

doi:10.3969/j.issn.2097-0013.2025.02.004

三峡库区三门洞滑坡监测数据挖掘

黄 珏^{1,2}, 叶润青^{1,2}, 付小林^{1,2}, 代贞伟^{1,2}, 吴 震^{1,3}, 陈 瑶^{1,3}

HUANG Jue^{1,2}, YE Run-Qing^{1,2}, FU Xiao-Lin^{1,2}, DAI Zhen-Wei^{1,2}, WU Zhen^{1,3}, CHEN Yao^{1,3}

1. 中国地质调查局武汉地质调查中心(中南地质科技创新中心), 湖北 武汉 430205; 2. 高坝大库运行安全湖北省重点实验室(中国长江三峡集团有限公司), 湖北 宜昌 443133; 3. 中国地质大学(武汉)地质调查研究院, 湖北 武汉 430074
1. Wuhan Center, China Geological Survey (Geosciences Innovation Center of Central South China), Wuhan 430205, Hubei, China;
2. Hubei Key Laboratory of Operation Safety of High Dam and Large Reservoir (China Three Gorges Corporation), Yichang 443133, Hubei, China; 3. Institute of Geological Survey, China University of Geosciences (Wuhan), Wuhan 430074, Hubei, China

摘要: 基于传统的关联规则挖掘方法分析三峡库区地质灾害监测大数据时,需要通过多次扫描数据才能挖掘出频繁项集,存在速度较慢、效率低的问题。本文采用 FP-Growth 算法,以三门洞滑坡为例,选择降雨和库水位两类影响因素,细分出 6 种影响因子,与滑坡位移监测数据进行关联规则挖掘。筛选出 8 条关联规则结果并对其进行分析,发现了滑坡位移与大气降雨、库水位之间存在较大的相关性。降雨量越大,连续降雨时间越长,滑坡位移越大;单位时间内库水位变动越大,滑坡位移越大,库水位下降对滑坡位移产生的影响比库水位上升要大。研究表明,FP-Growth 算法较传统关联规则挖掘方法效率更高,更加适合大数据分析,可提供大数据中隐含的、可理解的、有价值的信息,挖掘出的监测数据关联规则有助于滑坡变形分析。针对三峡库区地质灾害防治数据具有多源、异构、海量的特点,应更多的应用大数据分析技术,以发现数据中的规律,指导地质灾害防治工作。

关键词: 关联规则; FP-Growth 算法; 大数据分析; 三门洞滑坡; 三峡库区

中图分类号: P642.22

文献标识码: A

文章编号: 2097-0013(2025)-02-0288-07

Huang J, Ye R Q, Fu X L, Dai Z W, Wu Z and Chen Y. 2025. Mining of Monitoring Data for Sanmendong Landslide in the Three Gorges Reservoir Area. *South China Geology*, 41(2): 288–294.

Abstract: When traditional association rule mining methods is used to analyze geological hazard monitoring big data in the Three Gorges Reservoir area, multiple scanning of the data is required to mine frequent itemsets, resulting in slow speed and low efficiency. By adopting the FP-Growth algorithm, and with the Sanmendong landslide as an example, two types of influencing factors, i.e. rainfall and reservoir water level, are selected, and six influencing factors are further subdivided for association rule mining with landslide displacement monitoring data. Eight association rule results are screened out and analyzed to find that there is a significant correlation between landslide displacement and atmospheric rainfall as well as reservoir water level, i.e. the greater the rainfall is, the longer the continuous rainfall time lasts, the greater the landslide displacement is; the greater the change of the reservoir water level per unit time is, the greater the landslide displacement is. The impact of a decrease in reservoir water level on landslide displacement is greater than

收稿日期: 2025-03-31; 修回日期: 2025-04-30

基金项目: 水利部国家重大水利工程建设基金三峡后续工作地质灾害防治项目(000121 2024C C60 003)

第一作者: 黄珏(1984—), 男, 工程师, 主要从事三峡库区地质灾害监测预警相关研究工作, E-mail: huangjue@mail.cgs.gov.cn

that of an increase in reservoir water level. The research results indicate that, being more efficient than traditional association rule mining methods and more suitable for big data analysis, the FP-Growth algorithm can provide implicit, comprehensible and valuable information in big data, and the association rules mined from monitoring data are helpful for landslide deformation analysis. Given the characteristics of multi-source, heterogeneous, and massive geological disaster prevention and control data in the Three Gorges Reservoir area, big data analysis techniques should be more widely applied in geological disaster prevention and control work in the Three Gorges Reservoir area.

Key words: association rules; FP-Growth algorithm; big data analysis; Sanmendong landslide; the Three Gorges Reservoir area

三峡库区地质灾害自动化专业监测网已于 2016 年基本完成建设,专业监测设备、普适性监测设备已运转多年。大量不同类型的监测数据通过物联网传输,地质灾害监测频次大幅提升,逐渐形成了多源、异构、海量的地质灾害防治大数据(许强, 2020; 贾琰棋等, 2023)。2024 年三峡库区共有 257 处专业监测点, 2732 台普适型监测设备, 948 处气象观测站点, 全年完成入库数据量约为 1.12 亿条。

不少学者开展了基于大数据的地质灾害数据关联规则挖掘。关联规则挖掘是基于数据挖掘技术,在数据集里面挖掘出隐含的、可理解的、有效的信息(尚敏等, 2021; 朱鸿鹄等, 2022),发现数据集中某些属性同时出现的规律和模式,从而得到不同属性的关联关系。数据挖掘中常用的关联规则挖掘方法包括 Apriori 算法、FP-growth 算法。吴婷和牛瑞卿(2011)选取灾害点的灾害类型、灾害规模、灾害体的物质类型、高程差、水系岸别 5 个属性及叠加分析结果利用 Apriori 算法进行关联规则挖掘;段功豪等(2014)、马俊伟等(2019)利用动态监测对滑坡相关监测指标进行了充分的挖掘;孙义杰(2015)采用 Apriori 算法,分析了马家沟滑坡变形位移与各影响因子间的关联规则;Huang H F et al. (2016)分析了三峡库区地质灾害与降雨、库水位之间的关联规则;杨柯等(2020)利用 Apriori 算法分析第四系分布及厚度,提出了多源数据融合的第四系特征关联规则分析方法,吕梅洁(2024)采用 Apriori 算法对白水河滑坡进行关联规则挖掘。

传统关联规则挖掘 Apriori 算法需要通过构

造候选集,多次扫描原始数据才能挖掘出频繁项集,导致系统运转速度较慢(余飞娅和叶文波, 2021),算法自身产生的数据内存也会导致系统 I/O 过载,效率得不到保证(宣正邦等, 2022)。本文采用 FP-growth 算法通过将事务集存储到频繁模式树(Frequent Pattern Tree)中,使用该算法只需要对数据集进行两次扫描(李瑞等, 2024; 马健等, 2024),很大程度上加快了算法速度,有效解决了 Apriori 算法效率低的问题。

本文通过收集三门洞滑坡已有监测数据资料,对数据进行预处理,选取影响滑坡变形的主要诱发因子,对滑坡位移及因子数据进行聚类分析。在此基础上采用 FP-growth 算法开展关联规则挖掘:首先通过扫描数据集统计每个项的频率构建 FP-tree,再从 FP-tree 中挖掘频繁项集,从而获得关联规则结果。

1 关联规则挖掘

关联规则挖掘表达形式为: $A \rightarrow B$, 其中项集 A 为先决条件,即前项;项集 B 为对应关联结果,即后项。根据关联规则挖掘步骤,对三门洞滑坡监测数据进行挖掘分析:

1.1 关联规则挖掘步骤

(1)数据收集:收集已有监测数据资料,从监测项、监测频率等方面进行分类梳理归整,如地部位移、库水位、降雨量等数据。

(2)数据预处理:地质灾害监测数据存在误差以及由数据传输或设备问题导致的数据缺失(袁乾博等, 2020; 崔芳姿等, 2021),对滑坡监测大数

据进行预处理,是数据分析的前提。

数据分析前期的预处理工作非常重要,地质灾害监测数据可能存在数据异常的问题。若不对异常数据进行处理,这些“脏数据”会对数据挖掘的过程及结果产生负面影响,因此数据的预处理对于后续的挖掘效率是非常有帮助的,保证了数据的准确性、完整性、一致性,为下一步挖掘分析工作打好基础。

(3)聚类分析:把连续型数据转换成离散型数据。结合监测数据的特点,本文采用 K-means 算法将连续型数据转换成离散型数据,在后续数据挖掘分析中使用(刘鹏程,2018; 赵恩毅等, 2024)。

(4)关联规则挖掘:

本文采用 FP-Growth 算法对三门洞滑坡监测数据进行挖掘分析,生成频繁项集得到关联规则,计算支持度、置信度、提升度并对规则进行筛选(胡友鑫等, 2025),从而得到有效的关联规则结果。

1.2 FP-Growth 算法

FP-Growth 算法发现频繁项集的基本过程为首先构建 FP-tree,再从 FP-tree 中挖掘频繁项集(陈波等, 2022; 余欣媛, 2024)。

(1)构建 FP-tree: 扫描数据集(图 1a),根据事件

相关数据计算支持度:如 d 支持度为 1,z 支持度为 5(图 1b),删除支持度小的项:如删除支持度小于 3 的项(图 1c),按计数排序(图 1d),整理数据集(图 1e)。

通过扫描数据集统计每个项的频率,筛选出频繁项及其支持度,然后构建 FP-tree。FP-tree 的每个节点表示一个项,节点之间的链接表示项之间的关联。最终得到 FP-tree(图 2)。

(2)从 FP-tree 中挖掘频繁项集:例如,在上述 FP 树中,寻找以 s 结尾的频繁项集。首先找到所有包含 s 结点的路径,即 s 的前缀路径(图 3)。计算结点 s 关联的支持度计数,如果大于最低支持度,则 s 是频繁项集。由于 s 是频繁的,所以要解决发现以 s 结尾的其它频繁项集的问题。从 s 的条件 FP 树上收集其它项的前缀路径,并计算其频度计数,得到对应项集的支持度计数,进而判断其频繁性。以此类推,从 FP-tree 中挖掘频繁项集,在此基础上获得关联规则结果。

2 三门洞滑坡概况

2.1 滑坡概况

三门洞滑坡位于秭归县沙镇溪镇梅坪村,距

a							
1.	q	z	g	i	o		
2.	z	y	x	v	u	t	s
3.	z						
4.	q	x	m	n	r		
5.	y	q	x	z	p	s	o
6.	y	z	x	d	p	r	s

d	1
g	1
i	1
l	1
n	1
m	1
p	2
o	2
r	3
q	3
t	1
s	3
v	1
u	1
y	3
x	4
z	5

z	5
x	4
y	3
s	3
q	3
r	3

1.	z	q			
2.	z	x	y	r	s
3.	z				
4.	x	r	q		
5.	z	x	y	q	s
6.	z	x	y	r	s

图 1 整理数据集

Fig. 1 Organize the dataset

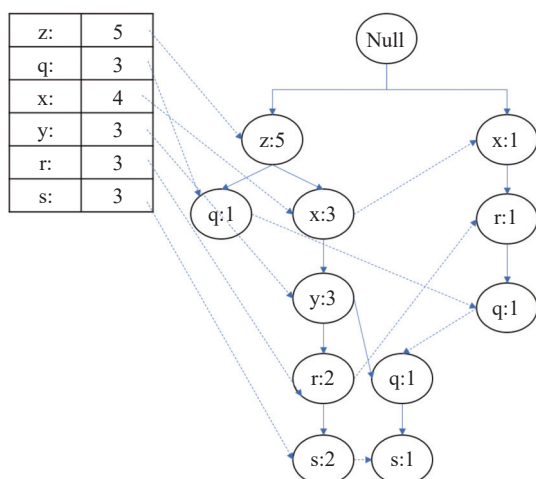


图 2 频繁模式树

Fig. 2 Frequent pattern tree

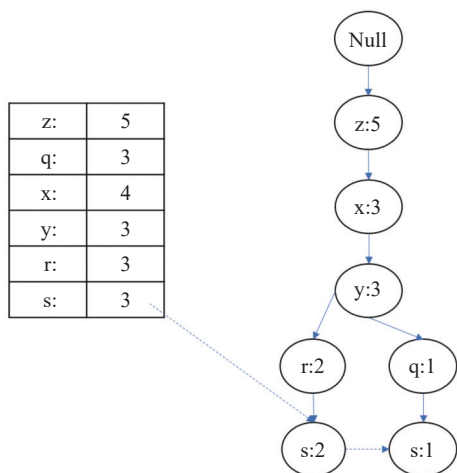


图 3 S 的前缀路径图

Fig. 3 Prefix-path of S

三峡支流青干河入江口约 8.5 km, 青干河右岸(易庆林等, 2018; 谢林冲等, 2020), 位于青干河右岸山坡上, 前缘直抵青干河, 剪出口高程 125 m, 后缘高程 430 m。滑坡体纵长约 830 m, 均宽 300 m, 滑坡总面积约 $24.9 \times 10^4 \text{ m}^2$, 滑坡体平均厚度约 20 m, 滑坡总体积约 $548 \times 10^4 \text{ m}^3$ 。滑坡坡面平均坡度 15° , 主滑方向 61° 。滑坡平面形态呈舌形状(图 4)。

三门洞滑坡滑体由 $Q^{\text{col+dl}}$ 构成, 崩坡积物为碎块石土。块石成分主要为石英砂岩、泥岩。滑带土为浅灰色粉质粘土夹碎石, 可塑-软塑状态, 湿, 稍密。碎石成分为浅灰绿色砂岩、泥岩, 粒径 0.5 ~ 2 cm 不等, 强风化, 多为次棱角状, 少数具一定磨圆度, 土石比为 8 : 2。

2.2 监测网点布置情况

2006 年 9 月在滑坡体北东方向 2 条相互基本平行的监测纵剖面上布设 6 个 GPS 监测点。2022 年监测点优化升级后, 在人工 GPS 监测点 ZG361、ZG363、ZG364 和 ZG365 旁分别新建 4 个 GNSS 自动监测点 (ZGX361、ZGX363、ZGX364 和 ZGX365), 基点共用三门洞滑坡东南侧的卧沙溪滑坡的基点。各监测点布置图详见图 4。

3 三门洞滑坡监测数据分析

3.1 因子选取

结合以往相关研究经验, 选取影响滑坡变形的的主要诱发因子, 并对其进行细分(李明亮等, 2024; 杨诗诗等, 2023)。根据三门洞滑坡现有监测数据, 选取了降雨和库水位两类影响因素, 并在这两类影响因素基础上, 细分出 6 种影响因子作为关联规则的前项, ZGX364 监测点位于滑坡中部, 受外界干扰相对较小, 异常数据较少, 故选取滑坡监测点 ZGX364 累计位移量(月)作为关联规则的后项。

影响因子分别为: 降雨量累计值(月)、库水位平均值(月)、降雨量日最大值(月)、连续降雨量最大值(月); 库水位上升最大值(月)、库水位下降最大值(月)。具体如表 1 所示:

3.2 聚类分析

根据关联规则挖掘步骤, 采用 K-means 算法将关联规则的前项、后项转换成离散型数据并划分等级。降雨量累计值(月)(R_m)、降雨量日最大值(月)($R_{d-\max}$)、连续降雨量最大值(月)($R_{m-\max}$)、库水位平均值(月)(W_d)、库水位上升最大值(月)($W_{m-\text{up}}$)、库水位下降最大值(月)($W_{m-\text{down}}$)、监测点累计位移量(月)对应定性化值及分级区间见表 2。

3.3 关联规则挖掘

采用 FP-Growth 算法对影响因子和监测点地表位移聚类结果进行关联规则分析, 设置最小支持度为 30%, 最小置信度为 70%, 在众多关联规则结果中再选择提升度大于 0.9 的规则, 经过筛选后得到 8 条关联规则, 见表 3。

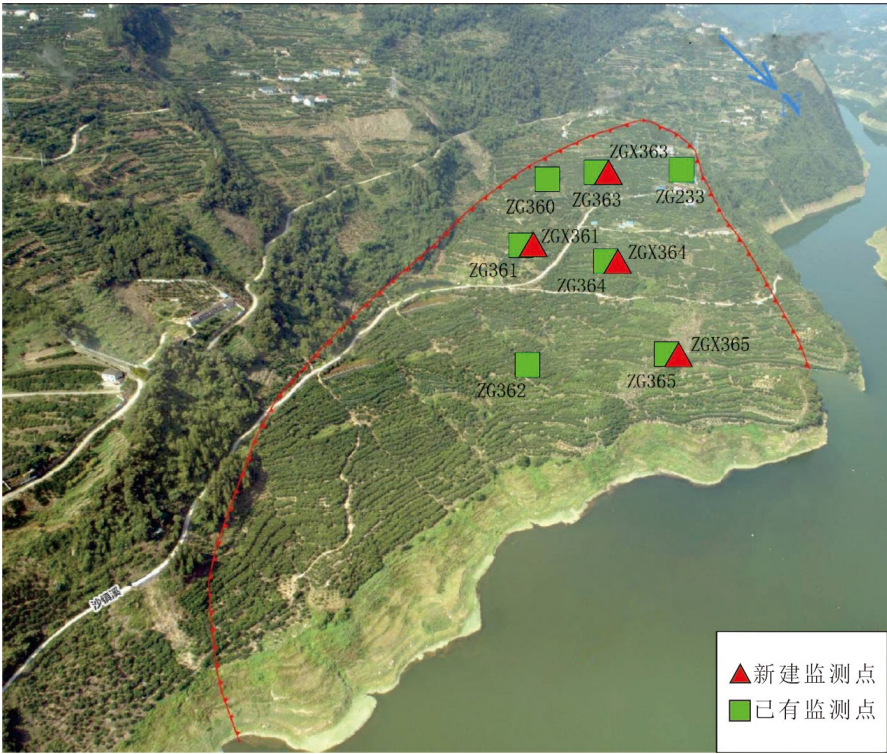


图 4 三门洞滑坡及监测点布置图

Fig. 4 Location and layout of monitoring points of Sanmendong landslide

表 1 影响因素

Table 1 Influencing factors

影响因素	影响因子	符号	单位
降雨	降雨量累计值(月)	R_m	mm
	降雨量日最大值(月)	R_{d-max}	mm
	连续降雨量最大值(月)	R_{m-max}	mm
库水位	库水位平均值(月)	W_d	m
	库水位上升最大值(月)	W_{m-up}	m
	库水位下降最大值(月)	W_{m-down}	m

结果显示: 规则 1—3 为关联规则后项监测点月累计位移细微变形阶段, 由关联规则结果可知, 当降雨量日最大值(月)达到 21.6 ~ 45.3 mm, 同时连续降雨量最大值(月)达到 29.6 ~ 91.2 mm, 监测点附近会发生细微变形。

规则 4—6 为关联规则后项监测点月累计位移较小变形阶段, 由关联规则结果可知, 中等程度的连续降雨同时伴有库水位下降会引起监测点附近发生较小变形。

规则 7—8 为关联规则后项监测点月累计位移较大变形阶段, 由关联规则结果可知, 单位时间内库水位大幅下降叠加连续降雨或强降雨会对滑坡稳定性造成较大的影响。

表 2 聚类分析结果

Table 2 Cluster analysis results

影响因子	定性化值	分级区间
$R_m(mm)$	大	(167.6, 477.8)
	中	(62.2, 162.3)
	小	(3.7, 59.5)
$R_{d-max}(mm)$	大	(45.9, 137.8)
	中	(21.6, 45.3)
	小	(1.1, 21.2)
$R_{m-max}(mm)$	大	(93.2, 200.3)
	中	(29.6, 91.2)
	小	(1.7, 28.2)
$W_d(m)$	大	(166.8, 174.5)
	中	(155.7, 165.3)
	小	(145.3, 153.9)
$W_{m-up}(m)$	大	(13.7, 19.9)
	中	(5.3, 10.6)
	小	(0.6, 4.5)
$W_{m-down}(m)$	大	(14.5, 19.6)
	中	(8.7, 12.6)
	小	(1.2, 6.6)
累计位移量(月)(mm)	Ⅲ	(118, 151.5)
	Ⅱ	(43, 85)
	Ⅰ	(0, 26)

表 3 关联规则结果

Table 3 Association rule results

规则ID	前项	后项	支持度(%)	置信度(%)	提升度(%)
1	$R_{d-max}=中 \& R_{m-max}=中$	I	47.5	91	1.16
2	$R_m=中 \& R_{d-max}=高$	I	45.6	87	1.20
3	$R_m=中 \& W_d=中$	I	43.1	83	1.21
4	$R_{d-max}=中 \& W_d=低 \& W_{m-down}=中$	II	41.8	100	1.23
5	$R_{m-max}=中 \& W_{m-down}=中$	II	39.6	100	1.26
6	$R_{d-max}=高 \& R_{m-max}=中 \& W_d=中$	II	36.2	93	1.28
7	$R_m=高 \& W_d=低 \& W_{m-down}=高$	III	31.3	79	1.31
8	$R_{m-max}=高 \& W_d=低$	III	33.5	71	0.96

4 讨论与结论

本文对三门洞滑坡降雨、库水位、地表位移监测数据进行关联规则挖掘:

(1)在降雨和库水位两类影响因素基础上, 细分出影响因子: 降雨量累计值(月)、降雨量日最大值(月)、连续降雨量最大值(月)、库水位平均值(月)、库水位上升最大值(月)、库水位下降最大值(月)作为关联规则的前项, 将监测点累计位移量(月)作为关联规则的后项。通过聚类分析, 将关联规则的前项、后项转换成离散型数据并划分等级。

(2)采用 FP-Growth 算法对影响因子和地表位移进行关联规则挖掘, 筛选出 8 条关联规则结果, 通过关联规则 1、2 可以发现, 滑坡位移与大气降雨之间存在较大的相关性, 降雨量越大, 连续降雨时间越长, 滑坡位移越大。同时, 滑坡位移与库水位也存在较大的相关性, 三峡水库正常蓄水后, 水位在 145~175 m 间浮动。通过关联规则 4、7 可以发现, 库水位下降对滑坡位移产生的影响比库水位上升要大, 单位时间内库水位降幅越大, 滑坡位移越大。若库水位快速下降同时叠加连续强降雨, 会加剧滑坡的变形, 更加影响滑坡的稳定性。

(3)根据关联规则分析结果, 每年 5~8 月汛期, 容易出现降雨和库水位两类影响因素叠加现象, 会对滑坡稳定性有较大影响, 在此期间需加强滑坡防治工作, 以防险情发生。

(4)通过数据挖掘得到的关联规则结果为滑

坡变形分析提供了数据分析结果依据, 关联规则分析可广泛应用于地质灾害分析与预警中。针对三峡库区地质灾害防治数据多源、异构、海量的特点, 大数据分析技术方法应更多的应用在三峡库区地质灾害防治工作中。

参考文献:

賁琰棋, 易武, 黄晓虎, 魏兆亨, 肖宇煌, 邓欣雨. 2023. 基于有效降雨量的“阶跃型”滑坡递进式预警模型研究 [J]. 华南地质, 39(3): 492-501.

陈波, 詹明强, 黄梓莘. 2022. 基于关联规则的库岸边坡监测数据挖掘方法 [J]. 长江科学院院报, 39(8): 58-64.

崔芳姿, 吴斌, 史学磊, 范振林, 任涛. 2021. 地质灾害监测数据综合处理与分析 [J]. 地理空间信息, 19(7): 88-92+8.

段功豪, 牛瑞卿, 咬登魁. 2014. 基于关联规则算法的滑坡判据动态提取研究 [J]. 安全与环境工程, 21(6): 47-52.

朱鸿鹄, 王佳, 李厚芝, 叶霄, 施斌, 张勤. 2022. 基于数据挖掘的三峡库区特大滑坡变形关联规则研究 [J]. 工程地质学报, 30(5): 1517-1527.

胡友鑫, 林茂彦, 罗剪秋, 陈超, 黄金煜. 2025. 复杂网络高阶结构的关联规则挖掘及其应用 [J]. 电子科技大学学报, 54(1): 152-160.

吕梅洁. 2024. 基于关联规则的滑坡监测数据挖掘研究 [D]. 河北地质大学硕士学位论文, 43-54.

刘鹏程. 2018. 大数据分析技术在地质灾害系统中的应用研究 [D]. 西安工业大学硕士学位论文, 10-16.

李明亮, 吕梅洁, 侯梦媛, 朱昊. 2024. 基于 Eclat 算法的八字门滑坡变形因素关联性分析 [J]. 长江科学院院报, 41(6): 150-155.

李瑞, 刘珊, 闫磊. 2024. 基于 FP-Growth 算法的新能源配电网 CPS 网络攻击检测方法 [J]. 电信科学, 40(11): 103-113.

- 马 健,谢雨晴,张丽岩,王 燕,周欢生.2024.基于 FP-growth 算法的交通事故数据关联规则挖掘研究 [J]. 科技创新与生产力,45(9):95-97.
- 马俊伟,唐辉明,邹宗兴,刘 晓,温 韬,王艳昆. 2019. 滑坡位移预测的智能算法与应用实例 [M]. 武汉:中国地质大学出版社,23-66.
- 余欣媛.2024.基于 FP-Growth+LSTM 的问题发现及行为预测模型 [J]. 科技与创新,(11):10-14.
- 尚 敏,廖 芬,马 锐,刘昱廷.2021.白家包滑坡变形与库水位、降雨相关性定量化分析研究 [J]. 工程地质学报,29(3):742-750.
- 孙义杰. 2015. 库岸边坡多场光纤监测技术与稳定性评价研究 [D]. 南京大学博士学位论文,90-123.
- 谢林冲,王明清,周 超.2020.基于降雨-库水联合作用下三门洞滑坡演化过程与机制 [J]. 水电能源科学,38(6):128-132.
- 吴 婷,牛瑞卿.2011.三峡库区巴东段地质灾害因子关联规则挖掘 [J]. 计算机工程与应用,47(11):240-244.
- 许 强.2020.对滑坡监测预警相关问题的认识与思考 [J]. 工程地质学报,28(2):360-374.
- 宣正邦,黄树成,王 逊.2022.一种基于 Hadoop 的 Apriori 改进算法 [J]. 计算机与数字工程,50(3):459-462.
- 余飞娅,叶文波.2021.基于 FP-Growth 算法的计量主站告警分析研究 [J]. 电气自动化,43(6):30-32+35.
- 杨 柯,叶润青,付小林,牛瑞卿.2020.三峡库区第四系特征关联规则分析 [J]. 人民长江,51(6):118-124.
- 杨诗诗,叶润青,付小林,吴润泽,熊 能,文天龙.2023.三峡库区降雨型滑坡预警雨量阈值研究 [J]. 华南地质,39(3):445-454.
- 袁乾博,肖诗荣,李春霞,夏 宇,王 涛.2020.基于 GLSSVM 模型的三门洞滑坡变形预测研究 [J]. 人民长江,51(7):130-135.
- 易庆林,覃世磊,文 凯,周 宝,曾怀恩.2018.三峡库区三门洞滑坡变形特征与影响因素分析 [J]. 三峡大学学报 (自然科学版),40(5):41-46.
- 赵恩毅.2024.大数据中的数据清洗与预处理技术研究 [J]. 信息记录材料,25(3):195-197.
- Huang H F, Yi W, Lu S Q, Yi Q L, Zhang G D. 2016. Use of monitoring data to interpret active landslide movements and hydrological triggers in Three Gorges reservoir[J]. Journal of Performance of Constructed Facilities, 30(1): 1-11.