0.01278	0.01255	0.01840	0.01178	0.01102	0.01046

表 8 Dataset II 上随 β 变化发生 I 类漂移模式的分布拟合误差

Table 8 Distribution fitting error of type I drift a STAPs on Dataset II with β changes

分布\β	0.006	0.0065	0.007	0.0075	0.008	0.0085	0.009	0.0095	0.01
Beta	0.19307	0.14354	0.13626	0.18399	0.17971	0.16804	0.15532	0.16913	0.15291
Logistic	**0.18617**	**0.13541**	**0.12936**	0.17976	**0.17149**	**0.16122**	0.15009	0.16786	0.14904
Normal	0.18696	0.13689	0.12983	**0.17948**	0.17232	0.16195	**0.14938**	0.16724	**0.14792**
Exponential	0.20014	0.15157	0.14439	0.19391	0.18907	0.17658	0.16800	0.17974	0.16205
Weibull	0.18779	0.13701	0.13098	0.18068	0.17371	0.16339	0.14988	**0.16681**	0.14832

表 9 Dataset II 上随 β 变化发生 II 类漂移模式的分布拟合误差

Table 9 Distribution fitting error of type II drift STAPs on Dataset II with β changes

分布\β	0.006	0.0065	0.007	0.0075	0.008	0.0085	0.009	0.0095	0.01
Beta	**0.01582**	**0.01738**	**0.01970**	**0.02213**	**0.02335**	**0.02848**	**0.02805**	**0.02525**	**0.02255**
Logistic	0.04203	0.04342	0.04084	0.04523	0.04593	0.05459	0.06162	0.06330	0.05765
Normal	0.05344	0.05199	0.05266	0.05844	0.05808	0.06585	0.08349	0.08528	0.07749
Exponential	0.04637	0.04543	0.04289	0.04843	0.04876	0.04915	0.05112	0.04084	0.03937
Weibull	0.02154	0.02316	0.02160	0.02892	0.02947	0.03508	0.03550	0.02731	0.02669

(3)可读性

表 10 列出了在 Datasets I 和 Datasets II 上漂移程度居前五的、长度为 2 的 STAP 漂移检测结果，模式末尾的括号内标注的是该模式归一化后的漂移程度。对比表 10(a)和表 11(b)以及表 11(a)和表 11(b)，本组实验的参数设置与第 1 组实验相同。

表 10 Dataset I 上长度为 2 且漂移程度居前五的状态转移模式

Table 10 STAPs of length 2 and with top-5 drift degrees on Dataset I

(a) Type I

{(CO(GT),f),(PT08.S1(CO),d),(C6H6(GT),f)}→{(NOx(GT),h)}(0.24%)
{(CO(GT),f),(PT08.S1(CO),d),(C6H6(GT),f)}→{(NO2(GT),h)}(0.24%)
{(CO(GT),f),(PT08.S1(CO),d),(C6H6(GT),f)}→{(NOx(GT),h),(NO2(GT),h)}(0.24%)
{(T,b)}→{(PT08.S1(CO),c),(PT08.S4(NO2),b)}(0.24%)
{(T,b)}→{(CO(GT),f),(PT08.S5(O3),e)}(0.22%)

(b) Type II

{(NOx(GT),d)}→{(T,d)}(0.62%)
{(T,d)}→{(NOx(GT),d)}(0.6%)
{(CO(GT),D)}→{(T,d)}(0.53%)
{(PT08.S4(NO2),b)}→{(T,d)}(0.52%)
{(T,d)}→{(RH,B)}(0.5%)

表 11 Dataset II 上长度为 2 且漂移程度居前五的状态转移模式

Table 11 STAPs of length 2 and with top-5 drift degrees on Dataset II

(a) Type I

{(oil pressure(Mpa),d)}→{(production fluid(m3),B)}(0.54%)
{(oil pressure(Mpa),d)}→{(producing water(m3),C)}(0.54%)
{(oil pressure(Mpa),d)}→{(gassiness(104m3),f)}(0.5%)
{(oil pressure(Mpa),d)}→{(containingwater(%),b)}(0.49%)
{(oil pressure(Mpa),c)}→{(oilpressure(Mpa),C)}(0.47%)

(b) Type II

{(production fluid(m3),B)}→{(frequency(Hz),H)}(0.17%)
{(oil-producing(m3),A)}→{(frequency(Hz),H)}(0.17%)
{(gassiness(104m3),c)}→{(production fluid(m3),b),(oil-producing(m3),b)}(0.17%)
{(frequency(Hz),H)}→{(production fluid(m3),B)}(0.17%)
{(frequency(Hz),H)}→{(oil-producing(m3),A)} (0.17%)

由此可知，I 类和 II 类漂移的 STAP 模式明显不同，具有较强的多样性。以表 10 对应的空气质量数据为例，模式“{(T,b)}→{(CO(GT),f),(PT08.S5(O3),e)}(0.22%)”读作温度下降 2%～4%后，在 1 到 2 个单位时间间隔内，CO(GT)气体含量会下降 32%～64%且 PT08.S5(O3)气体含量会下降 16%～32%。该模式从频繁变为不频繁，可能性变小了 0.22%。模式“{(CO(GT),D)}→{(T,d)}(0.53%)”读作在 CO(GT)气体含量上升 8%～16%后，在 1 到 2 个单位时间间隔内，温度会下降 8%～16%。该模式从不频繁变得频繁，可能性增加了 0.53%。其余漂移模式的解释方法同理。

结束语 针对多元时序数据上的模式漂移检测问题，本文提出了一种增量式的三支漂移检测算法。相对于已有工作，本文工作可在不增加任何参数的情况下，同时获得频繁模式和发生漂移的模式。漂移检测的三分策略关注的是在增量更新前后，模式的频繁信息发生变化的类型，即频繁变不频繁、不频繁变频繁以及保持频繁或不频繁 3 类。保持频繁的模式被保存，保持不频繁的被抛弃。通过在多个领域的数据集上的实验验证了算法的正确性、实用性，以及漂移检测结果的可解释性。实验中漂移程度均为归一化后的结果。

未来将在以下 4 个方面推进本文工作：

(1)研究一次性匹配条件和无重叠区间匹配条件下的状态转移模式漂移检测算法。

(2)采用深度优先或并行技术提升模式的支持度计算效率。

(3)引入更丰富的离散化技术如 SAX，以获得不同的漂移检测结果。

(4)探索更合理的模式漂移检测性能的评价指标、机制和方案。

参 考 文 献

[1] PAWLAK Z. Rough sets [J]. International Journal of Computer & Information Sciences, 1982, 11(5): 341-356.

[2] YAO Y Y. Three-way decisions and cognitive computing [J]. Cognitive Computation, 2016, 8(4): 543-554.

[3] YAO Y Y. The geometry of three-way decision[J/OL]. Applied Intelligence, 2021: 1-28. https://doi.org/10.1007/s10489-020-02142-z.

[4] LI J H, HUANG C C, QI J J, et al. Three-way cognitive concept learning via multi-granularity[J]. Information Sciences, 2017, 378: 244-263.

[5] MAO H, ZHAO S F, YANG L Z. Relationships between three-way concepts and classical concepts[J]. Journal of Intelligent & Fuzzy Systems, 2018, 35(1): 1063-1075.

[6] DENG X F, YAO Y Y. Decision-theoretic three-way approximations of fuzzy sets[J]. Information Sciences, 2014, 279: 702-715.

[7] YAO Y Y. Interval sets and three-way concept analysis in incomplete contexts[J]. International Journal of Machine Learning and Cybernetics, 2017, 8(1): 3-20.

[8] FANG Y, MIN F. Cost-sensitive approximate attribute reduction with three-way decisions[J]. International Journal of Approximate Reasoning, 2019, 104: 148-165.

[9] MIN F, LIU F L, WEN L Y, et al. Tri-partition cost-sensitive active learning through kNN[J]. Soft Computing, 2019, 23(5): 1557-1572.

[10] YE X, LIU D. An interpretable sequential three-way recommendation based on collaborative topic regression[J/OL]. Expert Systems with Applications, 2021, 168. https://doi.org/10.1016/j.eswa.2020.114454.

[11] ZHANG H R, MIN F, SHI B. Regression-based three-way recommendation[J]. Information Sciences, 2017, 378: 444-461.

[12] MIN F, ZHANG S M, CIUCCI D, et al. Three-way active learning through clustering selection[J]. International Journal of Machine Learning and Cybernetics, 2020, 11(5): 1033-1046.

[13] YUE X D, CHEN Y F, MIAO D Q, et al. Tri-partition neighbor-

hood covering reduction for robust classification[J]. International Journal of Approximate Reasoning,2017,83:371-384.
[14] YU H,WANG X C,WANG G Y,et al. An active three-way clustering method via low-rank matrices for multi-view data[J]. Information Sciences,2020,507:823-839.
[15] MIN F,ZHANG Z H,ZHAI W J,et al. Frequent pattern discovery with tri-partition alphabets[J]. Information Sciences,2020,507:715-732.
[16] LI H X,ZHANG L B,HUANG B,et al. Sequential three-way decision and granulation for cost-sensitive face recognition[J]. Knowledge-Based Systems,2016,91:241-251.
[17] REN R S,WEI L. The attribute reductions of three-way concept lattices[J]. Knowledge-based systems,2016,99:92-102.
[18] ZHOU B,YAO Y Y,LUO J G. Cost-sensitive three-way email spam filtering[J]. Journal of Intelligent Information Systems,2014,42(1):19-45.
[19] ZHUANG D E H,LI G C L,WONG A K C. Discovery of temporal associations in multivariate time series[J]. IEEE Transactions on Knowledge and Data Engineering,2014,26(12):2969-2982.
[20] ZHANG Z H,MIN F. Frequent state transition patterns of multivariate time series[J]. IEEE Access,2019,7:142934-142946.
[21] ZENG S C,ZHANG Z H,MIN F,et al. A three-way incremental updating method of state transition pattern [J]. Journal of Zhengzhou University (Natural Science Edition),2020,52(1):16-23.
[22] MIN F,WU Y X,WU X D. The Apriori property of sequence pattern mining with wildcard gaps[J]. International Journal of Functional Informatics and Personalized Medicine,2012,4(1):15-31.
[23] WU X D,ZHU X Q,HE Y,et al. PMBC:pattern mining from biological sequences with wildcard constraints[J]. Computers in Biology and Medicine,2013,43(5):481-492.
[24] WU Y X,TONG Y,ZHU X Q,et al. NOSEP:Nonoverlapping sequence pattern mining with gap constraints[J]. IEEE Transactions on Cybernetics,2017,48(10):2809-2822.
[25] QIAN Y K,CHEN M,YE L X,et al. Network-wide anomaly detection method based on multiscale principal component analysis[J]. Journal of Software,2012 (2):361-377.
[26] ZHOU D H,WEI M H,SI X S. A survey on anomaly detection,life prediction and maintenance decision for industrial processes [J]. Acta Automatica Sinica,2013,39(6):711-722.
[27] MAO J L,JIN C Q,ZHANG Z G,et al. Anomaly detection for trajectory big data:advancements and framework[J]. Journal of Software,2017,28(1):17-34.
[28] YOU C C,FENG X P,LIU L J,et al. An abnormal chest X-ray diagnostic report detection method based on topic model[J]. Computer Engineering & Science,2020,42(4),741-748.
[29] MEI Y D,CHEN X,SUN Y Z,et al. A method for software system anomaly detection based on log in formation and CNN-text [J]. Chinese Journal of Computers,2020,43(2):366-380.
[30] CHU G,HU X G,ZHANG Y H. Semantic-based Concept Drift Detection Algorithm for Text Data Stream[J]. Computer Engineering,2018,44(2):24-30.
[31] ZHOU Y J,XU C,LI J G. Unsupervised anomaly detection method based on improved CURE clustering algorithm[J]. Journal on Communications,2010,31(7):4-23.
[32] LI N,GUO G D,CHEN L F. Concept drift detection method with limited amount of labeled data[J]. Journal of Computer Applications,2012,32(8):2176-2185.
[33] CHENG G,QIAN D X,GUO J W,et al. A classification approach based on divergence for network traffic in presence of concept drift[J]. Journal of Computer Research and Development,2020,57(12):2673-2682.
[34] HU M,BAI X,XU W,et al. Review of anomaly detection algorithms for multidimensional time series[J]. Journal of Computer Applications,2020,40(6) 1553-1564.
[35] LIAN Y F,DAI Y X,WANG H. Anomaly detection of user behaviors based on profile mining[J]. Chinese Journal of Computers,2002,25(3):325-330.
[36] TIAN X G,GAO L Z,SUN C L,et al. Anomaly detection of program behaviors based on system calls and homogeneous markov chain models[J]. Journal of Computer Research and Development,2007(9):1538-1544.
[37] XIAO H,HU Y F. Data mining based on segmented time warping distance in time series database[J]. Journal of Computer Research and Development,2005,42(1):72-78.
[38] KEOGH E,LONARDI S,CHIU W. Finding Surprising Patterns in a Time Series Database In Linear Time and Space[C]// Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York:ACM Press,2002:550-556.
[39] YU B J,XIA Z G,WANG J L. Anomaly detection algorithm based on gaussian process model[J]. Computer Engineering and Design,2016,37(4):914-920.

SHEN Shao-peng, born in 1993, postgraduate, is a member of China Computer Federation. His main research interests include reinforcement learning and anomaly detection.

ZHANG Zhi-heng, born in 1990, Ph.D, is a member of China Computer Federation. His main research interests include time-series analysis, three-way decision and cost-sensitive learning.

(责任编辑:李亚辉)