北京大学学报(自然科学版) 第 55 卷 第 2 期 2019 年 3 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 55, No. 2 (Mar. 2019)
doi: 10.13209/j.0479-8023.2019.001

基于模式识别方法的湖泊水质污染特征聚类研究

任婷玉 梁中耀 陈会丽 刘永†

北京大学环境科学与工程学院, 水沙科学教育部重点实验室, 北京 100871; † 通信作者, E-mail: yongliu@pku.edu.cn

摘要 构建耦合自组织映射神经网络(SOFM)和随机森林(RF)的方法, 对中国 63 个湖泊 11 年的 9 种水质指标(5110 条数据)进行模式识别。首先采用 SOFM 对湖泊进行聚类, 以识别污染状况, 然后采用 RF 分析水质指标对湖泊类别的决定效果, 以确定代表性指标。SOFM 的结果显示, 湖泊可以按污染程度分为 3 类。RF 的结果发现, 在分类准确率为 80%时, 根据高锰酸盐指数和叶绿素 *a* 浓度即可判定湖泊污染程度。该方法可从庞杂的数据中识别出反映水体污染特征的水质指标, 为快速认知水体污染状况及选取监测指标提供参考。

关键词 模式识别; 水质污染; 自组织映射神经网络; 随机森林

Clustering of Lake Variables Based on Pattern Recognition Method

REN Tingyu, LIANG Zhongyao, CHEN Huili, LIU Yong†

College of Environmental Science and Engineering, Key Laboratory of Water and Sediment Sciences Ministry of Education, Peking University, Beijing 100871; † Corresponding author, E-mail: yongliu@pku.edu.cn

Abstract The self-organizing feature map (SOFM) and random forest (RF) method were integrated to recognize water quality patterns of nine water quality indicators for 63 lakes in China for 11 years (5110 data). The SOFM was built firstly to cluster lakes to identify the pollution conditions. Then, the RF was used to explore the good-of-fitness of water quality variables on the clustering result and to determine the important water quality indicators. The result of SOFM shows that the lakes can be clustered into three types. And the result of RF shows that permanganate index and chlorophyll *a* can determine the pollution condition when the classification accuracy is 80%. The integrated method can identify the water quality indicators reflecting the pollution conditions from complex data. In practice, the method can be used to determine the pollution conditions and direct the monitoring indicators.

Key words pattern recognition; water pollution; self-organizing feature map; random forest

湖泊水质污染和富营养化是全球性的环境问题[1−3], 给生态系统和人体健康带来严重威胁[4]。近年来, 随着时空精度的提高, 水质监测数据得到大量积累, 从庞杂的数据中识别反映湖泊污染特征的水质指标对于快速认知污染状况具有重要意义[5]。统计是常用的分析方法, 但当数据量较大时, 该方法受限于数据分布的假设和较差的拟合效果。模式识别方法具有运算便捷、结果可靠等优点, 可广泛应用于大量数据的特征识别中, 适用于根据大量监测数据对水质污染状况和主要指标的识别[6−9]。

模式识别是一类采用统计学习算法对大量数据模式特征进行分析的方法, 包括非监督聚类方法和监督分类方法。自组织特征映射网络(self-organizing feature map, SOFM)是一种常用的非监督聚类方法, 具有很强的自组织、自学习、自适应、容错和记忆联想功能, 广泛应用于水质评价中[10−12]。研究表明, SOFM 网络在聚类过程中可有效地避免权重系数的影响, 计算过程简单, 对水质的聚类效果良好, 评估结果可靠[13−14]。随机森林(random forest, RF)是一种常用的监督分类方法, 具有预测准

国家自然科学基金(51779002)资助
收稿日期: 2018–03–16; 修回日期: 2018–07–21; 网络出版日期: 2019–02–28

确率高、对异常值和噪声的容忍度好以及不易出现过拟合等优点[15], 可用于代表性水质指标的识别[16]。研究表明, 与人工神经网络和支持向量机的评价结果相比, RF 方法在分类预测阶段和交叉验证阶段的分类准确率均较高[17]。RF 可解决其他机器学习算法稳健性不足和过学习等问题, 对数据的前提条件要求宽松, 调节参数少, 训练速度快, 在水质评价分析中值得推广[18]。

尽管非监督聚类和监督分类方法在水资源管理中均有较广泛的应用, 然而将两种方法的优势结合起来进行水质污染特征分析的研究尚不多见。为识别湖泊水质的污染特征和代表性水质指标, 本研究提出耦合 SOFM 和 RF 的方法, 对我国 63 个湖泊 2006—2016 年的 9 种水质指标共 5110 条数据进行模式识别, 以期为湖泊水质污染状况的识别和水质监测指标的选取提供参考。

1 材料和方法

1.1 研究对象

本文选择《中国环境状况公报》中的 63 个湖泊(图 1)作为研究对象, 水质数据的时间跨度为 11 年(2006—2016 年), 频次为每月 1 次, 删除部分缺失的数据后, 共有 5110 条监测数据。由于富营养化和有机物污染是我国湖泊面临的主要水环境问题, 因此选取总氮(TN)、氨氮(NH_3-N)、总磷(TP)、叶绿素 *a* (Chla)、化学需氧量(COD)、高锰酸盐指数(COD_{Mn})和生化需氧量(BOD_5)作为水质指标; 同时, 由于一般理化指标中的 pH 和溶解氧(DO)能够综合反映水质的污染状况, 因而也将其纳入分析, 即共对 9 项水质指标进行聚类分析。

1.2 研究思路

本研究采用 SOFM 和 RF 两种模式的识别方法进行数据分析, 其中 SOFM 可将湖泊聚类为不同类别, 将该类别作为 RF 分类的因变量, 从而实现两种模式识别方法的耦合。在进行 SOFM 分析之前, 需对数据进行预处理(图 2)。

1.3 数据处理

首先对水质指标进行 Shapiro-Wilk 正态性检验[19]。结果表明: 在置信水平为 0.05 时, 大部分湖泊的 9 项水质指标均服从偏态分布(表 1)。为避免异常值对聚类结果的影响, 可选用中位数而非平均值来代表各水质指标的集中程度[20–21]。由于各水质指标值的量级存在差异, 模式识别方法对输入变量的量级比较敏感[22], 因此对变量(y)进行归一化(式(1))。

$$x_i = 2 \times \frac{(y_i - y_{\min})}{y_{\max} - y_{\min}} - 1, \tag{1}$$

其中, x_i 和 y_i 分别为变换之后和变换之前水质变量的第 i 个监测数据, $y_{\max}$ 和 $y_{\min}$ 分别为监测数据的最

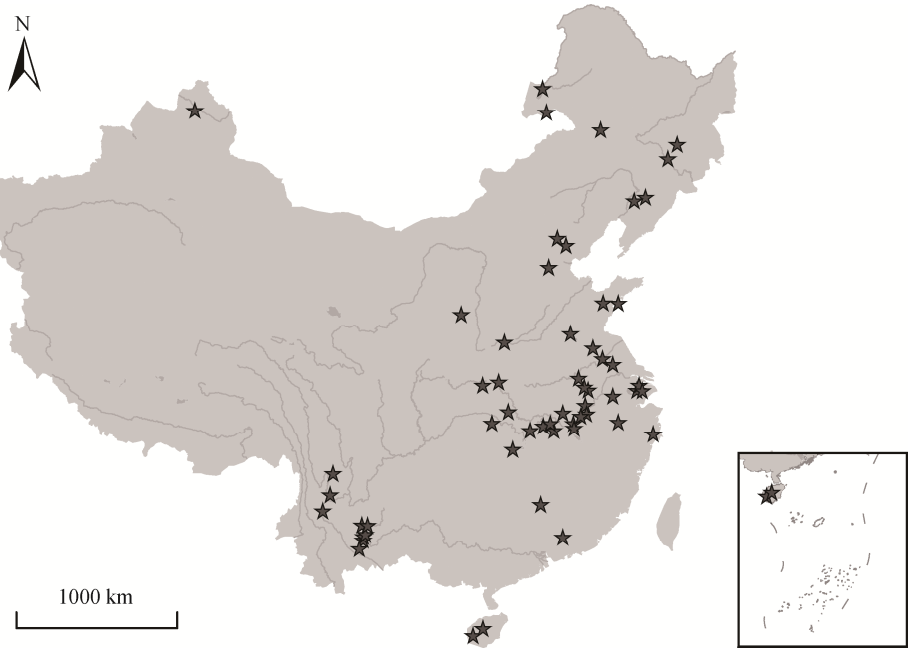


图 1 湖泊的空间分布

Fig. 1 Locations map of lakes

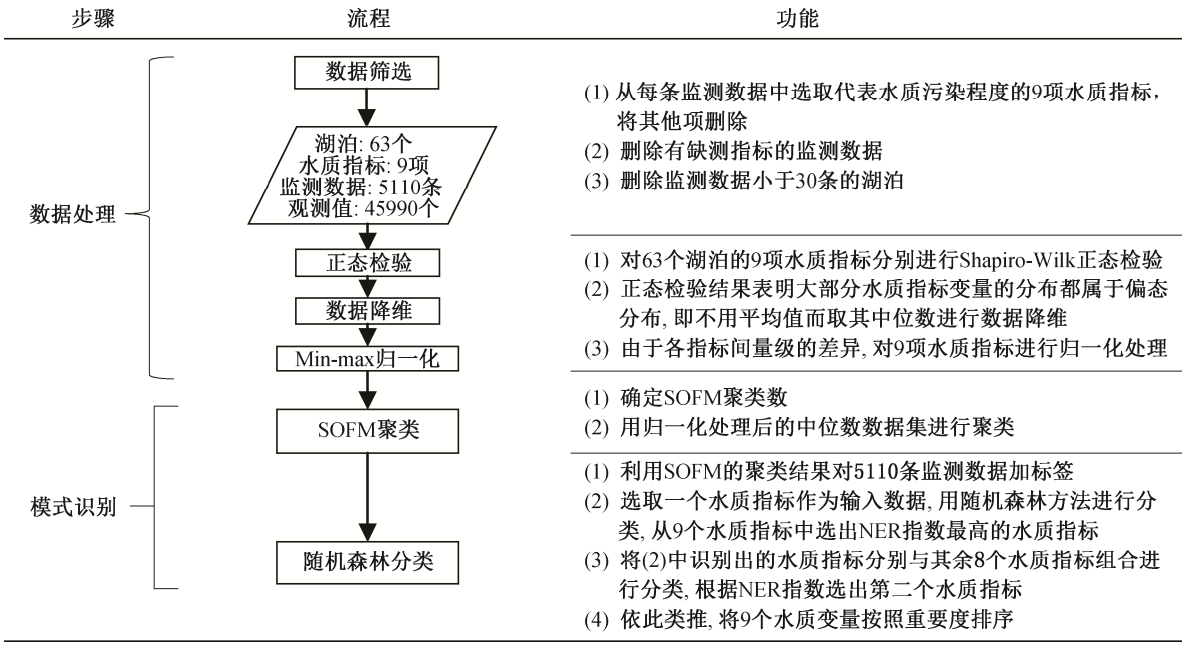


图 2　研究思路

Fig. 2　Research flow chart

表 1　不同湖泊正态分布的比例(%)

Table 1　Ratio of normal distributions in different lakes (%)

分布类型	pH	DO	COD_{Mn}	COD	BOD_5	NH_3-N	TN	TP	Chla
正态分布	46.0	55.6	66.7	73.0	79.4	93.7	88.9	90.5	90.5
偏态分布	54.0	44.4	33.3	27.0	20.6	6.4	11.1	9.5	9.5

大值和最小值。

1.4 模式识别方法

1.4.1 SOFM 神经网络构建

无监督聚类是数据类别未知, 根据样本间的相似性特征对数据进行聚类[23]。SOFM 网络具有拓扑保持能力和自组织概率分布特性, 并能提取输入数据的特征, 聚类结果更加客观[24]。当输入不同的样本后, 训练开始时 SOFM 网络输出层产生响应的神经元是随机的, 但经过自组织训练后, 会形成有序的神经元排列, 功能相近的神经元分布较近, 功能不同的神经元分布较远。主要步骤包括初始化输入层与输出层的连接权重, 确定输出层神经元优胜邻域, 输入训练样本, 通过自学习更新优胜邻域与连接权值[24]。

在以往的研究中, 一般都尝试不同的 SOFM 网络聚类数, 然后根据结果的合理性来确定。本研究根据 $5\times\sqrt{L}$ 公式(L 为湖泊个数, 本文为 63 个)[25], 将输出层设置为 5×8 的六边形结构, 由竞争层神经元间连接权重及产生响应神经元的空间分布位置来确定最终的分类数, 最后将输出层神经元个数设置为最终确定的分类数, 输出聚类结果。

1.4.2 RF 方法

监督分类指部分数据类别已知, 用已知类别的数据训练分类器, 对未知类别的数据进行分类[10]。RF 是一种基于分类树算法的分类方法, 其优点是运算速度快, 分类准确率极高; 可克服指标之间可能存在的多元共线性问题, 且不需要降维; 可对导致水质污染的水质指标进行重要度排序; 对离群值和缺省值不敏感[26]。

RF 方法可对导致湖泊水质污染的主要控制因子(水质指标)进行识别。该方法默认使用“袋外数据”(out of bag, oob)误差给出分类过程中各变量的重要性。当训练样本中各类样本数目相近时, oob 误差的识别效果较好; 当各类别样本数目差别较大时, 该指数易高估变量的重要性。所以, 本研究选用正确率指数(NER 指数)作为判别准则, 该指数能够反映分类结果的准确性, 并根据 RF 重分类结果对不同指标进行重要性排序。排序过程如下: 将已知分类(SOFM 聚类结果)的全部数据作为训练数据输入 RF, 构建分类函数对输入数据进行重分类, 若重分类结果与 SOFM 聚类结果一致, 则视为分类正确。据此, 首先采用单个变量作为 RF 的输入样本, 获得各变量对应的 NER, 选择具有最大 NER 的变量(即为第一重要的水质指标)分别与其他 8 个变量

组合作为 RF 输入样本，并选择具有最大 NER 的变量组合即可筛选出第二重要的水质指标，依此类推，进行筛选，即可对不同指标的重要度进行排序(图 2)。NER 指数计算公式为

$$\mathrm{NER}=\frac{1}{n}\sum_{i=1}^{n}\frac{x_i}{m_i}\times 100\%, \tag{2}$$

其中，n 为分类总数，m_i 为训练样本中第 i 类的样本总数，x_i 为用 RF 重分类后与 SOFM 聚类结果一致的样本数。

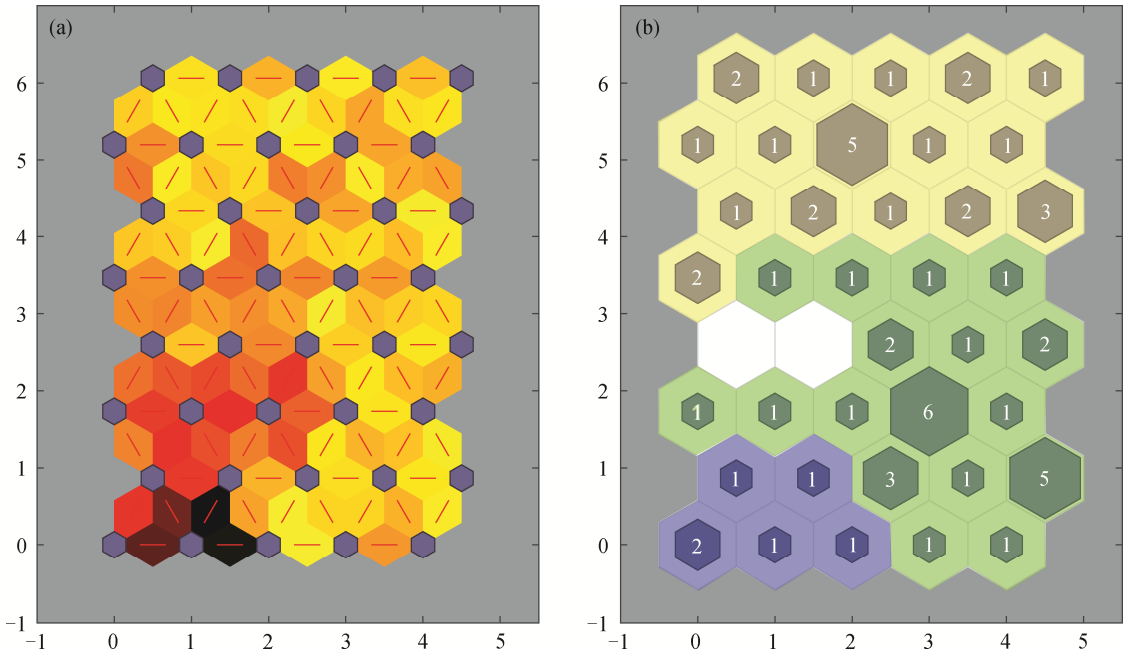
本研究的数据分析用 R 软件和 MATLAB 软件实现。正态性检验和数据归一化用 R 软件的自带软件包实现，SOFM 神经网络在 MATLAB 中利用 newsom 函数构建，RF 采用 R 软件的 randomForest 软件包实现。

2 结果与讨论

2.1 湖泊水质污染程度聚类

将 SOFM 网络输出层设置为 5×8 的六边形结构，对 63 个湖泊进行聚类，输出层神经元间的连接权重如图 3(a)所示，图 3(b)为输出层产生响应的神经元分布。综合图 3(a)和(b)，可认为输出层代表 3 种模式，即图 3(b)中黄色、绿色和蓝色 3 个区域。

根据上述分析，全国 63 个湖泊的聚类结果如表 2 所示。其中，第 1 类湖泊有 6 个，第 2 类湖泊有 27



(a) 蓝色正六边形为输出层神经元，长条形六边形表示神经元间的连接，其颜色深浅与连接权重相关，颜色越深表示连接权重越小、神经元间差异越大，即对应的输入层中湖泊水质污染程度相差越大; (b) 六边形内的数字代表在该输出层神经元上产生响应的湖泊个数

图 3 SOFM 网络输出层的神经元间连接权重(a)和产生响应的神经元空间分布(b)

Fig. 3 Connection weight between neurons (a) and locations map of responsive neurons (b) in SOFM network output layer

表 2 SOFM 网络聚类结果

Table 2 SOFM network clustering results

类别	湖泊名称
1	达赉湖，滇池草海，滇池外海，杞麓湖，星云湖，异龙湖
2	白洋淀，贝尔湖，察尔森水库，巢湖，巢湖东半湖，巢湖西半湖，大伙房水库，东平湖，董铺水库，斧头湖，富水水库，洪泽湖，崂山水库，龙感湖，骆马湖，南漪湖，太湖，太湖北部沿岸区，太湖东部沿岸区，太湖湖心区，太湖西部沿岸区，乌伦古湖，峡山水库，小浪底水库，阳澄湖，于桥水库，漳河水库
3	白莲河水库，莱子湖，程海，大广坝水库，丹江口水库，淀山湖，东江水库，洞庭湖，洱海，抚仙湖，高邮湖，隔河岩水库，洪湖，黄龙滩水库，梁子湖，泸沽湖，密云水库，磨盘山水库，南四湖，鄱阳湖，千岛湖，升金湖，松花湖，松涛水库，瓦埠湖，王瑶水库，武昌湖，新丰江水库，阳宗海，长潭水库

个，第3类湖泊有30个。对3类湖泊的各项水质指标求平均值(表3)，发现各类湖泊pH值差别不大，第2类湖泊的DO最高，其余7项水质指标浓度均为第1类高于第2类、第2类高于第3类，且第1类湖泊的COD、TP和Chla浓度极高，可达第2类湖泊的10倍以上。由此可知，第1类湖泊污染程度较严重，第2类湖泊污染程度中等，第3类湖泊污染程度较轻，水质良好。

不同类别湖泊的空间分布如图4所示，可以看出，污染程度较高的湖泊主要分布在云贵高原，这些湖泊属断裂陷落型湖泊，水深岸陡，对入湖污染物的净化能力较弱；入湖支流水系较多但出流水系较少，面源污染入湖的渠道多，但不利于污染物的排出，导致污染物的累积[27]。中等污染程度的湖泊主要分布于人口密度较大的东部平原地区，该区域长期以来对资源进行不合理的开发，对环境缺乏保护，例如围湖造田、放水发电、对水生生物过度捕捞、农业及生活废污水的排放等行为，致使水体污染严重，生物资源锐减，湖泊的生态环境遭到一定程度的破坏，因而该类湖泊水质污染的主要原因为流域经济发展与环境保护的不协调[28]。第3类湖泊所处流域的人口密度较小，经济发展程度较低，外源负荷输入较少，湖泊污染程度较轻[29]。

由地表水环境质量标准(GB3838—2002)可得63个湖泊的水质类别(表4)。其中，第1类的6个湖泊均为劣Ⅴ类，水质较差；第2类的27个湖泊均匀分布于Ⅱ类至劣Ⅴ类之间，水质中等；第3类的30个湖泊中有60%属于Ⅲ类，水质良好。由此也验证了SOFM聚类结果的可靠性以及将其作为先验信息进行RF分类的合理性。

2.2 湖泊水质主要控制因子识别

用RF对能够反映水质污染程度的主要水质指标进行识别。结果表明，COD_{Mn}的NER指数最高为68.85%，是9项水质指标中对水质污染程度影响最

表3　三类湖泊各项水质指标平均值

Table 3　Water quality indicators' average of three types of lakes

类别	pH	DO	COD_{Mn}	COD	BOD_5	NH_3-N	TN	TP	Chla	污染程度
1	8.8	7.2	13.7	64.4	5.5	1.16	2.84	0.3	0.075	高
2	8.0	9.2	4.1	19.6	2.4	0.19	1.53	0.05	0.007	中
3	7.9	7.7	2.7	11.6	1.6	0.17	0.83	0.03	0.005	低

说明：除pH外，其他水质指标的单位均为mg/L。

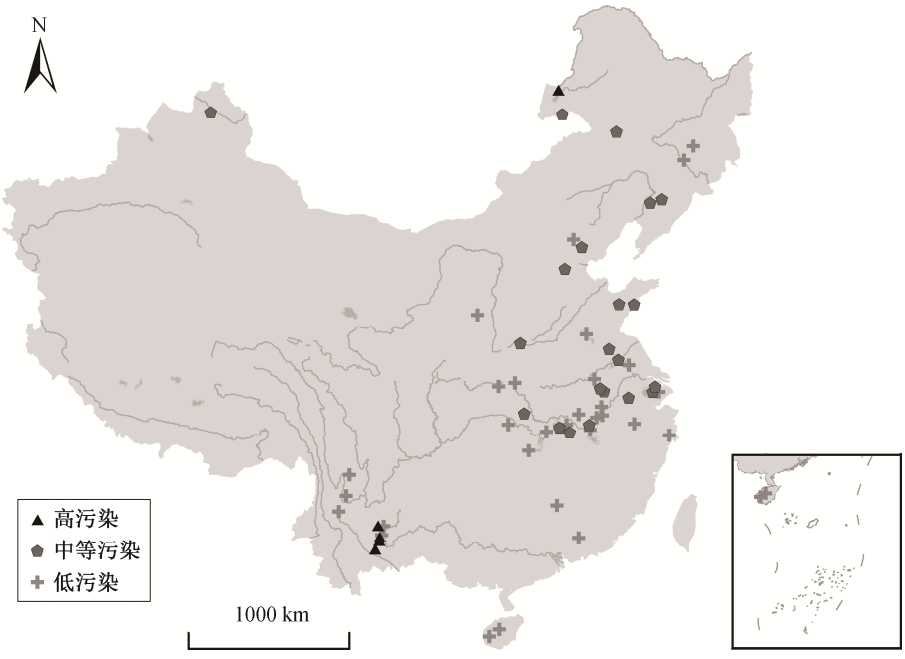


图4　三类湖泊的空间分布

Fig. 4　Locations map of three types of lakes

表 4 各湖泊的水质类别统计

Table 4 Water quality statistics for each lake

湖泊类别	水质类别比例/%					
	Ⅰ类	Ⅱ类	Ⅲ类	Ⅳ类	Ⅴ类	劣Ⅴ类
1	0	0	0	0	0	100.0
2	0	7.4	18.5	18.5	25.9	29.6
3	3.3	16.7	60.0	10.0	6.8	3.3

大的指标; 将 COD_{Mn} 与其他 8 项指标依次组合作为训练数据, 可得 COD_{Mn} 和 Chla 的 NER 指数最高, 为 79.54%, 即 Chla 是对污染程度影响第二重要的指标。同理, 可得水质指标对污染程度决定性的重要度排序: COD_{Mn}>Chla>DO>TN>TP>COD>BOD_5>pH>NH_3-N (图 5)。

由上述结果可知, 当只用 COD_{Mn} 和 Chla 进行分类时, 准确率接近 80%。COD_{Mn} 可表征水体中的有机物含量, Chla 是常见的表征水体中藻类浓度(即湖泊富营养化程度)的指标[30]。当选取这两个指标对湖泊水质污染程度进行识别时, 由表 3 可以看出, 第 1 类湖泊属于高 COD_{Mn}、高 Chla 型, 第 2 类湖泊属于中 COD_{Mn}、中 Chla 型, 第 3 类湖泊属于低 COD_{Mn}、低 Chla 型, 这与由 9 个水质指标进行模式识别得出的湖泊污染程度结果一致。若控制准确率为 90%, 则需选用 COD_{Mn}, Chla, DO, TN, TP 和 COD 这 6 个水质指标。

因此, 在湖泊水质监测中应特别重视 COD_{Mn} 和 Chla 的监测和分析, 可适当地增加这两项水质指标的监测频次, 提高对湖泊水质的认知, 同时可适当减少其他对水质污染程度代表性较低的水质指标的监测频率, 在保证污染识别精度的前提下, 有效降低监测费用[9]。基于 RF 的识别结果(图 5), 决策者可根据自身对水质分类准确率的不同需求, 选取相应的重点监测指标。

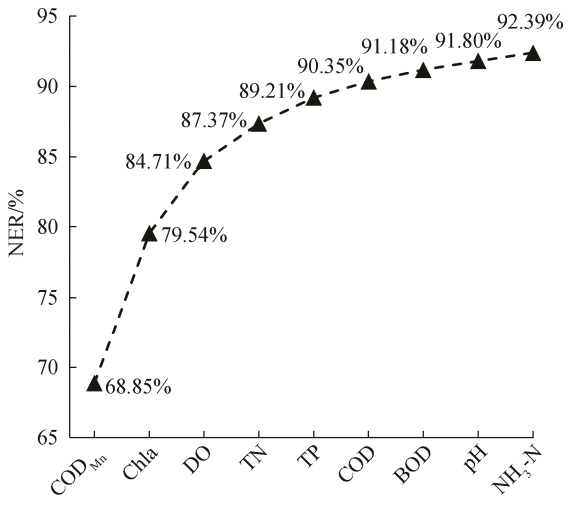


图 5 水质指标的累积 NER 指数值

Fig. 5 NER index of each water quality indicator

在以往的研究中, SOFM 通常采用多个聚类数进行试验, 选取聚类结果较合理的作为最终的聚类数。该方法受主观因素影响较大, 无法保证聚类结果的准确性。本研究从 SOFM 的原理出发, 根据输出层神经元连接权重与产生响应神经元的空间离散程度确定聚类数目, 增强了聚类结果的可信度。用各指标的中位数进行 SOFM 聚类, 会造成大量的数据损失, 将 SOFM 与 RF 两种模式识别方法进行耦合, 可以充分利用数据集中的每一条监测数据。根据 SOFM 聚类结果, 在整个数据集上用 RF 进行监督分类, 可以克服常见的模式识别方法中由于输入数据维度的限制而需对输入样本进行维度压缩导致的无法对所有数据进行聚类的缺点。

3 结论

本文对我国 63 个湖泊 11 年的 9 种水质指标进行模式识别, 根据水质污染程度, 63 个湖泊可分为 3 类。水质指标对污染程度决定性的重要度排序为 COD_{Mn}>Chla>DO>TN>TP>COD>BOD_5>pH>NH_3-N。在分类准确率为 80%时, 选取 COD_{Mn} 和 Chla 两项水质指标即可识别湖泊污染程度。因此, 在湖泊水质监测中, 可适当地增加对 COD_{Mn} 和 Chla 的监测频次, 减少其他水质指标的监测频率, 达到在保证湖泊水质精确评价的前提下降低监测费用的效果。

本研究提出的耦合 SOFM 和 RF 的方法能够对所有水质数据进行分析, 并识别水质的污染程度和代表性水质指标。本研究验证了该方法的合理性, 未来可采用该方法对其他水体污染特征和主要控制指标进行识别。

参考文献

[1] Barnett T P, Pierce D W, Hidalgo H G, et al. Human-induced changes in the hydrology of the western United States. Science, 2008, 319: 1080–1083

[2] Harper D, Zalewski M, Pacini N. Ecohydrology: processes, models and case studies: an approach to the sustainable management of water resources. Trowbridge: Cromwell Press, 2008

[3] Kozaki D, Rahim M H B A, Ishak W M F B W, et al. Assessment of the river water pollution levels in Kuantan, Malaysia, using ion-exclusion chromatographic data, water quality indices, and land usage patterns. Air Soil & Water Research, 2016, 9: 1–11

[4] Wetzel R G. Limnology: lake and river ecosystems. Eos Transactions American Geophysical Union, 2001, 21(2): 1–9

[5] Lavine B K, Rayens W S. Comprehensive Chemometrics. Amsterdam: Elsevier, 2009

[6] Bücker A, Crespo P, Frede H G, et al. Identifying controls on water chemistry of tropical cloud forest catchments: combining descriptive approaches and multivariate analysis. Aquatic Geochemistry, 2010, 16(1): 127–149

[7] Juahir H, Zain S M, Aris A Z, et al. Spatial assessment of Langat River water quality using chemometrics. J Environ Monit, 2010, 12(1): 287–295

[8] Shrestha S, Kazama F. Assessment of surface water quality using multivariate statistical techniques: a case study of the Fuji river basin, Japan. Environmental Modelling & Software, 2007, 22(4): 464–475

[9] Sotomayor G, Hampel H, Vázquez R F. Water quality assessment with emphasis in parameter optimisation using pattern recognition methods and genetic algorithm. Water Research, 2018, 130: 353–362

[10] 刘勇健, 沈军. 自组织神经网络法综合评价水质. 勘察科学技术, 2003(4): 22–25

[11] Tan P N, Steinbach M, Kumar V. 数据挖掘导论(完整版). 范明, 范宏建, 译. 北京: 人民邮电出版社, 2011

[12] 郑晓君, 罗妮娜, 裴洪平. 利用SOFM网络评价杭州西湖水质的时空变化. 生物数学学报, 2007, 22(2): 317–322

[13] Zhang Xianqi, Feng Wenhong. Self-organizing neural networks evaluation model and its application // International Conference on Artificial Intelligence and Education. Hangzhou, 2010: 52–55

[14] 刘娅, 朱文博, 李双成. 基于SOFM神经网络的京津冀地区水源涵养功能分区. 环境科学研究, 2015, 28(3): 369–376

[15] 方匡南, 吴见彬, 朱建平, 等, 随机森林方法研究综述. 统计与信息论坛, 2011, 26(3): 32–38

[16] 明均仁, 肖凯. 基于R语言的面向需水预测的随机森林方法. 统计与决策, 2012(9): 81–83

[17] 康有, 陈元芳, 顾圣华, 等. 基于随机森林的区域水资源可持续利用评价. 水电能源科学, 2014, 32(3): 34–38

[18] 张颖, 高倩倩. 基于随机森林分类算法的巢湖水质评价. 环境工程学报, 2016, 10(2): 992–998

[19] Shapiro S S, Wilk M B. An analysis of variance test for normality. Biometrika, 1965, 52(3): 591–599

[20] Carpenter M. The new statistical analysis of data. Journal of the American Statistical Association, 1996, 42(2): 205–206

[21] Helsel D R, Hirsch R M. Statistical methods in water resources. Technometrics, 2002, 174(1): 466–467

[22] Todeschini R, Ballabio D, Consonni V. Distances and other dissimilarity measures in chemometrics. Hoboken: John Wiley & Sons, 2015

[23] Frank I E, Todeschini R. The Data Analysis Handbook. Technometrics, 1994, 38(2): 193

[24] 叶敏婷, 王仰麟, 彭建, 等. 基于SOFM网络的云南省土地利用程度类型划分研究. 地理科学进展, 2007, 26(2): 97–105

[25] Astel A, Tsakovski S, Barbieri P, et al. Comparison of self-organizing maps classification approach with cluster and principal components analysis for large environmental data sets. Water Research, 2007, 41(19): 4566–4578

[26] 李欣海. 随机森林模型在分类与回归分析中的应用. 应用昆虫学报, 2013, 50(4): 1190–1197

[27] 于洋, 张民, 钱善勤, 等. 云贵高原湖泊水质现状及演变. 湖泊科学, 2010, 22(6): 820–828

[28] 孟庆义. 国内湖泊水质污染及富营养化治理. 北京水务, 2001(5): 45–47

[29] 蒋火华, 吴贞丽. 世界典型湖泊水质探研. 世界环境, 2000(4): 35–37

[30] 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. 环境科学学报, 2014, 34(3): 754–762