DOI:10.11705/j.issn.1672-643X.2022.05.01

# 基于粗糙集-贝叶斯网络的流域水质评价方法

白 云[1], 李 勇[2], 张万娟[3], 贺 嘉[4]
(1. 重庆工商大学 管理科学与工程学院, 重庆 400067; 2. 贵州医科大学 环境污染与疾病监控教育部重点实验室, 贵州 贵阳 550025; 3. 西北农林科技大学 经济管理学院, 陕西 杨凌 712100; 4. 重庆工商大学 长江上游经济研究中心, 重庆 400067)

**摘 要**: 针对现实流域水质评价中存在不完整和不确定信息这一问题,基于智能互补思想,提出了粗糙集-贝叶斯网络的流域水质评价方法。使用粗糙集理论提取影响流域水质状态的主要因子,得到最小属性约简集以降低建模复杂度;然后构造贝叶斯网络并进行模型训练,获得其网络结构和条件概率表,实现流域水质的概率决策推理;最后对嘉陵江流域重庆段的3个水质监测断面进行实例分析,验证该方法的有效性和准确性。结果表明:该方法可以正确进行流域水质评价推理,相比于其他方法(贝叶斯网络、灰色-贝叶斯网络、粗糙集-朴素贝叶斯),其具有最高的准确率(>0.97)、精确率(>0.86)、召回率(>0.86)和$F1$值(>0.86),为流域水环境管理提供了有效的技术支持。

**关键词**: 水质评价; 粗糙集理论; 贝叶斯网络; 属性约简; 不确定性
**中图分类号**:X824 **文献标识码**: A **文章编号**: 1672-643X(2022)05-0001-10

# Watershed water quality assessment method based on rough set and Bayesian network

**BAI Yun[1], LI Yong[2], ZHANG Wanjuan[3], HE Jia[4]**
(1. *School of Management Science and Engineering, Chongqing Technology and Business University, Chongqing* 400067, *China*; 2. *Key Laboratory of Environmental Pollution Monitoring and Disease Control, Ministry of Education, Guizhou Medical University, Guiyang* 550025, *China*; 3. *College of Economics & Management, Northwest A&F University, Yangling* 712100, *China*; 4. *Research Center for Economy of Upper Reaches of the Yangtze River, Chongqing Technology and Business University, Chongqing* 400067, *China*)

**Abstract**: According to the intelligent complementary strategy, a new watershed water quality assessment method based on rough set (RS) and Bayesian network (BN) was presented for the water quality assessment containing incomplete and uncertain information. Firstly, RS was used to extract the main factors affecting watershed water quality, so as to obtain the minimum attribute reduction set, which can be used to reduce the modelling complexity. Then, the BN was constructed and trained based on the attribute reduction set, and its network structure and conditional probability table were obtained to realize the probabilistic decision reasoning of watershed water quality. Finally, the model evaluation indexes were used to analyse three water quality monitoring sections in the Chongqing section of the Jialing River to verify the correctness and effectiveness of this method. The results show that this method is applicable to the watershed water quality assessment, and has the highest accuracy (>0.97), precision (>0.86), recall (>0.86), and $F1$-measure (>0.86) compared with other methods (BN, GRA-BN, RS-NB). This method can provide an effective technical support for the water environment management in the watershed.

**Key words**: water quality assessment; rough set (RS); Bayesian network (BN); attribute reduction; uncertainty

**收稿日期**:2021-10-26; **修回日期**:2022-04-26
**基金项目**:国家自然科学基金项目(71801044); 国家社会科学基金项目(18CJY005)
**作者简介**:白 云(1985-),男,山西阳泉人,博士,副教授,硕士生导师,研究方向为水资源管理。
**通讯作者**:李 勇(1991-),男,重庆合川人,博士,研究方向为环境大数据分析。

# 1 研究背景

近年来,随着社会经济的飞速发展和城市规模的不断扩大,我国面临的水环境污染、水资源短缺和流域生态安全等问题日益突出,制约了区域经济社会的可持续发展,威胁着居民的饮用水安全[1-3]。在水环境的规划管理和污染治理中,对流域水环境质量进行科学的综合评价意义重大。目前,用于水质评价的方法主要有单因子指数法[4]、综合污染指数法[5]、多元统计分析法[6]、指数平滑法[7]、神经网络[8]以及灰色系统评价法[9],以上方法在实际应用中各有利弊。单因子指数法将水质最差的类别作为评价结果,结果过于消极;综合污染指数法综合性强,但无法科学地确定各水质指标的权重;多元统计分析和指数平滑等统计方法结构简单,但不适用于非线性数据处理和多指标综合建模;神经网络和灰色系统等智能方法具有较强的自适应学习能力,但对数据要求严格[10]。流域水环境是一个动态系统,水质信息复杂多变且充满不确定性,作为当前概率推理和不确定性知识表达领域最有效的模型之一,贝叶斯网络能够更好地解决水环境不确定性问题,在水质预测评价等领域得到了更多的应用[11-13]。然而,由于水质监测数据本身所具备的不完整性、大量性、高维性和动态性等特点,单一的贝叶斯网络在水质评价方面还存在一定的不足。

粗糙集理论作为一种信息分析处理理论,其最大优点在于在没有先验知识的情况下也能够将不完整信息进行有效分析,从中发现隐含信息和潜在规律,从而获得最小知识表达[14]。近年来,粗糙集理论因其独特的数据分析能力而备受关注,在诊断、分类、预测、知识获取、数据挖掘等领域均有成功的应用[15-18]。目前,有学者已将粗糙集与贝叶斯网络结合起来开展研究,如 Xiao 等[19]将其应用于高速织机故障诊断,Xu 等[20]将其应用于地震危险性定量评估,而将二者共同应用于流域水质分析的实例还未见报道。本文将粗糙集理论和贝叶斯网络联合应用于流域水质评价,在粗糙集属性约简的基础上,充分利用贝叶斯网络在不确定系统中的概率推理能力,合理评价嘉陵江流域重庆段的水环境质量,以期为流域水环境的科学管理及防控提供参考。

# 2 数据来源与研究方法

## 2.1 研究区概况

嘉陵江位于东经 103°45′ ~ 109°00′、北纬 29°20′ ~ 34°25′之间,全长 1 345 km,流域总面积为 $16\times10^4$ km$^2$,是长江支流中流域面积最大,长度仅次于汉江的河流;其中,重庆境内长约 153 km,是汇入长江的最后一个省级行政区,沿岸不仅有一些工业园和企业,还有许多城镇和居民点,工业废水、生活污水、农业面源污染物等多种水源均可进入嘉陵江,在一定程度上影响了流域水质。嘉陵江流域在重庆境内有 3 个水质监测断面,分别为北温泉、梁沱和大溪沟断面,均为重庆主城区重要饮用水水源地,且位于三峡库区上游,水质标准要求高,对其进行水质评价、了解其水质状况及水质变化情况尤为重要。嘉陵江流域重庆段水系分布及水质监测断面位置见图 1。

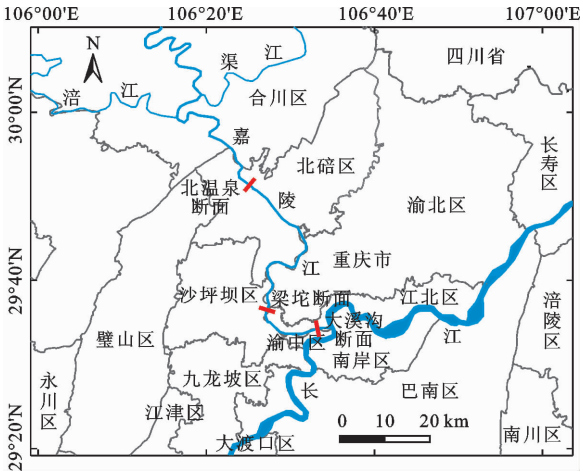



图 1 嘉陵江流域重庆段水系分布及水质监测断面位置

## 2.2 数据来源

本文的研究范围为嘉陵江流域重庆段的 3 个监测断面,研究时段为 2010 - 2016 年。监测断面水质数据(内环境因素)来源于重庆市生态环境监测中心,其水质监测频率为每周 1 次,包括 pH 值、溶解氧(DO)、高锰酸盐指数($COD_{Mn}$)、氨氮($NH_3$—N)和总磷(TP)等 5 个水质指标。根据《地表水环境质量标准》(GB 3838—2002)[21]将水质划分为五类。当 5 个水质指标数据均缺失时,则舍弃该周的数据,若部分缺失则使用均值/最大概率值法[22]对其进行填补,北温泉、梁沱和大溪沟最终分别有 258 周、349 周和 327 周数据,数据缺失较为严重。选择季节(春、夏、秋、冬)、温度(平均温度、最高温度、最低温度)、降水量、天气(雨、阴、晴)作为可能影响水质状况的外环境因素,数据收集于全球天气网(http://lishi.tianqi.com/chongqing/)。

## 2.3 研究方法

2.3.1 粗糙集理论 粗糙集理论(rough set, RS)是由 Pawlak 提出的可处理不精确和不完整信息的一种数据分析处理方法,其主要功能之一是对指标属性进

行约简[23]。设有一个知识表达系统 $S = (\boldsymbol{U}, \boldsymbol{A}, \boldsymbol{V}, f)$。其中:$\boldsymbol{U}$ 为论域,是由 $n$ 个对象组成的非空有限集合;$\boldsymbol{A}$ 为属性集,是条件属性集 $\boldsymbol{C}$ 和决策属性集 $\boldsymbol{D}$ 的并集;$\boldsymbol{V}$ 为属性值的集合;$f$ 为信息函数,$f:\boldsymbol{U}\times\boldsymbol{A}\rightarrow\boldsymbol{V}$,用以确定 $\boldsymbol{U}$ 中每个对象 $x$ 的属性值。具有条件属性和决策属性的知识表达系统也叫决策表。

粗糙集通过定义 1 个参数即依赖度 $r$ 来衡量 $\boldsymbol{D}$ 对 $\boldsymbol{C}$ 的依赖程度,公式如下[17]:

$$r = \gamma_C(\boldsymbol{D}) = \frac{card(\boldsymbol{POS}_C(\boldsymbol{D}))}{card(\boldsymbol{U})} \tag{1}$$

式中:$card(\cdot)$ 为集合中的元素个数;$\boldsymbol{POS}_C(\boldsymbol{D})$ 为 $\boldsymbol{D}$ 关于 $\boldsymbol{C}$ 的下近似集,即能够全部被 $\boldsymbol{D}$ 包含的 $\boldsymbol{U}$ 中所有按等价关系 $\boldsymbol{C}$ 划分的集合;$card(\boldsymbol{U})$ 为论域 $\boldsymbol{U}$ 的元素个数。

在不同的内部和外部环境条件下,影响综合水质类别的因素也将有所不同,所以水质评价决策表,就是在保证流域水质各项指标正确归类的前提下,把水质评价决策表中不重要或不相关的影响因素剔除掉,进而形成 1 个最优属性集的简约结果。

2.3.2　贝叶斯网络　贝叶斯网络(Bayesian network, BN)是由 Pearl 提出的一种基于图论和概率论的不确定知识表达模型,是智能化系统中处理不确定信息的主要方法之一[24]。对一个给定的水质指标时间序列,在贝叶斯网络建模前,首先计算水质节点的类别,公式如下:

$$P(B_{i,j} \mid A_j) = \frac{P(B_{i,j})P(A_j \mid B_{i,j})}{\sum_{i=1}^{5} P(B_{i,j})P(A_j \mid B_{i,j})} \tag{2}$$

式中:$i$ 为水质类别,$i = 1, 2, \cdots, 5$;$j$ 为水质指标个数,$j = 1, 2, \cdots, 5$;$B_{i,j}$ 为某水质指标 $j$ 为 $i$ 类水质的标准值;$A_j$ 为某水质指标 $j$ 的监测值;$P(B_{i,j})$ 为先验概率,即根据先验知识判断水样为 $i$ 类水的概率;$P(A_j \mid B_{i,j})$ 为条件概率,即水质属于 $i$ 类水时,监测值为 $A_j$ 的概率;$P(B_{i,j} \mid A_j)$ 为后验概率,即监测值为 $A_j$ 的条件下,水质类别为 $i$ 的概率。

对于先验概率 $P(B_{i,j})$,在没有任何先验信息条件下,水样属于某一类别的概率相等,即 $P(B_{1,j}) = P(B_{2,j}) = P(B_{3,j}) = P(B_{4,j}) = P(B_{5,j}) = 1/5$。

对于条件概率 $P(A_j \mid B_{i,j})$,计算方式采用距离法[10],公式如下:

$$P(A_j \mid B_{i,j}) = \frac{1/L_{i,j}}{\sum_{i=1}^{5} 1/L_{i,j}} \tag{3}$$

式中:$L_{i,j} = |A_j - B_{i,j}|$,表示监测值 $A_j$ 离标准值 $B_{i,j}$ 的距离,$L_{i,j}$ 越大,则水质属于 $i$ 类水的可能性就越小。

通过公式(3)得出各水质指标对水质类别的后验概率后,即可计算多个指标下水质的综合后验概率 $P_i$:

$$P_i = \sum_{j=1}^{5} w_j P(B_{i,j} \mid A_j) \tag{4}$$

式中:$w_j$ 为某水质指标 $j$ 的权重,可通过变异系数权重法求得[25],即:$w_j = \delta_j \Big/ \sum_{j=1}^{5} \delta_j$;$\delta_j$ 为某水质指标 $j$ 的变异系数,定义为标准差与均值的比率。

确定最终的综合水质类别,即将概率最大的水质类别作为最终所属类别 $c$:

$$P_c = \max P_i \quad (i = 1,2,\cdots,5) \tag{5}$$

2.3.3　粗糙集 - 贝叶斯网络评价方法构建　嘉陵江重庆断面水质评价的粗糙集 - 贝叶斯网络方法结构见图 2,具体步骤如下:

步骤 1:数据获取和预处理。收集内环境和外环境数据,对内环境数据缺失值采用直接删除或均值/最大概率值法[22]补齐的方式进行预处理。

步骤 2:数据离散化。根据水质分类标准限值和 Entropy/MDL 算法[26]分别对连续型的内、外环境数据进行离散化处理。

步骤 3:粗糙集属性约简。采用遗传算法[3]对流域水质的影响因子进行属性约简,得到每个属性的重要性。

步骤 4:贝叶斯网络结构学习。采用 K2 算法[27]和文献分析结果构造流域水质评价模型的贝叶斯网络结构。

步骤 5:贝叶斯网络参数学习。采用最大似然估计法[28]进行贝叶斯网络参数学习,得到网络中各指标节点的条件概率表。

步骤 6:证据推理。采用联结树推理引擎[29]对节点的连接证据进行推理,获得综合水质类别。

步骤 7:流域水质评价结果。

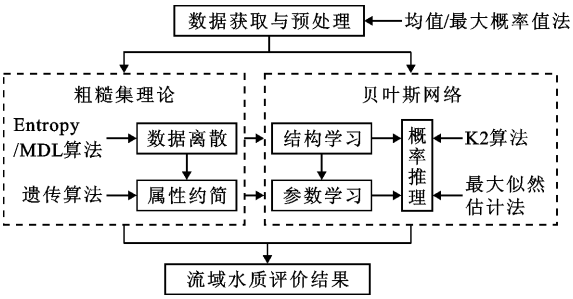



**图 2　粗糙集 - 贝叶斯网络水质评价方法结构**

2.3.4　评估指标　为了评估模型的水质预测效果,采用准确率($Ac$)、精确率($Pr$)、召回率($Re$)和

$F1$ 值作为模型评估指标[15]。公式如下：

$$Ac = \frac{ZP + ZN}{ZP + ZN + FP + FN} \tag{6}$$

$$Pr = \frac{ZP}{ZP + FP} \tag{7}$$

$$Re = \frac{ZP}{ZP + FN} \tag{8}$$

$$F1 = 2 \times \frac{Pr \cdot Re}{Pr + Re} \tag{9}$$

式中：$ZP$ 为被正确分类为正例的数量；$ZN$ 为被正确分类为负例的数量；$FP$ 为被错误分类为正例的数量；$FN$ 为被错误分类为负例的数量。

## 3 结果与分析

### 3.1 流域水质指标分析

图 3 为嘉陵江流域重庆段 3 个监测断面水质指标的时间序列图。

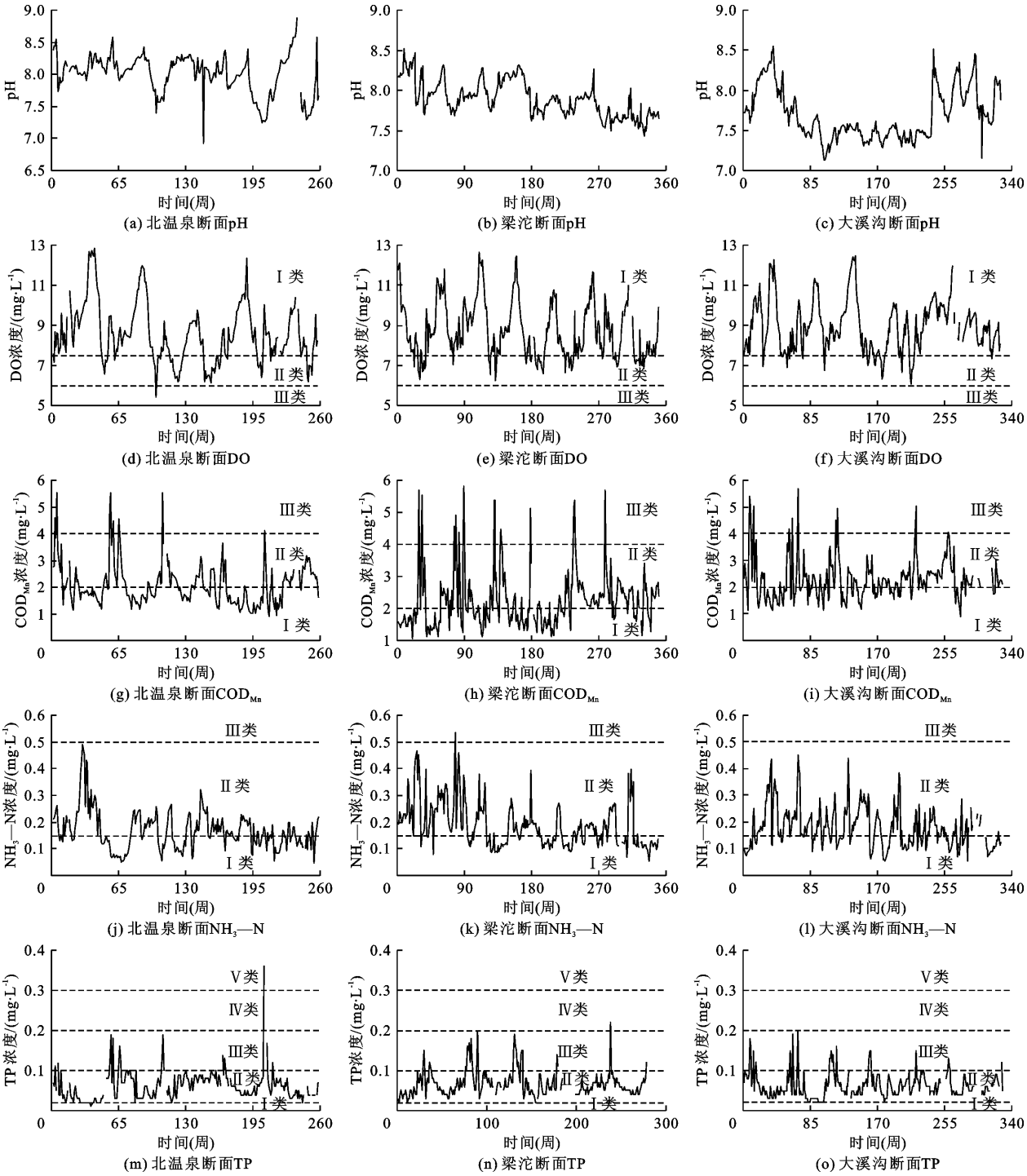



图 3 嘉陵江流域重庆段 3 个监测断面水质指标的时间序列

由图 3 可以看出,除个别周的水环境质量为Ⅲ类以外,3 个断面其他周的 DO 和 $NH_3$—N 常年保持在Ⅰ类(DO 浓度≥7.5 mg/L、$NH_3$—N 浓度≤0.15 mg/L)或Ⅱ类标准(7.5 mg/L > DO 浓度≥6.0 mg/L、0.5 mg/L≥$NH_3$—N 浓度 >0.15 mg/L);相对而言,$COD_{Mn}$和 TP 污染较为严重,部分数据处于Ⅲ类水质标准(6 mg/L≥$COD_{Mn}$浓度 > 4 mg/L、0.2 mg/L≥TP 浓度 >0.1 mg/L),北温泉断面有一突变点的 TP 甚至达到Ⅴ类水质标准(0.4 mg/L≥TP 浓度 >0.3 mg/L)。此外,就时间序列的基本特征而言,数据缺失较为严重,波动性较大且无明显周期性,DO、COD、$NH_3$—N 和 TP 平均浓度分别为 8.76、2.21、0.18、0.06 mg/L,标准偏差分别为 1.38、0.81、0.08、0.03 mg/L,较高的标准偏差证实了此数据的不确定性,增加了水质模型评价的难度。

### 3.2　基于粗糙集的水质因子属性约简

3.2.1　*数据离散*　粗糙集理论不能对连续型数据进行直接处理,在属性约简前,需对其进行离散化。对于 DO、$COD_{Mn}$、$NH_3$—N 和 TP 等内环境因素,根据《地表水环境质量标准》中的水质分类标准进行离散化,Ⅰ~Ⅳ类水依次定义为状态 1~5;对于 pH 值和连续型的外环境因素(平均温度、最高温度、最低温度、降水量),采用 Entropy/MDL 算法进行离散化;对于字符型离散数据则直接定义离散值,天气因素将雨天、阴天、晴天分别定义为状态 1、2、3,季节因素将春、夏、秋、冬分别定义为状态 1、2、3、4。内、外环境因素的离散化结果分别见表 1 和 2。

由表 1 可知,水质指标 DO 和 pH 值被分为两类,且状态 1 的数据较多,其余水质因子均是状态 2 较多;$COD_{Mn}$和 $NH_3$—N 有 3 种状态,但状态 3 的数据较少;TP 有 5 种状态,状态 1、4 和 5 的数据均较少。由表 2 可知,平均温度、最高温度、最低温度均有 3 种状态,数据分布较为均匀;降水量被分为 6 类,且状态 1(无降水)的数据最多;天气则是状态 1(雨天) > 状态 2(阴天) > 状态 3(晴天)。

3.2.2　*粗糙集属性约简*　数据离散化后,采用遗传算法对水质的影响因素进行粗糙集属性约简。算法参数设计如下:总群大小和最大迭代次数分别为 70 和 30,交叉概率和变异概率分别为 0.7 和 0.05。由于需要综合考虑内、外环境因素对水质的影响,因此选择重要性大于 0.4 的属性作为约简集(若重要性大于 0.6 则有断面未涉及外环境因素)。3 个断面的粗糙集约简结果见表 3。基于表 3 的约简结果,则可进行贝叶斯网络的建模和分析。

### 3.3　基于粗糙集 - 贝叶斯网络的水质评价

3.3.1　*贝叶斯网络结构学习*　选择数据集中的 80% 当作训练集进行网络建模,剩余的 20% 当作测试集进行网络验证。

**表 1　内环境因素的离散化结果**

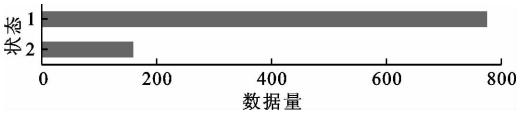
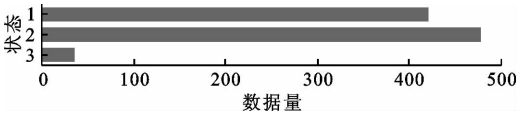
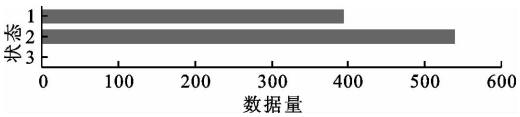
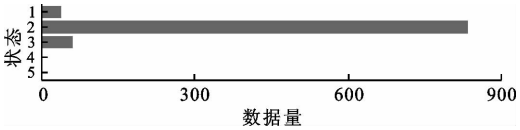
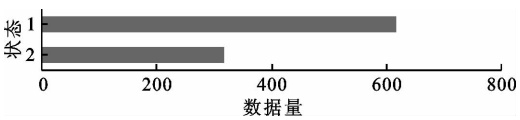
| 内环境因素 | 状态 | | | | | 数据分布 |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | |
| DO/(mg·L$^{-1}$) | ≥7.5 | ≥6 | | | |  |
| $COD_{Mn}$/(mg·L$^{-1}$) | ≤2 | ≤4 | ≤6 | | |  |
| $NH_3$—N/(mg·L$^{-1}$) | ≤0.15 | ≤0.5 | ≤1.0 | | |  |
| TP/(mg·L$^{-1}$) | ≤0.02 | ≤0.1 | ≤0.2 | ≤0.3 | ≤0.4 |  |
| pH | ≤8 | ≤9 | | | |  |

表2 外环境因素的离散化结果

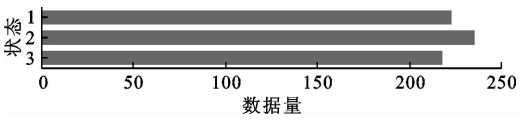
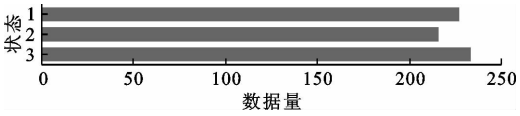
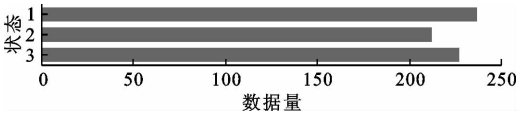
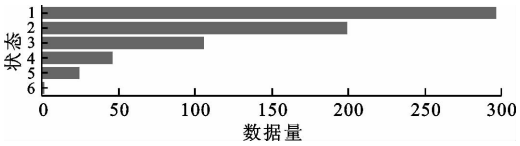
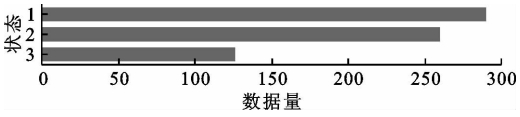
| 外环境因素 | 状态 1 | 2 | 3 | 4 | 5 | 6 | 数据分布 |
|---|---|---|---|---|---|---|---|
| 平均温度/℃ | [0,16) | [16,24) | [24,40) | | | |  |
| 最高温度/℃ | [0,19) | [19,28) | [28,45) | | | |  |
| 最低温度/℃ | [0,14) | [14,21) | [21,35) | | | |  |
| 降水量/mm | [0,0] | (0,5] | (5,15] | (15,30] | (30,70] | (70,140] |  |
| 天气 | 雨天 | 阴天 | 晴天 | | | |  |



表3 粗糙集属性约简结果

| 断面 | 属性集 |
|---|---|
| 北温泉断面 | $COD_{Mn}$、$NH_3$—N、TP、pH、最低温度、降水量、天气、季节 |
| 梁沱断面 | DO、$COD_{Mn}$、$NH_3$—N、TP、pH、最低温度、降水量、天气、季节 |
| 大溪沟断面 | DO、$COD_{Mn}$、$NH_3$—N、TP、最低温度、降水量、天气、季节 |

水质评价模型的贝叶斯网络结构通过K2算法和文献分析结果确定。首先,设置初始值,最大父节点数;节点状态数为离散节点的状态数;节点排序为节点的顺序,该模型设置的初始顺序为:外环境因素、内环境因素、水质类别。然后寻找每个节点在节点排序中的父节点,使用评分函数评价网络优劣,当新找到的父节点评分高于原结构时,将此节点作为当前节点的父节点,直到搜索完所有父节点,确定最终的网络结构。3个水质监测断面的贝叶斯网络结构见图4。

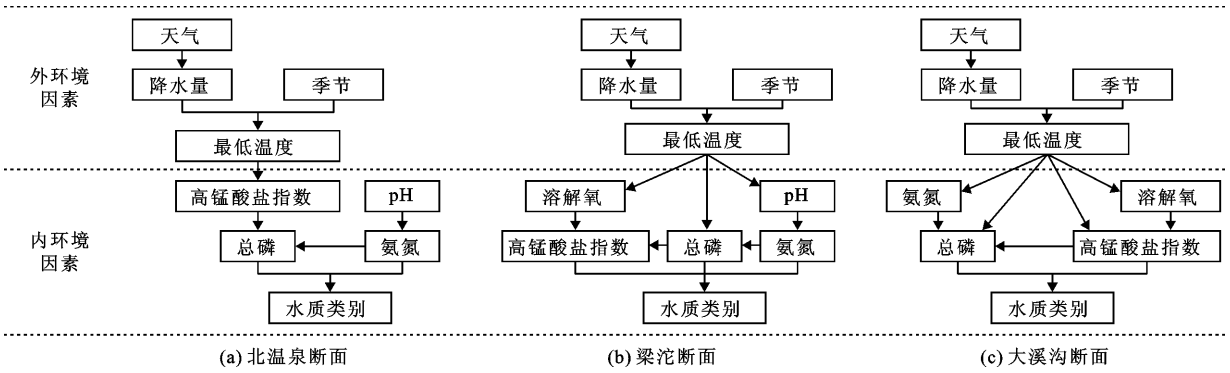



图4 3个水质监测断面的贝叶斯网络结构图

由图4可以看出,3个断面的网络结构均被分为两大部分,即外环境因素影响内环境因素,内环境因素影响最终水质类别。具体而言,3个断面的外环境影响因素相同,均涉及天气、降水量、季节和最

低温度,且因果关系一致;3 个断面的内环境因素存在差异,北温泉断面的 DO、大溪沟断面的 pH 值与水质类别没有直接或间接联系,而梁沱断面的水质类别与 5 个内环境因素均存在关联性;对于北温泉断面,最低温度影响 $COD_{Mn}$ 浓度,TP 和 $NH_3$—N 浓度直接影响水质类别的划分(图 4(a));梁沱断面的网络结构比北温泉断面复杂,最低温度可直接影响 DO、TP 和 pH 值,而直接影响水质类别的因素有 $COD_{Mn}$、TP 和 $NH_3$—N(图 4(b));对于大溪沟断面,$NH_3$—N、TP、$COD_{Mn}$ 和 DO 均受最低温度的直接影响,水质类别则直接受 TP 和 $COD_{Mn}$ 的影响(图 4(c))。值得注意,TP 与 3 个监测断面的综合水质类别均有直接关系,表明嘉陵江重庆段水质的主要污染指标是 TP,造成 TP 污染严重的原因包括山地地形、生活污染、工业污染和农业污染等[1,30]。

3.3.2　贝叶斯网络参数学习　贝叶斯网络结构建立后,使用最大似然估计法进行参数学习,得到网络节点的条件概率表,进而推算各个节点对目标节点的影响。首先分析外环境因素对水质状态的影响。分析发现,季节和最低温度对 3 个断面的水质影响较大,具体表现在夏季的水质最差,出现Ⅲ类及以上水质的概率明显大于其他季节(表 4);与夏季相对应,高温时水质状态最差,例如,随着最低温度的升高,Ⅲ类水质的概率逐渐增大,北温泉断面从 2% 增加到 10%,梁沱断面从 16% 增加到 37%,大溪沟断面从 0 增加到 10%(表 5)。相较而言,天气和降水量对水质状态的影响较小,本文通过梁沱断面的降水情况来举例分析,结果见表 6。

**表 4　不同季节对 3 个断面水质状态的影响占比** %

| 断面 | 季节 | Ⅰ | Ⅱ | Ⅲ | Ⅳ |
|---|---|---|---|---|---|
| 北温泉断面 | 春季 | 1 | 95 | 4 | 0 |
| | 夏季 | 0 | 89 | 10 | 1 |
| | 秋季 | 1 | 94 | 5 | 0 |
| | 冬季 | 0 | 97 | 3 | 0 |
| 梁沱断面 | 春季 | 12 | 68 | 20 | 0 |
| | 夏季 | 9 | 55 | 36 | 0 |
| | 秋季 | 12 | 65 | 23 | 0 |
| | 冬季 | 12 | 71 | 17 | 0 |
| 大溪沟断面 | 春季 | 3 | 96 | 1 | 0 |
| | 夏季 | 2 | 89 | 9 | 0 |
| | 秋季 | 3 | 94 | 3 | 0 |
| | 冬季 | 2 | 98 | 0 | 0 |

**表 5　不同温度对 3 个断面水质状态的影响占比** %

| 断面 | 最低温度/℃ | Ⅰ | Ⅱ | Ⅲ | Ⅳ |
|---|---|---|---|---|---|
| 北温泉断面 | 状态 1,[0,14) | 1 | 97 | 2 | 0 |
| | 状态 2,[14,21) | 0 | 95 | 4 | 1 |
| | 状态 3,[21,35) | 0 | 89 | 10 | 1 |
| 梁沱断面 | 状态 1,[0,14) | 12 | 72 | 16 | 0 |
| | 状态 2,[14,21) | 13 | 69 | 18 | 0 |
| | 状态 3,[21,35) | 9 | 54 | 37 | 0 |
| 大溪沟断面 | 状态 1,[0,14) | 2 | 98 | 0 | 0 |
| | 状态 2,[14,21) | 3 | 96 | 1 | 0 |
| | 状态 3,[21,35) | 2 | 88 | 10 | 0 |

**表 6　梁沱断面不同降水量对水质状态的影响占比** %

| 降水量/mm | Ⅰ | Ⅱ | Ⅲ |
|---|---|---|---|
| 状态 1,[0] | 11 | 66 | 23 |
| 状态 2,(0,5] | 11 | 65 | 24 |
| 状态 3,(5,15] | 11 | 64 | 25 |
| 状态 4,(15,30] | 11 | 65 | 24 |
| 状态 5,(30,70] | 10 | 60 | 30 |
| 状态 6,(70,140] | 11 | 62 | 27 |

由表 6 可知,当降水量≤30 mm(状态 1 ~ 状态 4)时,水质状态的概率变化很小,即降水量几乎不影响水质状态;当降水量 > 30 mm(状态 4 ~ 状态 6),降水量对水质状态影响波动较大,出现Ⅲ类水质的概率先增大后减小。分析其原因,可能与雨水的冲刷作用和稀释作用有关,一方面,降水可将环境中的灰尘、污水、垃圾带入流域,导致水质状态变差;另一方面,强降水又能稀释流域中的污染物浓度,但Ⅲ类水质的概率还是偏高[31-32]。

其次,分析水质节点对最终水质状态的影响,3 个断面水质节点的条件概率表见表 7 ~ 9。表 7 ~ 9 中的结果是当水质节点是某种组合时对水质类别进行分类的概率。与贝叶斯网络结构图不同,条件概率表除了可以看到影响水质类别的指标,还能清楚看见节点间的概率传递,节点的组合决定了流域水质的状态,每一组概率组合都有不同的水质类别概率。以北温泉断面为例,概率表中有 7 种可能的组合(表 7),当 TP 和 $NH_3$—N 的状态均为 1 时,水质各有 50% 的概率符合Ⅰ类和Ⅱ类标准;当 TP 和 $NH_3$—N 有一个指标的状态变为 2 时,水质类别 100% 为Ⅱ类;当 TP 和 $NH_3$—N 的状态均为 2 时,Ⅱ类水和Ⅲ类水的概率分别为 98% 和 2%。

表 7 北温泉断面水质节点的条件概率表 %

| TP | $NH_3$—N | 水质 | | | |
|---|---|---|---|---|---|
| | | Ⅰ | Ⅱ | Ⅲ | Ⅳ |
| 状态 1 | 状态 1 | 50 | 50 | 0 | 0 |
| 状态 2 | 状态 1 | 0 | 100 | 0 | 0 |
| 状态 1 | 状态 2 | 0 | 100 | 0 | 0 |
| 状态 2 | 状态 2 | 0 | 98 | 2 | 0 |
| 状态 1 | 状态 3 | 0 | 0 | 100 | 0 |
| 状态 2 | 状态 3 | 0 | 0 | 100 | 0 |
| 状态 2 | 状态 4 | 0 | 0 | 0 | 100 |

表 8 梁沱断面水质节点的条件概率表 %

| $COD_{Mn}$ | $NH_3$—N | TP | 水质 | | |
|---|---|---|---|---|---|
| | | | Ⅰ | Ⅱ | Ⅲ |
| 状态 1 | 状态 2 | 状态 1 | 0 | 100 | 0 |
| 状态 1 | 状态 1 | 状态 2 | 0 | 100 | 0 |
| 状态 2 | 状态 1 | 状态 2 | 3 | 97 | 0 |
| 状态 3 | 状态 1 | 状态 2 | 0 | 50 | 50 |
| 状态 1 | 状态 2 | 状态 2 | 2 | 98 | 0 |
| 状态 2 | 状态 2 | 状态 2 | 0 | 100 | 0 |
| 状态 3 | 状态 2 | 状态 2 | 0 | 33 | 67 |
| 状态 2 | 状态 1 | 状态 3 | 0 | 40 | 60 |
| 状态 3 | 状态 1 | 状态 3 | 0 | 0 | 100 |
| 状态 1 | 状态 2 | 状态 3 | 0 | 100 | 0 |
| 状态 2 | 状态 2 | 状态 3 | 0 | 63 | 37 |
| 状态 3 | 状态 2 | 状态 3 | 0 | 14 | 86 |
| 状态 3 | 状态 3 | 状态 3 | 0 | 0 | 100 |
| 状态 3 | 状态 2 | 状态 4 | 0 | 0 | 100 |

表 9 大溪沟断面水质节点的条件概率表 %

| TP | $COD_{Mn}$ | 水质 | | |
|---|---|---|---|---|
| | | Ⅰ | Ⅱ | Ⅲ |
| 状态 1 | 状态 1 | 67 | 33 | 0 |
| 状态 2 | 状态 1 | 0 | 100 | 0 |
| 状态 1 | 状态 2 | 1 | 99 | 0 |
| 状态 2 | 状态 2 | 1 | 99 | 0 |
| 状态 3 | 状态 2 | 0 | 100 | 0 |
| 状态 2 | 状态 3 | 0 | 57 | 43 |
| 状态 3 | 状态 3 | 0 | 25 | 75 |

3.3.3 水质评价结果 获得贝叶斯网络中各节点的条件概率表后,利用联结树推理引擎进行证据推理,实现最终水质评价过程。嘉陵江流域重庆段各监测断面水质综合评价结果如表 10 所示。由表 10 可知,3 个断面的水质评价类别均为Ⅱ类,但从隶属概率上可以发现,相对于北温泉和大溪沟断面,梁沱断面的水质状态较差,有 24% 的概率为Ⅲ类水质。

表 10 嘉陵江流域重庆段各监测断面不同水质类别占比及评价结果 %

| 断面 | Ⅰ | Ⅱ | Ⅲ | 水质类别 |
|---|---|---|---|---|
| 北温泉断面 | 1 | 94 | 5 | Ⅱ |
| 梁沱断面 | 11 | 65 | 24 | Ⅱ |
| 大溪沟断面 | 2 | 94 | 4 | Ⅱ |

## 4 讨 论

基于流域水质指标分析(图 3)发现,与大多数流域水质序列相似,嘉陵江流域重庆段水质序列存在着数据缺失、冗余、动态、不完整和不确定的特点,且受多种外环境因素的共同影响,综合评价难度较大。粗糙集(RS)理论作为一种新的处理不完整知识的数学方法,能在保留关键信息的前提下对数据进行约简;然而,如果约简后的信息存在重要属性缺失时,RS 理论将不能很好地开展属性约简和分类分析,决策效率相对较低。贝叶斯网络(BN)具有严格的概率论基础,其优秀的不确定性处理能力可以很好地解决上述问题,但是当属性较多时存在计算规模较大的问题。因此,基于智能互补的思想,应用 RS 理论降低 BN 模型的复杂度的同时,应用 BN 模型解决人工智能领域解释性差的问题,提出了粗糙集 - 贝叶斯网络(RS - BN)的流域水质评价方法,即使用 RS 理论发现水质时间序列规则并找到最小属性集,采用最小属性集进行 BN 建模得到网络结构和条件概率表,最终实现水质的证据推理并获得综合水质类别。

为了更好地验证本研究方法的可靠性,选用贝叶斯网络(BN)、灰色 - 贝叶斯网络(grey relation analysis - Bayesian network, GRA - BN)、粗糙集 - 朴素贝叶斯(rough set - naive Bayes, RS - NB)对相同数据集进行建模和比较分析。BN 和 GRA - BN 模型采用与 RS - BN 相同的贝叶斯网络学习和贝叶斯网络推理方法,主要差别体现在模型输入,BN 模型没有对输入属性进行约简,3 个断面均采用所有内、外环境因素建模;GRA - BN 模型采用灰色关联性进行属性约简,并选择关联程度大于 0.95 的属性进行建模。RS - NB 模型采用粗糙集属性约简集进行建模,北温泉、梁沱、大溪沟 3

个断面的 RS - NB 模型参数分别为 9、10、9。

4 种水质评价方法的性能指标对比见表 11。表 11 显示:(1)与 BN 模型相比,RS - BN 方法的四个指标均有提升,说明粗糙集理论可以有效降低建模的冗余信息和不完整信息,提高模型效率和精度;(2)与 GRA - BN 方法相比,RS - BN 方法拥有更好的预测效果,说明粗糙集理论拥有比灰色理论更好的信息属性约简能力;(3)与 RS - NB 方法相比,RS - BN 方法具有更好的水质评价性能,凸显了 BN 模型在不确定问题处理和概率推理方面的能力。所以,本文提出的 RS - BN 方法获得了最高的准确率、精确率、召回率和 *F*1 值,水质评价效果最好。本研究很好地印证了前人提出的先约简、后预测的建模思路。如吴雨辰等[17]对 9·6 北海道地震滑坡易发性评价时发现,RS 约简后神经网络模型的准确性将提高 47.96%。此外,RS - BN 方法应用于三个断面的模拟结果一致,均具有最高的模拟精度,相对而言,其他 3 种方法在不同断面表现出不同的模拟效果,说明RS - BN 方法具有更好的普适性,能够在其他断面或流域的水质评价中推广使用。

**表 11　4 种水质评价方法的性能指标对比**

| 断面 | 评价指标 | RS - BN | BN | GRA - BN | RS - NB |
|---|---|---|---|---|---|
| 北温泉断面 | 准确率 | 0.9884 | 0.9614 | 0.9690 | 0.9693 |
| | 精确率 | 0.8729 | 0.3275 | 0.5350 | 0.7369 |
| | 召回率 | 0.9656 | 0.3704 | 0.5145 | 0.8584 |
| | F1 值 | 0.9169 | 0.5379 | 0.5245 | 0.7930 |
| 梁沱断面 | 准确率 | 0.9733 | 0.9656 | 0.9656 | 0.9513 |
| | 精确率 | 0.9296 | 0.2994 | 0.5988 | 0.4719 |
| | 召回率 | 0.9569 | 0.5759 | 0.5759 | 0.5771 |
| | F1 值 | 0.9430 | 0.7018 | 0.5871 | 0.6756 |
| 大溪沟断面 | 准确率 | 0.9844 | 0.9450 | 0.9450 | 0.9469 |
| | 精确率 | 0.8669 | 0.3150 | 0.3150 | 0.6003 |
| | 召回率 | 0.8669 | 0.3333 | 0.3333 | 0.5479 |
| | F1 值 | 0.8669 | 0.4928 | 0.3239 | 0.5729 |

## 5　结　论

针对水质时间序列的不确定性问题,本文提出了 RS 与 BN 相结合的流域水质综合评价模型,通过嘉陵江流域重庆段 3 个监测断面水质指标实例研究得到如下结论:

(1)利用 RS 挖掘隐含在数据中的关键信息,获得 3 个断面水质的属性约简集,避免了数据缺失、不确定和动态性带来的影响,降低了建模复杂度。

(2)利用 RS 属性约简集进行 BN 建模,获得 3 个断面的贝叶斯网络结构和条件概率表,找到了对流域水质状况影响较大的内、外环境因素。

(3)RS - BN 方法既考虑了 RS 的属性约简功能,又考虑了 BN 的不确定性概率表达能力,能在最小准则下准确给出影响因子与水质指标之间的因果关系。

(4)RS - BN 方法模拟结果令人满意,4 种评价指标均优于现有评价方法(BN、GRA - BN、RS - NB),适用于水质快速评价,为流域水质评价提供了一种新思路。

**参考文献:**

[1] 刘录三, 黄国鲜, 王 璠, 等. 长江流域水生态环境安全主要问题、形势与对策[J]. 环境科学研究, 2020, 33(5): 1081 - 1090.

[2] HUANG Jiacong, ZHANG Yinjun, BING Haijian, et al. Characterizing the river water quality in China: Recent progress and on-going challenges[J]. Water Research, 2021, 201: 117309.

[3] 刘 洁, 张丰帆, 赵 泞, 等. 基于改进遗传算法的河流水污染源反演方法[J]. 环境科学学报, 2020, 40(10): 3598 - 3604.

[4] 黄 诚, 黄廷林, 李 扬, 等. 金盆水库暴雨径流时空演变过程及水质评价[J]. 环境科学, 2021, 42(3): 1380 - 1390.

[5] 傅 博, 黄国如. 广东江门四堡水库水质时空变化及综合评价研究[J]. 水资源与水工程学报, 2019, 30(5): 64 - 71.

[6] NONG Xizhi, SHAO Dongguo, ZHONG Hua, et al. Evaluation of water quality in the South - to - North Water Diversion Project of China using the water quality index (WQI) method[J]. Water Research, 2020, 178: 115781.

[7] 董 杰, 李 欣, 方运海, 等. 基于改进模糊综合 - 指数平滑法的地下水水质评价和预测[J]. 中国海洋大学学报(自然科学版), 2020, 50(1): 126 - 135.

[8] 张秀菊, 王柳林, 李秀平, 等. 基于 BP 神经网络的潇河流域水质预测[J]. 水资源与水工程学报, 2021, 32(5): 19 - 26.

[9] 张 婷, 王学雷, 耿军军, 等. 基于 MIKE21 和灰色模式识别模型的洪湖水质模拟与评价[J]. 长江流域资源与环境, 2018, 27(9): 2090 - 2100.

[10] 杨 咪, 屈文岗, 钱 会. 基于熵权的贝叶斯模型及其在水质评价中的应用[J]. 灌溉排水学报, 2018, 37(1): 85 - 90.

[11] LIU Shuci, RYU D, WEBB J A, et al. A Bayesian approach to understanding the key factors influencing temporal variability in stream water quality—A case study in the Great Barrier Reef catchments[J]. Hydrology and Earth System Sciences, 2021, 25(5): 2663 - 2683.

[12] LI R A, MCDONALG J A, SATHASIVAN A, et al. A multivariate Bayesian network analysis of water quality factors influencing trihalomethanes formation in drinking water distribution systems[J]. Water Research, 2021, 190: 116712.

[13] 陆 丹, 耿昭克, 闵 敏, 等. 基于区间型贝叶斯模型的湟水干流水质评价[J]. 水利水运工程学报, 2020(2): 15 - 21.

[14] WANG Guoqiang, LI Tianrui, ZHANG Pengfei, et al. Double-local rough sets for efficient data mining[J]. Information Sciences, 2021, 571: 475 - 498.

[15] ZHANG Wanjuan, WANG Xiaodan, CABRERA D, et al. Product quality reliability analysis based on rough Bayesian network[J]. International Journal of Perfomability Engineering, 2020, 16(1): 37 - 47.

[16] 卢 媛, 王 栋. 改进的粗糙集 - 集对分析的水质评价方法[J]. 华北水利水电大学学报(自然科学版), 2019, 40(1): 34 - 38.

[17] 吴雨辰, 周晗旭, 车爱兰. 基于粗糙集 - 神经网络的 IBURI 地震滑坡易发性研究[J]. 岩石力学与工程学报, 2021, 40(6): 1226 - 1235.

[18] ZHANG Pengfei, LI Tianrui, WANG Guoqiang, et al. Multi-source information fusion based on rough set theory: A review[J]. Information Fusion, 2021, 68: 85 - 117.

[19] XIAO Yanjun, ZHANG Heng, ZHOU Wei, et al. Research on the fault diagnosis method for high-speed loom using rough set and Bayesian network[J]. Journal of Intelligent & Fuzzy Systems, 2020, 39(1): 1147 - 1161.

[20] XU Bin, HU Jun, HU Ting, et al. Quantitative assessment of seismic risk in hydraulic fracturing areas based on rough set and Bayesian network: A case analysis of Changning shale gas development block in Yibin City, Sichuan Province, China[J]. Journal of Petroleum Science and Engineering, 2021, 200: 108226.

[21] 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: 中国环境科学出版社, 2002.

[22] ANIS M Z. Determination of the best mean fill[J]. Quality Engineering, 2003, 15(3): 407 - 409.

[23] PAWLAK Z, SKOWRON A. Rudiments of rough sets [J]. Information Sciences, 2007, 177(1): 3 - 27.

[24] PEARL J. Probabilistic reasoning in intelligent systems [M]. San Francisco: Morgan Kaufmann, 1988.

[25] 程卫国, 李亚斌, 苏 燕, 等. 不同赋权方法的综合水质标识指数法对比分析[J]. 灌溉排水学报, 2019, 38(11): 93 - 99.

[26] SULAIMAN N S, BAKAR R A. Rough set discretization: Equal frequency binning, Entropy/MDL and semi naives algorithms of intrusion detection system[J]. Journal of Intelligent Computing, 2017, 8(3): 91 - 97.

[27] HUANG Wencheng, KOU Xingyi, ZHANG Yue, et al. Operational failure analysis of high-speed electric multiple units: A Bayesian network - K2 algorithm - expectation maximization approach [J]. Reliability Engineering & System Safety, 2021, 205: 107250.

[28] CHAUDHARY A, HANTUSH M M. Bayesian Monte Carlo and maximum likelihood approach for uncertainty estimation and risk management: Application to lake oxygen recovery model [J]. Water Research, 2017, 108: 301 - 311.

[29] WU Jiagao, GUO Yahang, ZHOU Hongyu, et al. Vehicular delay tolerant network routing algorithm based on Bayesian network[J]. IEEE Access, 2020, 8: 18727 - 18740.

[30] 续衍雪, 吴 熙, 路 瑞, 等. 长江经济带总磷污染状况与对策建议[J]. 中国环境管理, 2018, 10(1): 70 - 74.

[31] 梁 爽, 陈 敏, 肖尚斌, 等. 鄂西长江喀斯特小流域氮磷输出特征[J]. 长江流域资源与环境, 2021, 30(10): 2471 - 2481.

[32] 包 鑫, 江 燕, 胡羽聪. 潮河流域降雨径流事件污染物输出特征[J]. 环境科学, 2021, 42(7): 3316 - 3327.