DOI:10.3876/j.issn.1000-1980.2020.06.004

基于数据挖掘的太湖蓝藻生长水环境关键因子研究

李 蓓,李勇涛,蔡 梅

(太湖流域管理局水利发展研究中心,上海 200434)

摘要:为探究太湖水体富营养化演变机理,识别影响太湖富营养化及蓝藻生长的水环境关键因子,对2006—2018年太湖多源监测序列数据开展数据准备和数据清洗,采用*K*-means均值聚类方法获取离散的布尔型关联规则挖掘候选数据集,构建基于Apriori算法的太湖水环境关键因子关联规则挖掘模型,对影响太湖水体富营养化的水环境关键因子进行识别。结果表明:表征太湖富营养化程度的Chl-a质量浓度与TP质量浓度、NH_3-N质量浓度、pH和COD_{Mn}质量浓度均呈现不同强度的关联性,其中Chl-a质量浓度在0~18.36 mg/m^{3}区间内与TP质量浓度在0~0.045 mg/L区间内关联性最强;从水环境治理角度看,若将太湖TP质量浓度控制在0.045 mg/L以下,则全湖Chl-a质量浓度小于18.36 mg/m^{3}的概率最大,可以有效控制蓝藻数量总体处于较少状态,避免太湖蓝藻水华大规模暴发。

关键词:蓝藻;关联规则挖掘;富营养化;聚类;关键因子;太湖

中图分类号:X524 **文献标志码**:A **文章编号**:1000-1980(2020)06-0506-08

Research on key factors of water environment for *cyanobacteria* growth in Taihu Lake based on data mining

LI Bei, LI Yongtao, CAI Mei

(*Water Conservancy Development Research Center of Taihu Basin Authority of Ministry of Water Resources, Shanghai* 200434, *China*)

Abstract: In order to explore the eutrophication evolution mechanism and identify the key factors of water environment that affect the eutrophication and *cyanobacteria* growth of Taihu Lake, the data preparation and data cleaning for the data of the multi-source monitoring sequence of Taihu Lake from 2006 to 2018 were carried out. The *K*-means clustering method was used to obtain the discrete Boolean association rule for the mining of candidate data sets, and a mining model of association rule for key factors of Taihu Lake was constructed based on the Apriori algorithm, from which the key factors of water environment that affect the eutrophication of Taihu Lake were identified. The results showed that the mass concentration of chlorophyll a, which characterizes the degree of eutrophication in Taihu Lake, has different degrees of correlation with total phosphorus, ammonia nitrogen, pH and permanganate index. Among them, the mass concentration of chlorophyll a in the range of 0-18.36 mg/m^3 has the strongest correlation with the total phosphorus in the range of 0-0.045 mg/L. From the perspective of water environment management, if the total phosphorus concentration in Taihu Lake is controlled below 0.045 mg/L, the probability that the mass concentration of chlorophyll a in the whole lake is below 18.36 mg/m^3 would be the highest, which can effectively control the number of *cyanobacteria* in the overall state of less and further avoid the large-scale outbreak of *cyanobacteria* bloom in Taihu Lake.

Key words: *cyanobacteria*; association rules mining; eutrophication; clustering; key factor; Taihu Lake

---

基金项目:国家重点研发计划(2018YFC0407903)

作者简介:李蓓(1972—),女,教授级高级工程师,博士,主要从事水资源配置、管理与保护研究。E-mail:libei@tba.gov.cn

通信作者:李勇涛,工程师。E-mail:liyongtao@tba.gov.cn

引用本文:李蓓,李勇涛,蔡梅.基于数据挖掘的太湖蓝藻生长水环境关键因子研究[J].河海大学学报(自然科学版),2020,48(6):506-513.
LI Bei, LI Yongtao, CAI Mei. Research on key factors of water environment for *cyanobacteria* growth in Taihu Lake based on data mining [J]. Journal of Hohai University(Natural Sciences),2020,48(6):506-513.

湖泊富营养化是世界性水环境问题,20 世纪 80 年代以来,受环太湖城市经济社会发展及人类活动影响,太湖水体富营养化问题日趋严重[1]。由于湖泊面积较大且环太湖出入湖河道众多,入湖污染负荷量远超湖体自净能力,2019 年环太湖河流 TP、TN 入湖污染负荷分别为太湖纳污能力的 3.6 倍和 4.28 倍。受诸多自然及人为因素影响,湖泊水环境复杂多变,富营养化治理难度较大。

2007 年无锡由于蓝藻暴发引发饮用水危机,蓝藻事件引起社会广泛关注。2008 年 5 月,国务院批复实施《太湖流域水环境综合治理总体方案》[2]。2013 年,为巩固治理成果,提升治理水平,国家发展和改革委员会牵头,会同有关部门深入调查研究,编制形成《太湖流域水环境综合治理总体方案(2013 年修编)》[3](以下简称《总体方案修编》),作为未来一个时期指导太湖流域水环境综合治理的基本依据。随着太湖流域大规模水环境综合协同治理工作的持续推进,太湖富营养化与蓝藻水华的发展势头得到初步遏制,但 2016 年以来,蓝藻水华仍呈现扩大趋势[4]。为进一步深入探究太湖水体富营养化影响因素及发生机理,郝晨林等[5]通过低通时序滤波轨线法识别出 2006 年、2011 年为太湖营养过程轨线转折点,且气温变化可能是导致太湖富营养化加剧的主要原因。部分研究[6-9]表明,氮、磷是造成太湖富营养化的关键因子,还有大量相关研究则指出蓝藻水华暴发受多种因素影响[10-12],但目前仍未形成共识。

长期以来,太湖流域水环境治理积累了大量的多源监测数据,通过数据挖掘来揭示数据间内在联系、趋势和模式已经成为水环境数据科学领域新的研究手段。“数据挖掘”概念最早由 Fayyad 等[13]提出,目前广泛应用于教育、生物、金融、医学、电子商务等领域。在水环境领域,曹钦[14]将基于约束的序列模式挖掘算法应用到三峡库区水环境安全预警决策中,曹敏杰[15]基于时空影响域和上下文约束的海洋生态环境关联规则挖掘分析研究,设计了基于关联规则的海洋生态环境时空挖掘分析框架,用于对赤潮现象进行分析与预测预警,均取得了良好效果。目前国内外采用数据挖掘手段开展水环境关键因子识别研究的成果较少,在水环境领域内的关联规则挖掘研究及应用还不够深入,本文拟利用数据挖掘技术,另辟蹊径,基于不断更新的、迅速发展的系列外部多源数据,采用关联规则挖掘算法,识别太湖蓝藻生长水环境关键因子,为太湖水环境治理提供依据。

1　研究区概况

太湖是中国第三大淡水湖泊,位于北纬 30°55′40″~31°32′48″、东经 119°52′32″~120°36′10″之间。太湖是流域防洪及水资源调配中心和流域内最重要的水源地,苏浙两省环湖大中城市均以太湖为主要饮用水水源地,也是上海市、浙江嘉兴市等下游城市的重要供水水源,同时,太湖通过环湖河道出湖水量为周边及下游地区提供工农业生产、生活用水。按照自然条件和湖区水质研究需要,一般将全湖划分为东太湖、东部沿岸区(含胥湖)、贡湖、梅梁湖、竺山湖、西部沿岸区、南部沿岸区和湖心区。太湖(不含五里湖)目前共布设 31 个监测点,分设在 8 个湖区,分别在梅梁湖 5 个、竺山湖2 个、贡湖 4 个、东太湖 3 个、湖心区 6 个、西部沿岸区 2 个、东部沿岸区 4 个和南部沿岸区 5 个(图 1)。

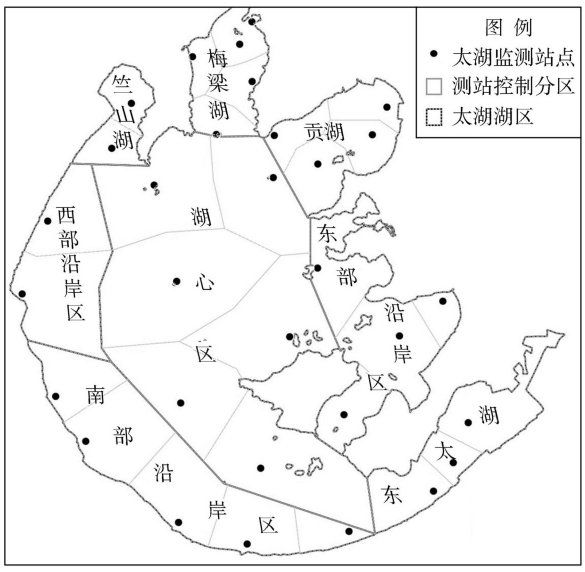

图 1　太湖监测站点及分区示意图

Fig. 1　Distribution of lake sub-regions and monitoring sites of Taihu Lake

2007—2018 年太湖蓝藻状况监测数据显示,2007 年以来,太湖藻类及蓝藻总体呈扩张趋势,至 2017 年达到顶峰,2018 年两项指标较 2017 年均有所缩减(图 2),其中:2007 年蓝藻占全湖藻类比重最低,占比为 31.21%;2011 年后蓝藻在全湖藻类中占比均不低于 80%,且呈现逐年上升趋势;2017 年太湖蓝藻占全湖藻类比重达到 96.23%,为近 12 年来最高。研究太湖蓝藻生长机理对太湖水体富营养化治理意义重大。

2　关联规则挖掘原理及方法

2.1　原理

关联规则挖掘是数据挖掘技术中重要的研究内容,

用于从海量数据中提取有用的关联信息,发掘数据背后隐藏的关联性[16]。Agrawal 等[17]最早提出了基于频繁项集的经典关联规则 Apriori 算法,其优点为适合事务数据库的关联规则挖掘,主要思想是利用一个逐层搜索的迭代方法,对数据库中所有事务数据项进行扫描来完成频繁项集的挖掘。其中 Apriori 算法采用了两个重要的性质:

性质 1 频繁项集的所有非空子集必为频繁项集。

性质 2 非频繁项集的超集一定是非频繁的。

设 $I=\{i_1,i_2,\cdots,i_n\}$ 是项的集合,事务数据集是由一系列具有唯一标志的事务组成,且每个事务均为 I 中的子集。设 X、Y 均为事务数据集中的子集,且互不相交:

$$X \subseteq I \qquad Y \subseteq I \qquad X \cap Y = \varnothing \tag{1}$$

式中:I——事务数据项集合;X、Y——事务数据集合,皆为 I 的子集,若两者有关联则可表示为 $X \rightarrow Y$,X、Y 分别为关联规则的前件和后件;$\varnothing$——空集。

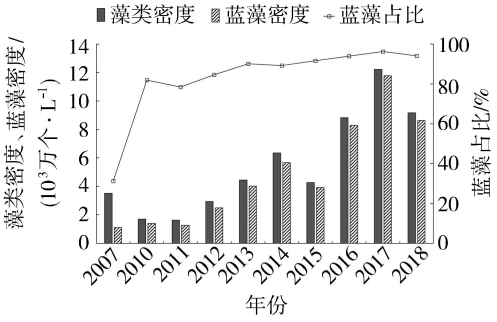

图 2 2007—2018 年太湖蓝藻状况

Fig. 2 Conditions of *cyanobacteria* in Taihu Lake from 2007 to 2018

2.1.1 支持度

关联规则 $X \rightarrow Y$ 支持度是指事务数据集中包含项目集 X 和 Y 的百分比,支持度描述了 X 和 Y 同时出现在事务中的概率,支持度越大则表示关联规则越重要。在挖掘过程中,通过设置最小支持度阈值,将支持度不满足要求的关联规则剪枝以提高算法效率。

$$\text{support}(X \rightarrow Y) = P(X \cap Y) \tag{2}$$

2.1.2 置信度

关联规则 $X \rightarrow Y$ 置信度是指在事务数据集中包含 $X \cup Y$ 的事务与包含 X 的事务之比,置信度越高表明该规则的可靠度越高。在挖掘过程中,通过设置最小置信度阈值,将置信度不满足要求的关联规则剔除,提高挖掘成果的可靠性。

$$\text{confidence}(X \rightarrow Y) = \frac{\text{support}(X \cup Y)}{\text{support}(X)} \tag{3}$$

2.1.3 算法选取

Apriori 算法的剪枝方法可大幅度减少候选项集,提高挖掘效率,国内外许多研究人员针对不同领域问题对 Apriori 算法进行了大量研究与改进[18-20]。常用关联规则挖掘算法主要特点见表 1。

表 1 常用关联规则算法及特点

Table 1 Common association rule algorithm and its characteristics

算法名称	算法描述
Apriori	关联规则最常用也是最经典的挖掘频繁项集的算法,通过连接产生候选项及其支持度,然后通过剪枝生成频繁项集
FP-growth	Apriori 算法的改进算法,提出了不产生候选频繁项集改用构造 FP-Tree 的方法
DHP	引入哈希树概念,提高算法效率
正交链表改进的 Apriori 算法	针对需多次扫描数据库需求,将数据库转化为关系矩阵,并用正交链表进行存储

经过对比分析,本研究主要针对可能与太湖蓝藻生长相关的多源外部数据进行挖掘,从候选数据体量及算法稳定性方面综合评价,拟采用综合性能较稳定的 Apriori 算法开展关键因子关联规则挖掘研究。

2.2 研究方法

浮游植物的大量生长是湖泊富营养化的重要现象,通常用 Chl-a 来表征湖泊富营养化程度[21],采用除 Chl-a 以外的多源水环境监测数据,通过数据清洗与数据离散后形成关联规则挖掘候选数据集。通过 Apriori 算法开展满足最小支持度和最小置信度的因子与 Chl-a 关联规则挖掘,并从中识别出与 Chl-a 关联性最强的因子,作为影响太湖水体富营养化程度的关键因子。

基于构建的数据集,挖掘关联规则的问题可以转换为寻找满足最小支持度和最小置信度阈值的强关联规则过程,分为两步:(a)生成频繁项集,由频繁项集生成满足最小支持度阈值的项集;(b)生成强关联规则,找出频繁项集中大于或等于最小置信度阈值的关联规则。关联规则挖掘流程见图 3。

3　太湖水环境关键因子关联规则挖掘模型构建

3.1　数据准备

为充分挖掘候选数据的内在联系,需尽可能多地考虑可能造成太湖蓝藻生长的因素,但候选数据的监测频次及样本体量应相当,以便对同一时间尺度内的相关因子进行横向关联挖掘。基于上述考虑,本文采用 2006—2018 年共 13 年太湖湖区 31 个监测站点逐月的水温、pH、DO、浊度、COD_{Mn}、TN、NH_3-N、TP 和 Chl-a 共 9 类指标序列数据作为关键因子关联规则挖掘研究的数据基础。每个测站同一时间的监测数据作为一条记录,数据标准化后存入数据库。经统计,累计获取监测记录 4 703 条。

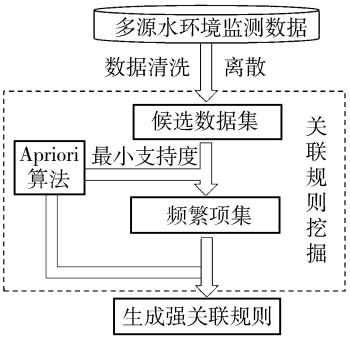

图 3　关联规则挖掘流程

Fig. 3　Flowchart of association rule mining

3.2　数据清洗与离散

3.2.1　异常数据剔除

实测数据往往存在数据缺失和数据异常情况,通过数据库管理技术自动识别并剔除包含空值及异常符号的记录,针对数据异常及极端情况,为保证数据离散及挖掘成果的合理性,此处采用拉依达准则法对候选数据中存在粗大误差的数据进行剔除,经数据清洗,获取有效记录 3 740 条。

3.2.2 数据离散

影响太湖富营养化的诸多环境因子均为通过实际监测获取的和时间相关的非连续数据点,需将多值关联规则问题转化为布尔型关联规则问题。考虑到监测数据按照 GB 3838—2002《地表水环境质量标准》[22] 规定的水质类别区间划分属于评价体系,由人为划定分级,较难体现数据本身的分布特性且易对挖掘结果产生干扰,因此采用无监督学习中的 *K*-Means 均值聚类算法(其中 *K* 为聚类簇数)。

结合数据本身分布特征进行聚类,参考水质指标分类,本次聚类各项按照不同数量簇别进行聚类,考虑到候选数据样本量,簇别太大虽可对候选数据离散级别进行细分,但较难挖掘出有效强关联规则,经综合考虑,拟采用 4、6、8 簇进行聚类,经聚类离散后形成候选数据集,分析不同簇别聚类条件对挖掘成果的影响(表 2)。

表 2　*K*-means 聚类结果

Table 2　*K*-means clustering results

簇别	水温/℃	pH	ρ(DO)/(mg·L^{-1})	浊度/cm	ρ(COD_{Mn})/(mg·L^{-1})	ρ(TN)/(mg·L^{-1})	ρ(NH_3-N)/(mg·L^{-1})	ρ(TP)/(mg·L^{-1})	ρ(Chl-a)/(mg·m^{-3})
4 簇	(0,11.02]	(0,7.93]	(0,7.72]	(0,13.11]	(0,3.302]	(0,1.279]	(0,0.145]	(0,0.045]	(0,18.36]
	(11.02,18.59]	(7.93,8.25]	(7.72,9.45]	(13.11,34.2]	(3.302,4.017]	(1.279,2.097]	(0.145,0.364]	(0.045,0.074]	(18.36,39.11]
	(18.59,25.41]	(8.25,8.59]	(9.45,11.26]	(34.2,60.45]	(4.017,4.798]	(2.097,3.103]	(0.364,0.795]	(0.074,0.112]	(39.11,79.2]
	>25.41	>8.59	>11.26	>60.45	>4.798	>3.103	>0.795	>0.112	>79.2
6 簇	(0,7.1]	(0,7.78]	(0,6.43]	(0,12.07]	(0,3.001]	(0,0.958]	(0,0.098]	(0,0.037]	(0,13.41]
	(7.1,12.64]	(7.78,8.06]	(6.43,7.99]	(12.07,29.27]	(3.001,3.524]	(0.958,1.432]	(0.098,0.201]	(0.037,0.057]	(13.41,23.67]
	(12.64,18.62]	(8.06,8.28]	(7.99,9.23]	(29.27,42.04]	(3.524,3.996]	(1.432,1.969]	(0.201,0.356]	(0.057,0.078]	(23.67,37.62]
	(18.62,23.35]	(8.28,8.5]	(9.23,10.59]	(42.04,59.98]	(3.996,4.5]	(1.969,2.589]	(0.356,0.618]	(0.078,0.102]	(37.62,58.32]
	(23.35,27.65]	(8.5,8.79]	(10.59,11.96]	(59.98,87.54]	(4.5,5.089]	(2.589,3.382]	(0.618,1.05]	(0.102,0.132]	(58.32,94.16]
	>27.65	>8.79	>11.96	>87.54	>5.089	>3.382	>1.05	>0.132	>94.16
8 簇	(0,5.39]	(0,7.5]	(0,5.74]	(0,10.38]	(0,2.739]	(0,0.853]	(0,0.077]	(0,0.032]	(0,10.86]
	(5.39,8.86]	(7.5,7.82]	(5.74,7.46]	(10.38,24.38]	(2.739,3.186]	(0.853,1.204]	(0.077,0.144]	(0.032,0.046]	(10.86,18.1]
	(8.86,12.68]	(7.82,8.03]	(7.46,8.44]	(24.38,32.97]	(3.186,3.568]	(1.204,1.579]	(0.144,0.232]	(0.046,0.061]	(18.1,27.15]
	(12.68,16.38]	(8.03,8.21]	(8.44,9.4]	(32.97,43.91]	(3.568,3.935]	(1.579,2.001]	(0.232,0.358]	(0.061,0.077]	(27.15,39.4]
	(16.38,19.79]	(8.21,8.38]	(9.4,10.44]	(43.91,60.01]	(3.935,4.321]	(2.001,2.487]	(0.358,0.554]	(0.077,0.094]	(39.4,56.57]
	(19.79,23.59]	(8.38,8.58]	(10.44,11.43]	(60.01,79.04]	(4.321,4.748]	(2.487,3.048]	(0.554,0.825]	(0.094,0.115]	(56.57,80.96]
	(23.59,27.72]	(8.58,8.84]	(11.43,12.44]	(79.04,101.15]	(4.748,5.233]	(3.048,3.693]	(0.825,1.215]	(0.115,0.144]	(80.96,114.82]
	>27.72	>8.84	>12.44	>101.15	>5.233	>3.693	>1.215	>0.144	>114.82

按照表 2 聚类离散成果,分别将水温、pH、DO 质量浓度、浊度、COD_{Mn}质量浓度、TN 质量浓度、NH_3-N 质量浓度、TP 质量浓度和 Chl-a 质量浓度按照 4 簇、6 簇、8 簇离散形成 3 套布尔型关联规则挖掘候选数据集。

3.3　关联规则挖掘成果

对候选数据集进行 Apriori 关联规则挖掘,针对不同簇数据集方案,选定适合的最小支持度和最小置信

度,获取关联规则挖掘结果见表3。

表3 关联规则挖掘成果

Table 3 Results of association rule mining

簇别	关联规则	支持度	置信度
$K=4$(最小支持度为0.25,最小置信度为0.7)	$0<\rho(TP)\leqslant 0.045$ mg/L→$0<\rho(Chl\text{-}a)\leqslant 18.36$ mg/m^3	0.291	0.866
	$0<\rho(TN)\leqslant 1.279$ mg/L→$0<\rho(NH_3\text{-}N)\leqslant 0.145$ mg/L	0.316	0.825
	$0<\rho(TP)\leqslant 0.045$ mg/L→$0<\rho(NH_3\text{-}N)\leqslant 0.145$ mg/L	0.27	0.804
	$0<\rho(DO)\leqslant 7.72$ mg/L→$0<\rho(NH_3\text{-}N)\leqslant 0.145$ mg/L	0.277	0.788
	$8.25<pH\leqslant 8.59$→$0<\rho(NH_3\text{-}N)\leqslant 0.145$ mg/L	0.286	0.771
	$7.93<pH\leqslant 8.25$→$0<\rho(Chl\text{-}a)\leqslant 18.36$ mg/m^3	0.259	0.76
	3.302 mg/L$<\rho(COD_{Mn})\leqslant 4.017$ mg/L→$0<\rho(NH_3\text{-}N)\leqslant 0.145$ mg/L	0.254	0.751
$K=6$(最小支持度为0.1,最小置信度为0.65)	$0<\rho(TP)\leqslant 0.037$ mg/L→$0<\rho(Chl\text{-}a)\leqslant 13.41$ mg/m^3	0.181	0.773
	$0<\rho(NH_3\text{-}N)\leqslant 0.098$ mg/L→$0<\rho(TP)\leqslant 0.037$ mg/L→$0<\rho(Chl\text{-}a)\leqslant 13.41$ mg/m^3	0.109	0.77
	$8.5<pH\leqslant 8.79$→$0<\rho(NH_3\text{-}N)\leqslant 0.098$ mg/L	0.111	0.694
	$0<\rho(TN)\leqslant 0.958$ mg/L→$0<\rho(NH_3\text{-}N)\leqslant 0.098$ mg/L	0.141	0.664
	3.001 mg/L$<\rho(COD_{Mn})\leqslant 3.524$ mg/L→$0<\rho(Chl\text{-}a)\leqslant 13.41$ mg/m^3	0.121	0.659
$K=8$(最小支持度为0.1,最小置信度为0.65)	$0<\rho(TP)\leqslant 0.032$ mg/L→$0<\rho(Chl\text{-}a)\leqslant 10.86$ mg/m^3	0.114	0.728

从挖掘成果可知,相同样本记录条件下,离散簇类越多,获得强关联规则所需支持度和置信度阈值越小;在支持度和置信度阈值相同条件下,离散簇类越多,获取的强关联规则越少。

4 成果分析

4.1 太湖水环境关键因子

太湖蓝藻生长是一个复杂、非线性的生态过程,涉及物理、化学等多方面影响因素,一些研究者[23-24]通过分析少数几个或某一类水环境因子与Chl-a质量浓度或蓝藻密度等指标间的关联关系,且存在数据量偏少的情况,可能导致结果存在片面性。从不同簇别离散的候选数据集挖掘成果可知,目前太湖水环境状况条件下,TP是影响太湖蓝藻生长的主要关键因子,此外NH_3-N、pH和COD_{Mn}也与Chl-a存在不同程度的关联关系,研究结论与文献[25-28]总体一致,表明了基于数据挖掘研究方法的有效性。

4.2 关联程度分析

Apriori算法明确了只有在支持度和置信度均较高的情况下该关联规则才属于强关联规则。从关联程度分布情况来看,如图4所示,气泡直径大小表明了该关联规则的强弱,当簇别K越大时,挖掘结果的关联性越小;支持度和置信度阈值设置越小,获得的强关联规则越多,且规则的可靠性越强。

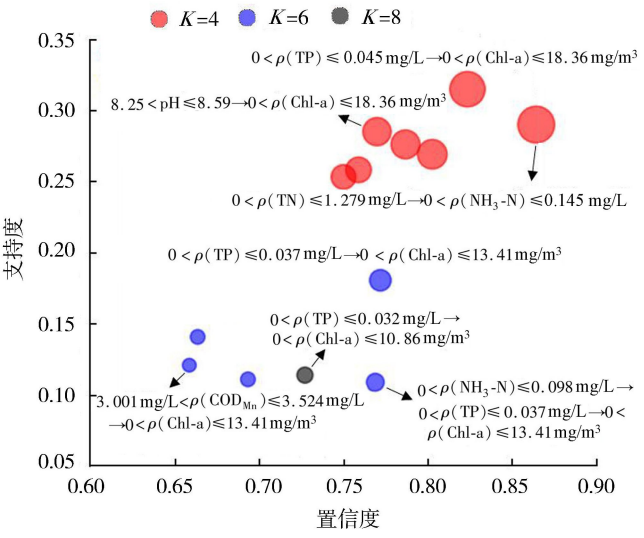

图4 挖掘结果关联程度分布

Fig. 4 Distribution of relevance degree of mining results

通过调整最小支持度和最小置信度等表征关联程度的参数的阈值,筛选获取候选数据集中蕴含的强关联规则。挖掘结果显示,Chl-a质量浓度与TP质量浓度、NH_3-N质量浓度、pH和COD_{Mn}质量浓度均呈现不同强度的关联性,按照关联性强度排序为:TP>pH> NH_3-N >COD_{Mn},其中:当$K=4$时,Chl-a质量浓度在0~18.36 mg/m^{3}区间内与TP质量浓度在0~0.045 mg/L区间内关联性最强,其支持度为0.291,置信度为0.866,与pH在7.93~8.25区间内关联性较强;当$K=6$时,Chl-a质量浓度在0~13.41 mg/m^{3}区间内与TP质量浓度在0~0.037 mg/L区间内关联性最强,其支持度为0.181,置信度为0.773,与COD_{Mn}质量浓度在3.001~3.524 mg/L和NH_3-N质量浓度在0~0.098 mg/L区间内关联性较强;当$K=8$时,Chl-a质量浓度在0~10.86 mg/m^{3}区间内与TP质量

浓度在 0 ~ 0.032 mg/L 区间内关联性最强,其支持度为 0.114,置信度为 0.728。

在关联规则挖掘研究中,合理设置关联程度参数能够加快算法计算效率,若最小支持度及最小置信度阈值设置不合理则较难获得理想的强关联规则,因此,在关联程度参数设置上需要首先设定最小支持度阈值,并结合挖掘成果调整最小置信度阈值,直至挖掘成果合理。

4.3　因子敏感区间

经分析,2006—2018 年水体监测数据关联规则挖掘成果中 TP 质量浓度处于 0 ~ 0.045 mg/L 区间内对 Chl-a 质量浓度在 0 ~ 18.36 mg/m^3 区间内敏感度最高。实测资料显示[29-31],2010—2017 年太湖北部湖泛易发区 Chl-a 质量浓度从 2010 年最低值 19.2 mg/m^3 呈逐年上升,至 2017 年达到最高,达到质量浓度为 56.6 mg/m^3,期间湖泛发生次数总体也呈现上升趋势(图 5)。受时空分布影响,全湖 Chl-a 质量浓度较太湖北部 Chl-a 质量浓度偏低,变化趋势与蓝藻密度一致,2010 年为太湖 2010—2019 年来蓝藻数量最少的年份,其中 2015 年和 2018 年各自较前一年略有下降,但总体仍呈现上升趋势。还有研究结论表示,当湖体中蓝藻细胞数量达到 2 000 万个/L 时[32],可被称为蓝藻水华,2010—2019 年太湖蓝藻细胞数量处于 2 000 万个/L 以内的 2010 年及 2011 年 Chl-a 质量浓度均高于 18.36 mg/m^3。因此,从水环境治理角度看,若将太湖 TP 质量浓度控制在 0.045 mg/L 以下,则全湖 Chl-a 质量浓度处于 18.36 mg/m^3 以下的概率最大,可以有效控制蓝藻数量总体处于较少状态,避免太湖蓝藻水华大规模暴发。

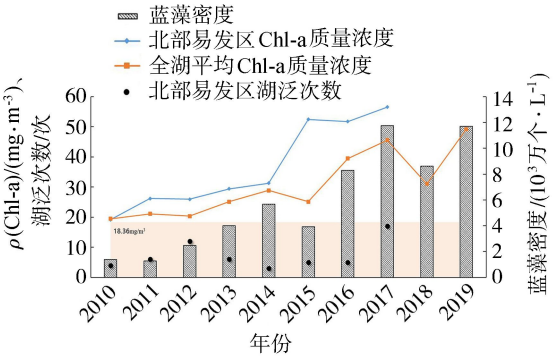

图 5　2010—2019 年太湖北部、全湖 Chl-a 质量浓度及蓝藻密度分布

Fig. 5　Distribution of chlorophyll a concentration and *cyanobacteria* density in northern Taihu Lake and whole lake from 2010 to 2019

对照《总体方案修编》确定的 2020 年控制目标(图 6),截至 2019 年,太湖 NH_3-N、TN 均已达到控制目标,但 COD_{Mn}、TP 尚未达到。由于近年来 TP 质量浓度呈现增长趋势,在流域治理中若进一步加大入湖磷负荷控制,将对改善太湖富营养化状况具有积极作用。太湖 NH_3-N 质量浓度 2019 年已回落至 0.087 mg/L,处于较强关联规则取值范围,COD_{Mn} 在 3.001 ~ 3.524 mg/L 范围内与 Chl-a 质量浓度在 0 ~ 18.36 mg/m^3 区间也呈现一定的关联性,因此,若进一步将太湖 COD_{Mn} 控制在 3.001 ~ 3.524 mg/L 范围内,则全湖 Chl-a 质量浓度小于 18.36 mg/m^3 的概率最大。本成果在《总体方案修编》确定的控制目标基础上,进一步明确了 TP 及

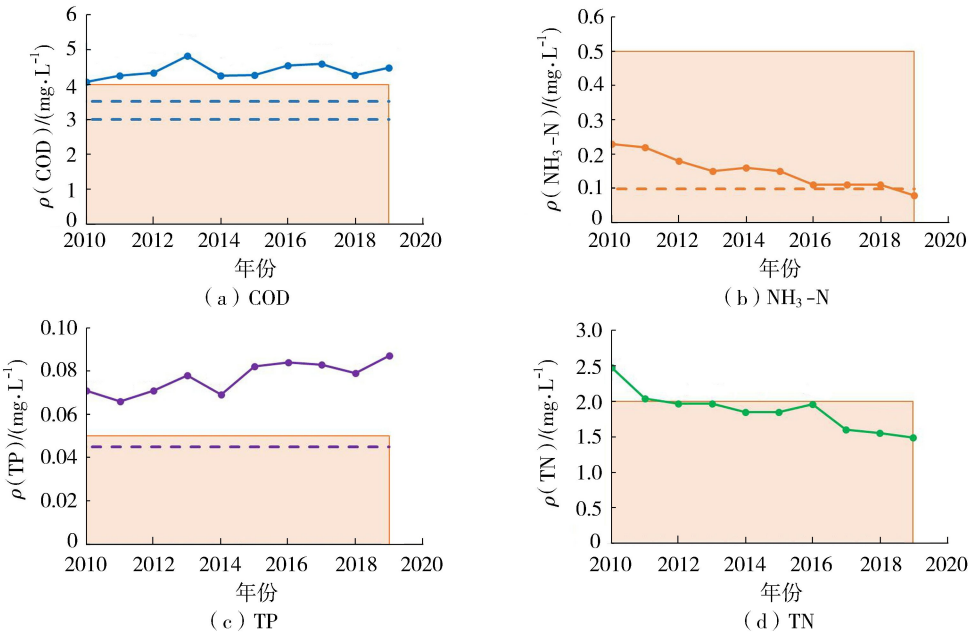

图 6　近年来太湖主要水质指标质量浓度变化情况及因子敏感区间

Fig. 6　Mass concentration changes of main water quality indicators and factor sensitive interval in Taihu Lake

COD_{Mn}等指标控制的具体范围,可为下阶段太湖水环境治理控制目标研究提供技术支撑。

5 结 语

通过构建基于 Apriori 算法的太湖水环境关键因子关联规则挖掘模型,对影响太湖水体富营养化的水环境关键因子进行识别。经分析,表征太湖富营养化程度的 Chl-a 质量浓度与 TP 质量浓度、NH_3-N 质量浓度、pH 和 COD_{Mn}质量浓度均呈现不同强度的关联性,关联性强度排序为 TP>pH> NH_3-N > COD_{Mn}。

TP 质量浓度处于 0 ~ 0.045 mg/L 区间内对 Chl-a 质量浓度在 0 ~ 18.36 mg/m^3区间内最为敏感。从水环境治理角度看,若将太湖 TP 质量浓度控制在 0.045 mg/L 以下,则全湖 Chl-a 质量浓度小于 18.36 mg/m^3的概率最大,可以有效控制蓝藻数量总体处于较少状态,避免太湖蓝藻水华大规模暴发。本研究可为下阶段太湖水环境治理控制目标研究提供技术支撑。

参考文献:

[1] 朱广伟. 太湖富营养化现状及原因分析[J]. 湖泊科学,2008,20(1):21-26. (ZHU Guangwei. Eutrophic status and causing factors for a large, shallow and subtropical Lake Taihu, China[J]. Journal of Lake Sciences, 2008,20(1):21-26. (in Chinese))

[2] 吕振霖. 太湖水环境综合治理的实践与思考[J]. 河海大学学报(自然科学版),2012,40(2):123-128. (LYU Zhenlin. Practice and thoughts on comprehensive treatment of water pollution in Taihu Lake[J]. Journal of Hohai University(Natural Sciences),2012,40(2):123-128. (in Chinese))

[3] 中华人民共和国国家发展和改革委员会. 太湖流域水环境综合治理总体方案(2013 年修编)[R]. 北京:中华人民共和国国家发展和改革委员会,2013.

[4] 秦伯强. 浅水湖泊湖沼学与太湖富营养化控制研究[J]. 湖泊科学,2020,32(5):1229-1243. (QIN Boqiang. Shallow lake limnology and control of eutrophication in Lake Taihu[J]. Journal of Lake Sciences, 2020,32(5):1229-1243. (in Chinese))

[5] 郝晨林,邓义祥,富国,等. 低通时序滤波轨线法在太湖营养过程识别中的应用[J/OL]. [2020-09-30](2020-09-19). https://doi.org/10.13198/j.issn.1001-6929.2020.10.05.

[6] 杨洋,刘其根,胡忠军,等. 太湖流域沉积物碳氮磷分布与污染评价[J]. 环境科学学报,2014,34(12):3057-3064. (YANG Yang, LIU Qigen, HU Zhongjun, et al. Spatial distribution of sediment carbon, nitrogen and phosphorus and pollution evaluation of sediment in Taihu Lake[J]. Acta Scientiae Circumstantiae, 2014,34(12):3057-3064. (in Chinese))

[7] 林泽新. 太湖流域水环境变化及缘由分析[J]. 湖泊科学,2002,14(2):111-116. (LIN Zexin. Analysis of water environmental change in Taihu watershed[J]. Journal of Lake Sciences, 2002,14(2):111-116. (in Chinese))

[8] 翟淑华,韩涛,陈方. 基于质量平衡的太湖氮、磷自净能力计算[J]. 湖泊科学,2014,26(2):185-190. (ZHAI Shuhua, HAN Tao, CHEN Fang. Self-purification capacity of nitrogen and phosphorus of Lake Taihu on the basis of mass balance[J]. Journal of Lake Sciences, 2014,26(2):185-190. (in Chinese))

[9] 白晓华,胡维平. 太湖水深变化对氮磷浓度和叶绿素 a 浓度的影响[J]. 水科学进展,2006,17(5):727-732. (BAI Xiaohua, HU Weiping. Effect of water depth on concentration of TN, TP and Chla in Taihu Lake, China[J]. Advances in Water Science, 2006,17(5):727-732. (in Chinese))

[10] 毕京博,郑俊,沈玉凤,等. 南太湖入湖口叶绿素 a 时空变化及其与环境因子的关系[J]. 水生态学杂志,2012,33(6):7-13. (BI Jingbo, ZHEN Jun, SHEN Yufeng, et al. Spatial-temporal characteristics of chlorophyll-a concentration and its relationship with environmental factors in the inlets of south Taihu Lake[J]. Journal of Hydroecology, 2012,33(6):7-13. (in Chinese))

[11] 王震,邹华,杨桂军,等. 太湖叶绿素 a 的时空分布特征及其与环境因子的相关关系[J]. 湖泊科学,2014,26(4):567-575. (WANG Zhen, ZOU Hua, YANG Guijun, et al. Spatial-temporal characteristics of chlorophyll-a and its relationship with environmental factors in Lake Taihu[J]. Journal of Lake Sciences, 2014,26(4):567-575. (in Chinese))

[12] 陆志华,李勇涛,钱旭,等. 考虑太湖水质指标的流域骨干工程调度方案[J]. 水资源保护,2018,34(6):44-48. (LU Zhihua, LI Yongtao, QIAN Xu, et al. Scheduling scheme of Taihu Basin key projects considering water quality index of Taihu Lake[J]. Water Resources Protection, 2018,34(6):44-48. (in Chinese))

[13] FAYYAD U M, PIATESKY S G. Advances in data mining and knowledge discovery in databases[M]. Palo Alto: AAAI/MIT Press, 1996.

[14] 曹钦. 序列关联挖掘算法研究及在水环境安全中的应用[D]. 重庆:重庆大学,2008.

[15] 曹敏杰. 浙江近岸海域海洋生态环境时空分析及预测关键技术研究[D]. 杭州:浙江大学,2015.

[16] AGRAWAL R. Mining association rule between sets of items in large database[C] // ACM SIGMOD conference on management of data. New York: Association for Computing Machinery, 1993:1-10.

[17] AGRAWAL R, SRIKANT R. Fast algorithms for mining association rules[C] //Proceedings of international conference on very large databases. Santiago:Association for Computing Machinery, 1994: 487-499.

[18] HAN J, PEI J, YIN Y. Mining frequent patterns without candidate generation[C] //ACM SIGMOD international conference on management of data. New York: Association for Computing Machinery, 2000:1-12.

[19] PARK J S, CHEN M S, YU P S. An effective hash-based algorithm for mining association rules[C] //ACM SIGMOD conference on management of data. New York: Association for Computing Machinery, 1995:175-186.

[20] 赵学健, 孙知信, 袁源,等. 一种正交链表存储的改进 Apriori 算法[J]. 小型微型计算机系统, 2016, 37(10):2291-2295. (ZHAO Xuejian, SUN Zhixin, YUAN Yuan, et al. An improved Apriori algorithm based on orthogonal list storage[J]. Journal of Chinese Computer Systems, 2016, 37(10):2291-2295. (in Chinese))

[21] CAMDEVYREN H, DEMYR N, KANIK A, et al. Use of principal component scores in multiple linear regression models for prediction of chlorophyll-a in reservoirs[J]. Ecological Modelling, 2005,181(4):581-589.

[22] 国家环境保护总局,国家质量监督检验检疫总局. 地表水环境质量标准:GB 3838—2002 [S]. 北京:中国标准出版社, 2002.

[23] 武胜利,刘诚,孙军,等. 卫星遥感太湖蓝藻水华分布及其气象影响要素分析[J]. 气象,2009,35(1):18-23. (WU Shengli, LIU Cheng, SUN Jun, et al. Remote sensing and analysis on meteorological factors of blue algal bloom in Lake Tai[J]. Meteorological Monthly, 2009,35(1):18-23. (in Chinese))

[24] CAO Huansheng, KONG Fanxiang, LUO Liancong, et al. Effects of wind and wind-induced waves on vertical phytoplankton distribution and surface blooms of microcystis aeruginosa in Lake Taihu[J]. Journal of Freshwater Ecology, 2006,21(2):231-238.

[25] 杨晓红,陈江,周李,等. 南太湖入湖口蓝藻水华时空分布规律及相关响应因子分析[J]. 中国环境监测,2011,27(2):92-96. (YANG Xiaohong, CHEN Jiang, ZHOU Li, et al. South Tai Lake's main lake inlet blue-green alga water bloom space and time distribution rule and related response factor analysis[J]. Environmental Monitoring in China,2011,27(2):92-96. (in Chinese))

[26] 黄炜,赵来军. 蓝藻水华相关因素识别、预测与治理[M]. 上海:复旦大学出版社,2015.

[27] 杨柳燕,杨欣妍,任丽曼,等. 太湖蓝藻水华暴发机制与控制对策[J]. 湖泊科学,2019,31(1):18-27. (YANG Liuyan, YANG Xinyan, REN Liman, et al. Mechanism and control strategy of *cyanobacteria*l bloom in Lake Taihu[J]. Journal of Lake Sciences, 2019,31(1):18-27. (in Chinese))

[28] WILHELM S W, FARNSLEY S E, LECLEIR G R. The relationships between nutrients, *cyanobacteria*l toxins and the microbial community in Lake Tai, China[J]. Harmful Algae, 2011, 10(2) :207-215.

[29] 王华,陈华鑫,徐兆安,等. 2010—2017 年太湖总磷浓度变化趋势分析及成因探讨[J]. 湖泊科学,2019,31(4):919-929. (WANG Hua, CHEN Huaxin, XU Zhaoan, et al. Variation trend of total phosphorus and its controlling factors in Lake Taihu, 2010—2017[J]. Journal of Lake Sciences, 2019,31(4):919-929. (in Chinese))

[30] 辛华荣,朱广伟,王雪松,等. 2009—2018 年太湖湖泛强度变化及其影响因素[J/OL]. [2020-10-08](2020-06-05). https://doi.org/10.13227/j.hjkx.202004172.

[31] 刘俊杰,陆隽,朱广伟,等. 2009—2017 年太湖湖泛发生特征及其影响因素[J]. 湖泊科学,2018,30(5):1196-1205. (LIU Junjie, LU Jun, ZHU Guangwei, et al. Occurrence characteristics of black patch events and their influencing factors in Lake Taihu during 2009 and 2017[J]. Journal of Lake Sciences,2018,30(5):1196-1205. (in Chinese))

[32] 孔繁翔,马荣华,高俊峰,等. 太湖蓝藻水华的预防、预测和预警的理论与实践[J]. 湖泊科学,2009,21(3):314-328. (KONG Fanxiang, MA Ronghua, GAO Junfeng, et al. The theory and practice of prevention, forecast and warning on *cyanobacteria* bloom in Lake Taihu[J]. Journal of Lake Sciences,2009,21(3):314-328. (in Chinese))

(收稿日期:2020-10-08　编辑:高建群)