

水质时间序列模式挖掘

夏 达 李士进

(河海大学 计算机与信息学院 江苏 南京 210098)

摘 要: 对水质时间序列进行数据挖掘,找出其蕴含的模式,对于水资源的改善有重要的现实意义。针对带间隔约束的有序时间序列的模式挖掘,现有算法多按左优先匹配以完备性为代价加快效率或枚举可能位置损失效率提高完备性。为了提高模式挖掘的效率同时保证一定的完备性,提出一种满足 One-Off 条件的带有间隔约束的单序列模式挖掘算法 FOFM (fast one-offing mining)。算法首先扫描序列获得长度为 1 的模式,再通过将当前长度的所有频繁模式进行两两比较,而后连接可连接的模式以形成新的模式,在模式连接的过程中记录候选模式最后事件的可能位置并通过回溯位置序列的方法检查模式的支持度,直至无法生成新的模式。实验结果表明,FOFM 算法在水质时间序列上相较于相关序列模式挖掘算法拥有较高的效率和一定的完备性。

关键词: 数据挖掘; 序列模式挖掘; 间隔约束; One-Off 条件

中图分类号: TP311

文献标识码: A

文章编号: 1673-629X(2018)05-0149-05

doi: 10.3969/j.issn.1673-629X.2018.05.034

Pattern Mining for Water Quality Time Series

XIA Da, LI Shi-jin

(School of Computer and Information, Hohai University, Nanjing 210098, China)

Abstract: It's important to mine water quality time series and find out its pattern in series for the improvement of water resources. For the mining of time series with interval constraints, most of existing algorithms reduce the completeness by left-first matching for the efficiency or reduce the efficiency by enumerating the possible position for the completeness. In order to improve the efficiency of pattern mining and maintain a high degree of completeness, we propose a fast one-offing mining algorithm with One-off condition and gap constraints. It first scans the sequence to obtain the pattern of length 1, and then obtains the candidate pattern through comparison on all frequent patterns of the current length after connection of the pattern. The possible position of the last event of the candidate pattern during the pattern connection is recorded and the support of the pattern by the backtracking sequence is checked, until a new pattern can't be generated. The experiment based on the water quality time series proves that the FOFM is more effective than the related sequential pattern mining algorithm with a certain completeness.

Key words: data mining; sequential pattern mining; gap constraints; One-Off condition

0 引言

水资源与人们的生产生活密切相关,水污染治理问题受到政府的高度重视^[1-2]。随着各地水质监测站点的建设及水质监测水平的提高,得到了大量的水质时间序列,对这些序列进行相应的序列模式挖掘研究,找出水质时间序列中隐藏的模式,对当前水质的保护及改善水资源环境研究相关水质对策都有极其重要的意义。

序列模式挖掘由 Agrawal 和 Srikant 提出,主要用于在给定的数据集中搜索反复出现的模式。随后,学

者们陆续提出了多种序列模式挖掘算法^[3-11]。文献[12]研究了满足 One-Off 条件的单序列模式挖掘问题,通过使用两种不同的搜索方法计算其支持度并取较大值,提高了模式挖掘的完备性。文献[13]结合了 One-Off 条件和通配符对单序列模式进行挖掘,并提出 MAIL 算法,算法同样采用两种策略结合计算模式支持度,有效提高了模式挖掘的效率及完备性。文献[14]在 MAIL 算法的基础上,提出了 OFMI 算法及 I-OFMI 算法,OFMI 是一种快速的带通配符和 One-Off 条件的序列模式挖掘算法,但同时它也不可避免地会

收稿日期: 2017-06-15

修回日期: 2017-10-10

网络出版时间: 2018-02-08

基金项目: 国家公益性科研项目(201501022); 江苏省重点研发计划项目(BE2015707)

作者简介: 夏 达(1992-),男,硕士研究生,通信作者,研究方向为数据挖掘;李士进,教授,CCF 会员,研究方向为模式识别、数据挖掘。

网络出版地址: <http://kns.cnki.net/kcms/detail/61.1450.tp.20180207.1917.096.html>

遗失部分模式。为了解决 OFMI 算法缺失模式的问题,文献[14]进一步提出了基于前向搜索和后向寻找的模式支持度的计算方法 I-OFMI,进一步提高了解的完备性。I-OFMI 算法相较于 OFMI 算法提升了模式发现的完备性,但同时也不可避免地牺牲了模式发现的效率。尽管与传统算法相比,OFMI 和 I-OFMI 算法分别提升了模式挖掘的效率及完备性,但较完备算法其效率仍较差。

水质时间序列具有高维性、复杂性、动态性等特点^[15]。水质的变化不仅受多种环境因素的影响,如气候变化、季节变化等^[16-18],一些突发事件如工厂违规排放污水、蓝藻暴发等也会导致水质急剧变化,这使得水质时间序列还具有一定的随机性^[18]。目前相关算法应用于水质时间序列存在完备性较差或效率较差的问题。因此,为了提高模式挖掘的效率,文中设计了一种新的算法应用于水质时间序列,并通过实验对其进行验证。

1 问题定义

定义 1: 给定序列 $S = s_1s_2 \cdots s_n$, 其序列长度为 n , 序列字符集记为 Σ , 代表序列中包含的所有不同字符, 序列字符集长度记为 $|\Sigma|$ 。例如序列: ccccbabbb, 其序列长度为 10, 序列字符集为 $\{a, b, c\}$, 序列字符集长度为 3。

定义 2: 通配符记为 $*$, 代表序列字符集中的任意字符。

定义 3: 间隔约束即通配符的个数范围, 最小间隔记为 N , 最大间隔记为 M , $M - N + 1$ 为间隔灵活度。

定义 4: 模式为序列字符集中的字符和通配符所组成的序列, 记为 $P = p_1p_2 \cdots p_m$, 其中 $p_i (1 \leq i \leq m)$ 为通配符或字符, 模式中通配符可以省略, 即 p_i 代表字符集中的字符, 其中 p_i 与 $p_{i-1} (2 \leq i \leq m)$ 的位置需满足相应间隔约束。文中的模式均为省略了通配符的模式。

定义 5: 对于位置序列 $I = i_1i_2 \cdots i_m, 1 \leq i_1 \leq \cdots \leq i_m \leq n$, 若满足对于任意的 $k (1 \leq k \leq m)$ 使得 $s_{i_k} = p_k$, 则称位置序列 I 为模式 P 在序列 S 中的一次出现。

定义 6: One-Off 条件即模式在序列中的任意两次出现均满足两次出现的位置序列不共用相同的位置, 例如 $bc bc$, $bc \{ (0, 1) \} \{ (2, 3) \}$ 满足 One-Off 条件而 $bc \{ (0, 1) \} \{ (0, 3) \}$ 不满足 One-Off 条件, 因为两次出现共用了位置 0。

定义 7: 模式在序列中所有满足 One-Off 条件的出现个数即为模式的支持度, 若模式的支持度大于相应的支持度阈值, 则称该模式为频繁模式。

定义 8: 对于模式 $P = p_1p_2 \cdots p_m$, 模式 $Q =$

$q_1q_2 \cdots q_t (t \leq m)$, 如果存在 $1 \leq i_1 \leq \cdots \leq i_t \leq m$ 满足 $p_{i_k} = q_k (1 \leq k \leq t)$, 则称 Q 为 P 的子模式, P 为 Q 的父模式。若该位置序列为一个公差为 1 的等差序列, 则称 Q 为 P 的连续子模式, P 为 Q 的连续父模式。

定义 9: 对于模式 $P = p_1p_2 \cdots p_m$, 模式 $Q = q_1q_2 \cdots q_t (t < m)$, 若 Q 为 P 的连续子模式且 $i_1 = 1$, 则称 Q 为 P 的前缀模式, 若 $t = m - 1$, 则称 Q 为 P 的最大前缀模式。

定义 10: 对于模式 $P = p_1p_2 \cdots p_m$, 将 p_m 在序列 S 中所有可能出现的位置序列记为模式 P 的尾序列。尾序列的大小即为 p_m 在序列 S 中所有可能出现的位置个数。

定义 11: 对于模式 $P = p_1p_2 \cdots p_m$, 模式 $Q = q_1q_2 \cdots q_m$, 若对于任意的 $k (2 \leq k \leq m)$ 均满足 $p_k = q_{k-1}$, 则称模式 P 与模式 Q 是可连接的, 其连接结果为 $p_1q_1q_2 \cdots q_m$ 。将模式 P 的尾序列记为新模式的前序列, 模式 Q 的尾序列记为新模式的后序列。

定理 1: 根据 Apriori 性质, 若模式 P 为频繁模式, 则 P 的所有非空连续子模式也一定是频繁的, 也即如果模式 P 为非频繁模式, 则 P 的所有连续父模式也一定是非频繁的。

2 一种新的序列模式挖掘算法

文中提出的 FOFM (fast one-offing mining) 算法首先扫描序列获得所有长度为 1 的频繁模式集, 并记录每个 1-项频繁模式的尾序列, 紧接着进行模式的连接过程。在由长度 $k-1$ 的频繁模式连接生成长度为 k 的候选模式集的过程中, 由两个符合可连接条件的长度为 $k-1$ 的频繁模式的尾序列进行连接, 连接过程中只需考虑 k -项模式的最后一个字符的可能位置即可, 在完成连接后再从 k -模式的最后可能位置序列通过反复提取其最大前缀模式的方法向 1-模式回溯, 回溯过程中遵循 One-Off 条件并采取右优先策略直至计算完成 k -模式的支持度。

FOFM 算法的具体步骤如下:

(1) 遍历序列 S , 获得序列 S 的字符集及字符集中每个字符对应的位置序列, 将结果存储在相应结构中。

(2) 检查字符集中每个字符所对应的位置序列的大小, 去掉小于最小支持度的字符, 使得每个字符为 1-频繁模式。

(3) 遍历当前长度的所有频繁模式进行两两比较, 将可连接的模式连接形成新的模式。

(4) 根据已经存储的结果, 获得用于连接的两个模式的位置序列, 对两个位置序列中的位置进行两两比较。如果间隔大于最大间隔, 则继续使用前序列中的下一个位置与后序列进行比较; 如果间隔满足间隔

约束,则存储该后序列中的当前位置并使用后序列的下一个位置继续进行比较;如果间隔小于最小间隔,则使用后序列的下一个位置继续进行比较,直到两个序列中某一个遍历完毕。

(5) 获得步骤4得到的新模式的尾序列后,如果新模式的尾序列的大小小于最小支持度,则说明新模式不是频繁模式,去除该模式,否则检查新模式的支持度,如果新模式的支持度满足最小支持度,则存储该模式及其尾序列。

(6) 所有当前模式处理完毕后,将新的频繁模式视为当前模式转入步骤3进行处理,直至无法连接形成新的模式。

支持度检查的具体步骤如下:

(1) 从新模式的尾序列开始,从大到小选取一个未被标记的位置,记录进标记数组。

(2) 获得当前模式的当前前缀模式的尾序列,从大到小与之前选取的位置比较,直到找到满足间隔约束的未被标记位置,将该位置记录进标记数组。

(3) 将最大前缀模式设为当前模式,重复步骤2,直到无法获得最大前缀模式,即模式长度为1。

以序列 $S = ccccbabbb$ 为例,间隔约束设为 $[0, 2]$, 最小支持度设为2。FOFM 算法首先遍历序列 S 获得 $c\{0, 1, 2, 3, 4\}$, $b\{5, 7, 8, 9\}$, $a\{6\}$ 。很明显模式 a 不符合最小支持度要求,去除模式 a 后, c 和 b 都是1-频繁模式。

对1-频繁模式 c 和 b 连接形成 $bb\{7, 8, 9\}$, $bc\{\}$, $cb\{5, 7\}$, $cc\{1, 2, 3, 4\}$, bc 不符合最小支持度要求被删除,对剩余模式 bb , cb , cc 检查其支持度, bb 存在 $\{(8, 9), (5, 7)\}$ 两个位置序列符合要求,同理 $cb\{\{(3, 5), (4, 7)\}\}$, $cc\{\{(3, 4), (1, 2)\}\}$ 符合要求,至此获得2-频繁模式 bb , cb , cc 。

对2-频繁模式两两连接形成 $bbb\{8, 9\}$, $cbb\{8, 9\}$, $ccb\{5, 7\}$, $ccc\{2, 3, 4\}$, 分别检查支持度得 $bbb\{\{(7, 8, 9)\}\}$, $cbb\{\{(4, 7, 9), (3, 5, 8)\}\}$, $ccb\{\{(3, 4, 7), (1, 2, 5)\}\}$, $ccc\{\{(2, 3, 4)\}\}$ 通过检查支持度删除不符合要求的 bbb 和 ccc 。

对3-频繁模式两两连接形成 $cbbb\{8, 9\}$, 检查支持度的 $cbbb\{\{(3, 4, 7, 9), (1, 2, 5, 8)\}\}$ 符合支持度要求。

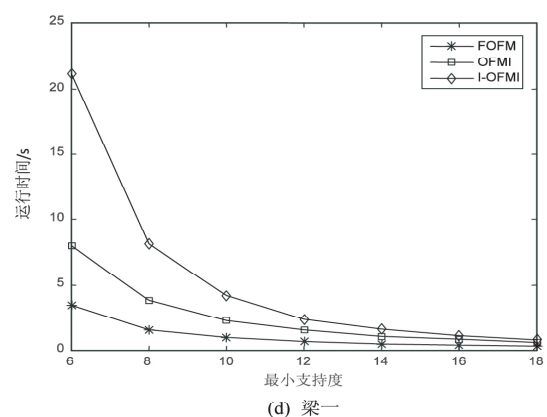
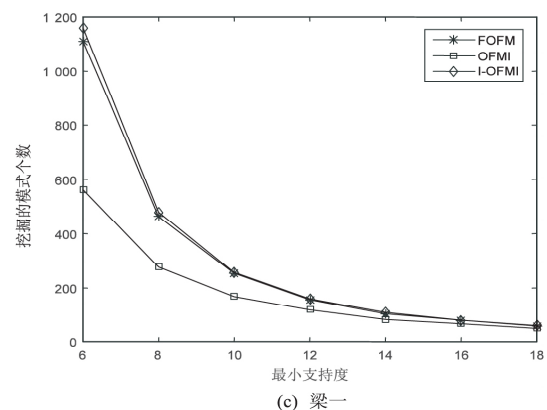
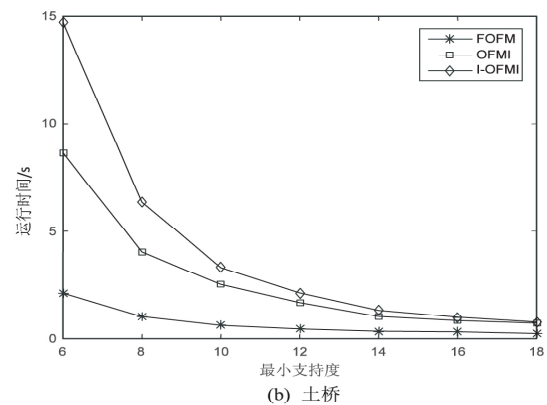
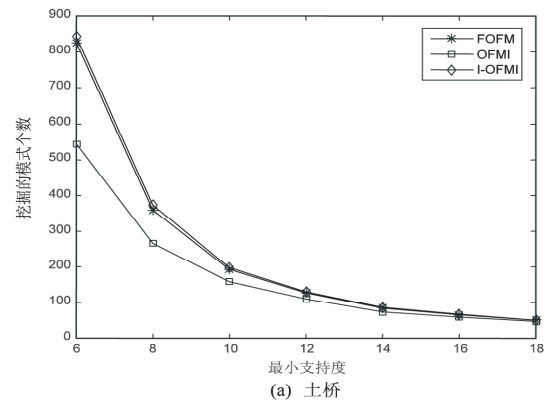
由于无法继续连接形成新的模式,FOFM 算法终止。

3 实验结果与分析

选取南京土桥,东台梁一,盐城新洋港3个水质站点2007-2016年的水质时间序列,序列1土桥长度为521,序列2梁一长度为511,序列3新洋港长度为

509,使用算法 OFMI、I-OFMI、FOFM 进行挖掘。为充分比较算法,分别在不同支持度、不同通配符的长度下对3种算法的模式挖掘数量及算法的运行时间进行对比。

实验一: \min_sup 分别设为6, 8, 10, 12, 14, 16, 18, 间隔约束为 $[0, 2]$ 结果如图1所示。



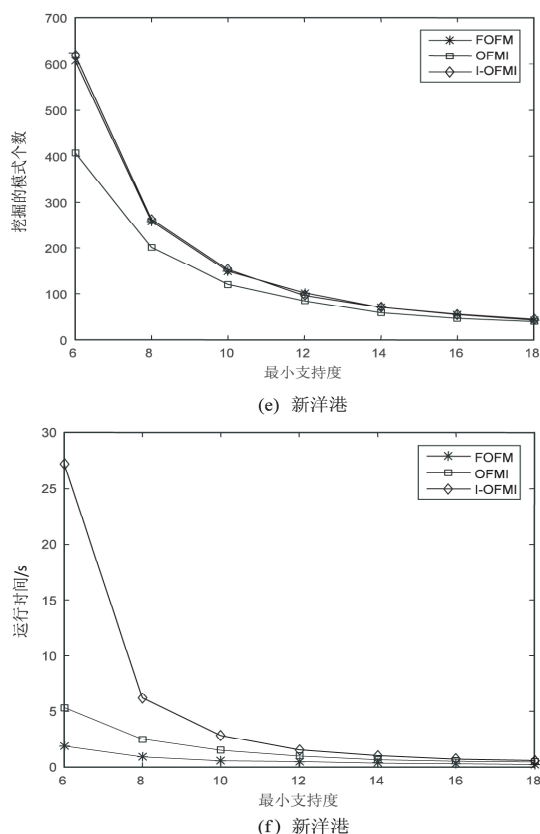


图 1 不同支持度下模式个数及运行时间结果对比

通过实验一可以发现,所有算法挖掘的模式个数和运行时间都随着最小支持度的增加而减少,这其中 OFMI 算法挖掘的模式个数最少,其效率处于 FOFM 算法与 I-OFMI 算法之间。I-OFMI 算法挖掘的模式较多但效率最差。文中提出的 FOFM 算法在不同支持度下运行速度都较快,FOFM 算法挖掘的模式个数与 I-OFMI 算法的挖掘结果差距较小,如序列 1 在最小支持度为 6 的情况下, I-OFMI 运行时间为 14.7 s 时,算法挖掘模式个数为 843,而 FOFM 算法运行时间为 2.1 s 时,挖掘模式个数为 826。相比 OFMI 算法, FOFM 算法挖掘的模式个数比 OFMI 算法更多,运行时间却更少。

实验二: min_sup 设为 20,最小间隔为 0,最大间隔长度分别为 2 3 4 5 6 7 8 结果如图 2 所示。

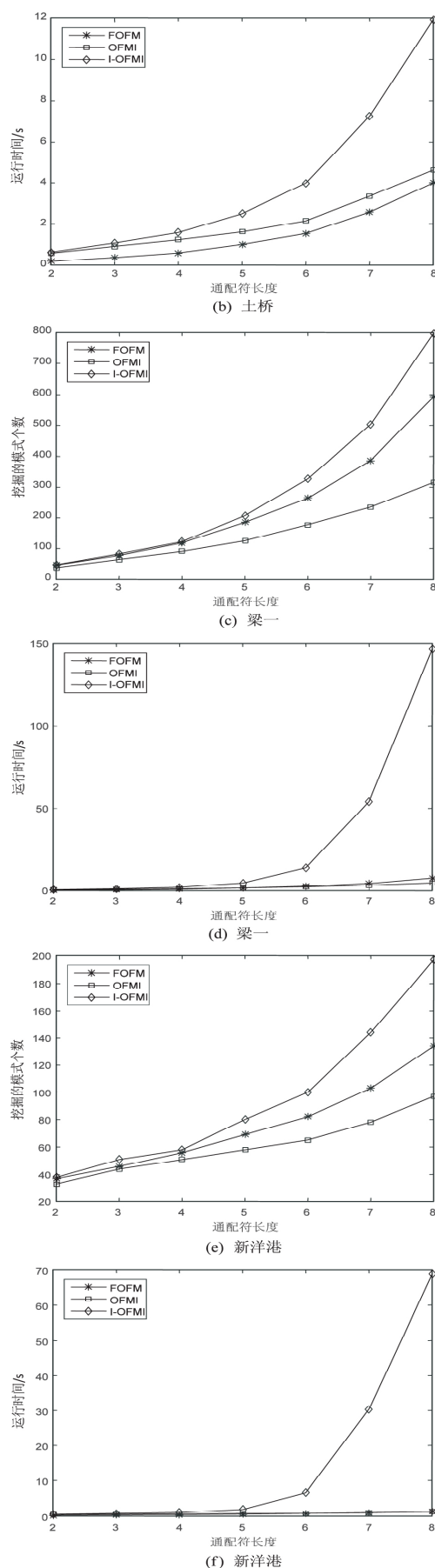
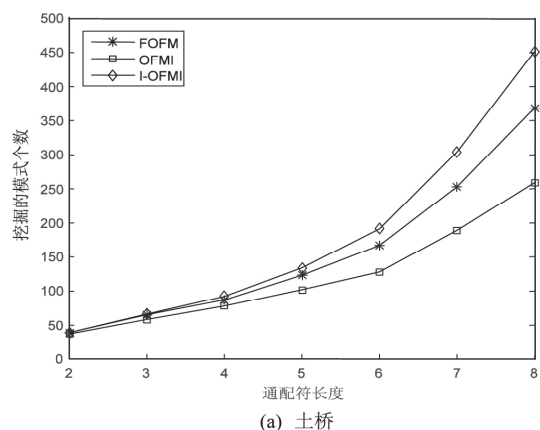


图 2 不同通配符长度下挖掘模式个数及运行时间结果对比

通过实验二可以发现,所有算法挖掘的模式个数和运行时间都随着通配符长度的增加而增加。I-OFMI 算法挖掘的模式个数较多,但算法运行时间消耗巨大。FOFM 算法在通配符长度较大时仍能保持一定的完备性,运行时间小于 I-OFMI 算法,如序列 2 在通配符长度为 8 时 FOFM 算法耗时 7.6 s, I-OFMI 算法耗时 146.6 s。相比 OFMI 算法,FOFM 算法挖掘的模式个数比 OFMI 算法更多,运行时间只有模式数量差距较大时才会大于 OFMI 算法,其他情况下均优于 OFMI 算法。

通过以上实验可以发现,FOFM 算法的运行效率明显优于 OFMI 算法及 I-OFMI 算法,在通配符长度较小时,FOFM 算法挖掘模式个数与 I-OFMI 算法差距较小,相比 OFMI 算法挖掘模式个数更多,这主要是因为 FOFM 算法在模式连接时选择保留模式的尾序列,避免了重复扫描序列和列举模式中事件的可能位置。

4 结束语

文中提出了一种新的带间隔约束的序列模式挖掘算法 FOFM,算法记录了模式连接形成的候选模式最后事件的可能位置,并采用回溯策略扫描模式的前缀模式以检查支持度。实验结果表明,FOFM 算法是一种快速的带通配符和 One-Off 条件的单序列模式挖掘算法,可以有效地挖掘满足 One-Off 条件的带间隔约束的序列模式,在一定通配符长度下其时间效率较高,同时保证了较高的完备性。但由于 FOFM 算法仅从后向前选取事件,在通配符长度较大时算法的完备性较差,有待进一步完善。

参考文献:

- [1] 张 晓.中国水污染趋势与治理制度[J].中国软科学,2014(10):11-24.
- [2] 马乐宽,王金南,王 东.国家水污染防治“十二五”战略与政策框架[J].中国环境科学,2013,33(2):377-383.
- [3] PEI Jian, HAN Jiawei, MORTAZAVI-ASL B, et al. PrefixSpan: mining sequential patterns efficiently by prefix-projected pattern growth[C]//Proceedings of the 17th international conference on data engineering. [s.l.]: IEEE, 2001: 215-224.
- [4] ZAKI M J. Sequence mining in categorical domains: incorporating constraints[M]//Proceedings of the ninth international conference on information and knowledge management. McLean, Virginia, USA: IEEE, 2001: 422-429.
- [5] JI Xiaonan, BAILEY J, DONG Guozhu. Mining minimal distinguishing subsequence patterns with gap constraints[C]//Proceedings of the fifth IEEE international conference on data mining. [s.l.]: IEEE, 2005: 194-201.
- [6] LI Chun, WANG Jianyong. Efficiently mining closed subsequences with gap constraints[C]//SIAM international conference on data mining. Atlanta, Georgia, USA: IEEE, 2008: 313-322.
- [7] ZHANG Minghua, KAO Ben, CHEUNG D W, et al. Mining periodic patterns with gap requirement from sequences[J]. ACM Transactions on Knowledge Discovery from Data, 2007, 1(2): 7.
- [8] ZHU Xingquan, WU Xindong. Mining complex patterns across sequences with gap requirements[C]//Proceedings of the 20th international joint conference on artificial intelligence. Hyderabad, India: Morgan Kaufmann Publishers Inc., 2007: 2934-2940.
- [9] KEMMAR A, LEBBAH Y, LOUDNI S, et al. Prefix-projection global constraint and top-k approach for sequential pattern mining[J]. Constraints, 2017, 22(2): 265-306.
- [10] HUYNH B, VO B, SNASEL V. An efficient method for mining frequent sequential patterns using multi-core processors[J]. Applied Intelligence, 2017, 46(3): 703-716.
- [11] BANDARU S, NG A H C, DEB K. Data mining methods for knowledge discovery in multi-objective optimization: part B - new developments and applications[J]. Expert Systems with Applications, 2017, 70: 119-138.
- [12] HE Yu, WU Xindong, ZHU Xingquan, et al. Mining frequent patterns with wildcards from biological sequences[C]//IEEE international conference on information reuse and integration. [s.l.]: IEEE, 2007: 329-334.
- [13] XIE Fei, WU Xindong, HU Xuegang, et al. Sequential pattern mining with wildcards[C]//IEEE international conference on tools with artificial intelligence. Arras, France: IEEE, 2010: 241-247.
- [14] 吴信东, 谢 飞, 黄咏明, 等. 带通配符和 One-Off 条件的序列模式挖掘[J]. 软件学报, 2013, 24(8): 1804-1815.
- [15] 刘祥明. 水质时间序列数据挖掘及其应用集成研究[D]. 重庆: 重庆大学, 2011.
- [16] 张永勇, 花瑞祥, 夏 瑞. 气候变化对淮河流域水量水质影响分析[J]. 自然资源学报, 2017, 32(1): 114-126.
- [17] 方晓波, 骆林平, 李 松, 等. 钱塘江兰溪段地表水质季节变化特征及源解析[J]. 环境科学学报, 2013, 33(7): 1980-1988.
- [18] 梁中耀, 刘 永, 盛 虎, 等. 滇池水质时间序列变化趋势识别及特征分析[J]. 环境科学学报, 2014, 34(3): 754-762.