DOI:10.13409/j.cnki.jdpme.20231217002

基于优化FP-Growth算法的滑坡频繁因素组合挖掘*

李佳颖[1]，郝彬超[1]，王卫东[2]，王智超[1]，曹禄来[3]，韩 征[2]，朱崇政[1]

（1.湘潭大学土木工程学院，湖南 湘潭 411105；2.中南大学土木工程学院，湖南 长沙 410083；
3.创辉达设计股份有限公司，湖南 长沙 410000）

摘要：滑坡影响因素复杂多样，挖掘滑坡的频繁因素组合能宏观快速地初步判识滑坡易发区域。以四川省凉山彝族自治州内586处滑坡灾害为样本数据，从地质条件、水文条件、地形条件、气象条件和人类工程活动五个方面收集12个滑坡影响因素，基于卡方检验剔除与滑坡灾害弱相关的影响因素，耦合分析滑坡区域与影响因素区划，针对大数据挖掘算法仅能以历史滑坡次数等离散型变量为挖掘依据的局限性，引入特征参数优化频繁模式树（FP-Growth）算法，使其能以历史滑坡面积和历史滑坡密度等连续型变量为挖掘依据，挖掘滑坡频繁二级因素组合，利用卡方检验与频率比检验挖掘结果准确性。结果表明：基于历史滑坡密度的优化关联规则算法能更好地挖掘滑坡频繁二级因素组合，其中，“高程＜1 769 m、地表起伏度62～140 m”的区域滑坡最频繁，需要对滑坡灾害重点关注与防治。针对原始关联规则算法仅能以滑坡次数为挖掘依据的局限，优化算法以考虑滑坡范围的影响，深入研究多种影响因素对滑坡的综合作用，为滑坡灾害的快速判识与防灾减灾提供参考。

关键词：大数据挖掘技术；优化关联规则算法；FP-Growth算法；滑坡影响因素；频繁组合挖掘
中图分类号：P642.22　　**文献标识码：**A　　**文章编号：**1672-2132(2025)03-0532-10

Mining of Landslide Frequent Factor Combinations Based on Optimized FP-Growth Algorithm

LI Jiaying[1], HAO Binchao[1], WANG Weidong[2], WANG Zhichao[1], CAO Lulai[3], HAN Zheng[2], ZHU Chongzheng[1]

(1. College of Civil Engineering, Xiangtan University, Xiangtan 411105, China;
2. College of Civil Engineering, Central South University, Changsha 410083, China;
3. Trans Figure. Design Co., Changsha 410000, China)

Abstract: The landslide influencing factors (LIFs) are complex and diverse. Mining frequent combinations of these factors can macroscopically and quickly identify landslide-prone areas. A total of 586 landslides in Liangshan Yi Autonomous Prefecture, Sichuan Province, were used as the sample dataset. Twelve LIFs were selected from five aspects: geological conditions, hydrological conditions, terrain conditions, meteorological conditions, and human engineering activities. The chi-squared test

* **收稿日期：**2023-12-17；**修回日期：**2024-05-06
基金项目：国家自然科学基金项目（51478483，52078493）、湖南省教育厅科研项目一般项目（23C0033）、湘潭大学科研项目（23QDZ07）资助
作者简介：李佳颖(1994—)，女，讲师，博士。主要从事滑坡等地质灾害与道铁选线规划方面的研究。E-mail:Jiaying_li@xtu.edu.cn
通信作者：王卫东(1971—)，男，教授，博士。主要从事滑坡等地质灾害研究。E-mail: csuwwd@csu.edu.cn

was used to eliminate LIFs weakly related to landslides, and coupled analysis was conducted between landslide areas and factor zoning. To address the limitation that big data mining algorithms could rely only on discrete variables such as the frequency of historical landslides, feature parameters were introduced to optimize the FP-Growth algorithm. This enabled it to utilize continuous variables such as historical landslide area and density as mining inputs. Frequent secondary factor combinations were mined, and their accuracy was verified using chi-squared and frequency ratio tests. The results showed that the optimized association rule algorithm based on historical landslide density was more effective in identifying frequent secondary factor combinations. Specifically, regions characterized by "elevation < 1 769 m and surface relief of 62-140 m" experienced the highest landslide frequency, requiring focused attention and mitigation efforts. This study addresses the limitation of conventional association rule algorithms that rely solely on landslide frequency as mining inputs. It optimizes the algorithm to incorporate the influence of landslide extent and conducts in-depth analysis of the combined effects of multiple LIFs on landslides, providing references for the rapid identification of landslide disasters and disaster mitigation.

Keywords: big data mining technology; optimized association rule algorithm; FP-Growth algorithm; landslide influencing factors; frequent combination mining

0 引 言

滑坡是多种影响因素共同作用的结果,固定区域必然存在滑坡发生频繁的影响因素组合,挖掘对滑坡影响最为显著的因素组合对于宏观、快速、初步判识滑坡高敏感性区域范围有很好的指导意义[1-2]。因此,依据历史滑坡与其影响因素等大量客观数据挖掘滑坡频繁因素组合是现阶段的重要问题[3-4]。

目前研究主要通过计算因素权重和统计对比的方法寻找滑坡频繁因素组合[5-6]。刘伟淇等[7]利用层次分析法及CRITIC法分别确定影响因素主观和客观权重,根据综合赋权信息量法叠加因子图层,生成滑坡敏感性地图。黄发明等[8]利用相关系数、线性回归、主成分分析等方法优化环境因子组合筛选,将其作为滑坡易发性预测模型的输入变量。然而,此类方法仅评估单个影响因素对滑坡的权重并简单叠加,并未考虑多种影响因素之间的相互联系和影响[9]。因此,如何科学合理地分析多种影响因素对滑坡灾害的综合作用是一个难题。针对此问题,需要直接建立因素组合与滑坡灾害的关系模型,计算因素组合对滑坡的重要性,以此判识其对滑坡的影响程度[10]。崔成涛等[11]假采用层次分析法计算出主观权重,通过变异系数法确定客观权重,然后利用博弈论原理得到组合权重,从而评价滑坡易发性。但此类方法需要先假定滑坡影响因素组合,假定过程主要依赖主观经验,或者需要计算每种因素组合的综合权重[12-14]。而滑坡灾害影响因素复杂多样,并且每种影响因素存在多种二级因素,因此各种二级因素组合数量庞大,需要一种能快速准确地从大量数据中提取有用信息的方法。

面向大数据挖掘的关联规则算法能在大量数据集中寻找项目间或项目集间的频繁模式与关联过程,因此也能用于分析滑坡与影响因素之间的关系[15]。其中,频繁模式树(FP-Growth)算法能简洁直观地发现大量数据中的频繁项集和关联规则[16],不同于其它关联规则算法需要多次扫描数据库,FP-Growth算法利用分治策略,扫描一遍数据库后,将项集压缩到频繁模式树中,再将频繁模式树分化并分别挖掘[17]。因此,FP-Growth算法只需要扫描两次数据库,且不需要候选项集,极大提高关联规则算法的效率,适用于海量数据的管理分析[18]。但原始FP-Growth算法是依据项目出现次数挖掘满足支持度和置信度要求的项集,因此原始算法只能通过计算滑坡发生次数挖掘滑坡频繁二级因素组合,难以应用于需要考虑滑坡灾害与影响因素之间空间关联的问题。

鉴于此,本文拟在原始算法的数据集中引入作为连续变量的特征参数,优化目前最常用的关联规则算法之一——FP-Growth算法,通过计算特征参数的累计值挖掘滑坡频繁的二级因素组合。以四

川省凉山彝族自治州为研究区域，将历史滑坡区域与12个滑坡影响因素作为数据集，将历史滑坡面积与历史滑坡密度作为特征参数优化原始FP-Growth算法，挖掘滑坡频繁的二级因素组合，确定影响因素与滑坡之间的关联规则，与原始FP-Growth算法的挖掘结果进行对比，验证二级因素组合与滑坡之间的相关性，深入研究多种影响因素对滑坡的综合作用，为滑坡灾害的快速判识与防灾减灾提供参考。

1 关联规则算法建模

1.1 FP-Growth算法

关联规则算法的目的是发现数据项集之间存在的关联关系，属于数据挖掘研究的一种[19]。关联规则挖掘的数据集为{TID：itemset}，其中TID为事务标识，itemset为TID的项，该两个参数是寻找关联规则和频繁项集的数据基础[20]。在数据集中，不可分割的最小单位信息即为项，项的集合即为项集。支持度和置信度是算法中常用的两个指标，用于衡量项集之间的关联程度。支持度是指某个项集在数据集中出现的频率，即该项集在数据集中出现的次数与总事务数之比；置信度是指在某个条件下出现关联规则的概率，即同时包含两个事务的项集出现的次数与包含其中一个事务的项集出现的次数之比。早期关联规则算法研究使用先验算法挖掘关联规则，需要对数据库进行多次扫描，并产生大量的候选频繁项集，时间成本与空间成本较大[21-22]。FP-Growth算法是一种使用压缩数据结构的频繁模式树存储查找频繁项集所需全部信息的方法，在先验算法的基础上使用了高级的数据结构——频繁模式树，有效减少扫描次数，提高算法的运行效率[23]。频繁模式树是一种由频繁项头表和项前缀树构成的特殊前缀树，FP-Growth算法的主要工作就是构建频繁模式树和在其中挖掘频繁项集[24]。

FP-Growth算法通过不断迭代频繁模式树的构造和投影过程，构建每个频繁项的条件投影数据库和模式树，直到构建的模式树为空或只包含一条路径。构建频繁模式树后，对其进行自下而上的搜索，所以整个运行过程只需要遍历两遍数据集，有效提高算法效率。在算法构造路径时，对于相同的项，路径可能部分重叠，路径重叠的越多，频繁模式树结构的压缩效果越好[25]。FP-Growth算法主要包括以下4个步骤：

（1）第一次扫描数据库，得到整个数据库的项，剔除小于最小支持度的项，将项按频繁程度降序排列。

（2）第二次扫描数据库，构建频繁模式树，树的根节点为空，各项为子节点，按照步骤（1）的顺序形成由根节点到子节点的分支，将频繁项集逐步加入树结构中，如果结构中已存在该项集，则增加该项集的值，如果不存在，则增加一个分支。频繁模式树的构建流程如图1所示。

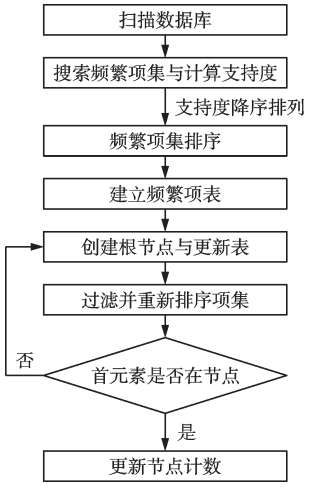

图1 频繁模式树构建流程

Fig.1 Flowchart of frequent pattern tree construction

（3）按照由下而上的顺序找到每个项的条件模式基，递归调用树结构，剔除小于最小支持度的项集。如果树结构路径单一，则直接列举所有项集；如果树结构路径不单一，则继续调用直到路径单一。

（4）计算各个频繁项集的置信度，找到满足条件的强关联规则。

FP-Growth算法利用频繁模式树的结构将关联规则挖掘分解为多个子问题，通过递归数据结构减小了算法运行计算量与时间成本。本文以历史滑坡次数为挖掘依据挖掘滑坡频繁二级因素组合，支持度与置信度的计算公式如下所示：

$$Support(A,B)=P(A\&B) \tag{1}$$

$$Confidence(A\Rightarrow B)=P(B|A)=\frac{P(A\&B)}{P(A)} \tag{2}$$

式中，$P(A\&B)$为二级因素A与B同时出现的概率；$P(B|A)$为在二级因素A出现的基础上二级因

素B出现的概率;$P(A)$为二级因素A的出现概率。

1.2 优化的FP-Growth算法

原始FP-Growth算法虽然能从大量数据中挖掘满足支持度和置信度要求的项集,但是只能以计算项数的方式挖掘频繁因素组合,难以分析滑坡范围的影响。因此,需要在原有的数据集{TID: itemset}中引入特征参数,形成新的数据集{TID: itemset, characteristic}。利用作为连续变量的特征参数挖掘频繁项集,挖掘频繁项集的基础不再是项集出现次数,而是相应特征的累计值。而滑坡与因素组合的空间关联特征在滑坡关联规则分析中不可忽视,为表征滑坡影响范围和滑坡影响概率分别与因素组合的空间关联性,本文以历史滑坡面积与历史滑坡密度为特征参数优化FP-Growth算法,以历史滑坡面积为特征参数的支持度与置信度的计算公式如下所示:

$$Support(A,B)=\frac{Area(A\&B)}{\sum Area} \tag{3}$$

$$Confidence(A\Rightarrow B)=\frac{Area(A\&B)}{Area(A)} \tag{4}$$

式中,$Area(A\&B)$为二级因素A与B所在区域的滑坡面积;$\sum Area$为研究区域总滑坡面积;$Area(A)$为二级因素A所在区域的滑坡面积。

以历史滑坡密度为特征参数的支持度与置信度的计算公式如下所示:

$$Support(A,B)=\frac{Density(A\&B)}{\sum Density} \tag{5}$$

$$Confidence(A\Rightarrow B)=\frac{Density(A\&B)}{Density(A)} \tag{6}$$

式中,$Density(A\&B)$为二级因素A与B所在区域的滑坡密度;$\sum Density$为研究区域总滑坡密度;$Density(A)$为二级因素A所在区域的滑坡密度。

1.3 评价指标

1.3.1 卡方检验

卡方检验是一种以χ^2分布为基础的假设检验方法,主要通过计算实际值与理论值之间的偏差判断假设的正确性,卡方检验值越大,则偏差程度越大;卡方检验值越小,则偏差程度越小[26]。卡方检验值的计算公式如下所示:

$$\chi^2=\sum_{i=1}^{n}\frac{(x_i-T)^2}{T} \tag{7}$$

式中,x_i为实际观测值;T为理论推测值。

本文利用卡方检验验证滑坡与影响因素的相关性与挖掘结果的准确性。卡方检验假设影响因素与滑坡无关,显著性是指当原假设为正确时人们却把它拒绝的概率或风险。由于卡方检验值与显著性的对应关系,卡方检验值越大,显著性越小,说明影响因素与滑坡无关的假设越不正确,二者越相关;卡方检验值越小,显著性越大,影响因素与滑坡越不相关。因此,根据卡方检验显著性可以确定影响因素与滑坡的相关性和挖掘结果的准确性。

同时,为避免因素组合区域面积的影响,计算各个组合所在区域的滑坡密度以分析挖掘结果的准确性,滑坡密度越大,该因素组合与滑坡灾害的关联越显著;反之说明该因素组合与滑坡灾害的关联越不显著。将其与整个研究区域的滑坡密度进行对比,挖掘因素组合的滑坡密度大于研究区域的滑坡密度,说明其挖掘结果更准确,挖掘方法更滑坡频繁二级因素组合的挖掘研究。

1.3.2 频率比

滑坡频率比为二级因素组合滑坡密度与研究区域滑坡密度之比,利用二级因素组合的频率比验证组合与滑坡的相关性。滑坡频率比越大,说明二级因素组合所在区域出现滑坡的可能性越大,反之则越小。如果区域的滑坡频率比大于1则说明该区域出现滑坡可能性较大,如果频率比小于1则说明该区域出现滑坡可能性较小。频率比的计算公式如下所示:

$$FR=\frac{P(LF_i)}{P(F_i)}=\frac{A_{LF_i}/A_{F_i}}{A_L/A}=\frac{A_{LF_i}\cdot A}{A_L\cdot A_{F_i}} \tag{8}$$

式中,A_{LF_i}为二级因素组合F_i所在区域的滑坡面积;A_{F_i}为二级因素组合F_i所在区域的面积;A_L为研究区域滑坡面积;A为研究区域总面积。

2 凉山彝族自治州概况与滑坡数据

2.1 凉山彝族自治州自然概况

凉山彝族自治州地处四川省南部,总面积约60 400平方千米(图2),位于四川省西南部,地处北纬26°03′~29°18′,东经100°03′~103°52′。地质构造

复杂，地貌复杂多样，界于四川盆地和云南省中部高原之间，地势西北高，东南低，北部高，南部低。凉山彝族自治州属亚热带季风气候，四季虽不明显，但干湿季节分明，冬半年日照充足，少雨干暖；夏半年云雨较多，气候凉爽。凉山地形与大气环流的复杂使凉山彝族自治州干雨季明显，导致凉山气候的复杂性、多样性。境内河流众多，有金沙江、雅砻江和大渡河三大水系。由于复杂的自然环境与季节性强降雨的影响，凉山彝族自治州的滑坡灾害频繁发生。根据国土资源部《全国地质灾害通报》，凉山彝族自治州地质灾害频发，滑坡灾害占比最高，以中小型浅层滑坡为主。滑坡分布如图2所示。

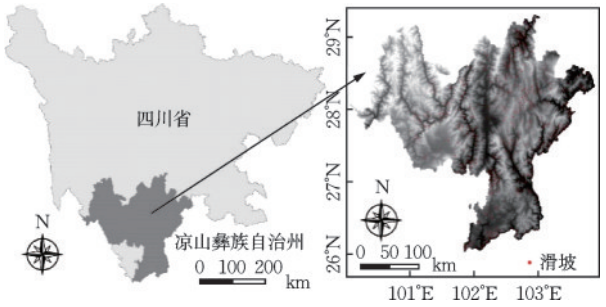

图2　凉山彝族自治州位置及滑坡分布

Fig.2　Location of Liangshan Yi Autonomous Prefecture and landslide distribution

2.2　滑坡与影响因素分析

滑坡是多种影响因素作用的结果，根据实地调查与相关研究收集了12个影响因素，地质方面包括岩性、与地质结构线距离；水文方面包括归一化植被指数、与水系距离；地形方面包括高程、坡度、地表起伏度、地形地貌、地形湿度指数；气象方面包括年平均降雨量；人类活动工程包括与铁路距离、与公路距离。各影响因素数据源简介如下：(1)通过高分辨率扫描设备扫描四川省地质图，经过数字化和配准得到岩性分布；(2)由中国科学院地理空间数据云提供的数字高程模型得到高程；(3)通过GIS平台处理DEM数据得到坡度、地表起伏度和地形湿度指数；(4)通过国家基础地理信息系统得到地形地貌、地质结构线和水系数据；(5)通过中国科学院资源环境科学与数据中心得到归一化植被指数；(6)通过中国气象局气象数据中心得到降雨数据；(7)通过中国交通运输部得到公路路网与铁路路网数据。

在GIS平台基于各个数据源数字化各个影响因素图层，其中岩性与地形地貌具有固定类别，为离散型影响因素，其它为连续型影响因素。为了后续分析，将连续型影响因素划分为多个二级因素。其中，作用于研究区域全范围的影响因素利用自然断点法划分，如高程和坡度等；作用范围受限的影响因素利用等间距统计法划分，如与水系距离、与铁路距离、与公路距离和与地质结构线距离等。各影响因素的二级因素分布如图3所示。

3　滑坡频繁因素组合挖掘

3.1　因素相关性分析

卡方检验是一种用途广泛的假设检验方法，能确定影响因素与滑坡的相关性，计算各个影响因素的卡方检验值与显著性，结果见表1。

表1　影响因素卡方检验值与显著性

Table 1　Chi-squared test values and significance of influencing factors

影响因素	χ^2	显著性	影响因素	χ^2	显著性
岩性	59.46	0.000	高程	401.97	0.000
坡度	11.59	0.021	地表起伏度	10.37	0.035
地形地貌	67.05	0.000	地形湿度指数	4.22	0.378
归一化植被指数	2.76	0.599	年平均降雨	77.29	0.000
与水系距离	3.97	0.411	与铁路距离	4.10	0.403
与公路距离	3.50	0.477	与地质构造线距离	38.41	0.000

基于试错法剔除显著性大于0.4的影响因素，筛选出8个影响因素：岩性、高程、坡度、地表起伏度、地形地貌、地形湿度指数、年平均降雨量和与地质构造线距离。

3.2　二级因素编码

挖掘二级因素组合之前，需要对滑坡单元的二级影响因素进行编码。基于二级影响因素分布图，各个影响因素的编码见表2。利用编码表征各个二级影响因素，各个二级因素组合也能用编码组合表示，例如编码组合“11，23，31，44，51，84”表示二级因素组合“稳定岩、高程2 393～2 990 m、坡度＜9.08°、地表起伏度＞717 m、平原地形、与地质构造线距离1 500～2 000 m”。

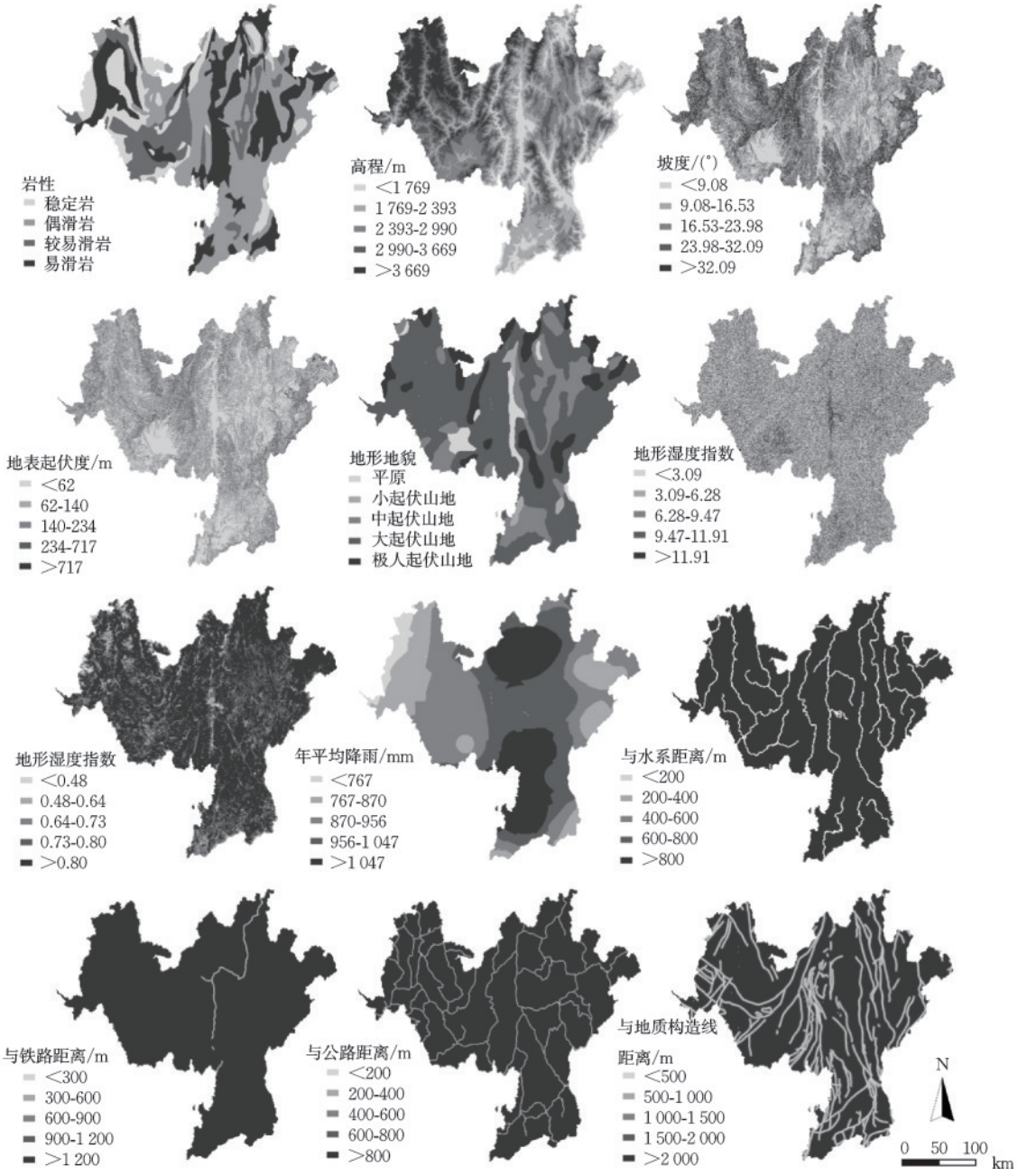

图3　二级影响因素分布

Fig.3　Distribution of secondary influencing factors

用于挖掘的影响因素有8种，每种影响因素被划分为4到5个二级因素，因此存在近百万种因素组合。因素组合中的二级因素越多，组合出现的频率越低，相应的支持度也越低。因此，算法设置的最小支持度和最小置信度需要考虑因素组合的出现频率。如果最小支持度和最小置信度设置过大，则满足要求的二级因素组合种类过少，不利于关联规则挖掘；如果设置过小，则挖掘的二级因素组合难以满足频繁项集的要求。

3.3　基于不同挖掘依据的挖掘结果

3.3.1　基于滑坡个数挖掘滑坡频繁因素组合

叠加分析各个影响因素图层和历史滑坡图层，将表2中二级因素编码导入586个历史滑坡单元中，以带有二级因素编码的滑坡个数作为挖掘依据，利用试错法设置最小支持度与最小置信度，使其既能用于挖掘关联规则，又满足频繁项集的要求。利用Python运行FP-Growth算法挖掘二级因素组合，生成组合的关联规则，计算各个影响因素

表2　影响因素编码

Table 2　Coding of influencing factors

影响因素	二级影响因素	编码	影响因素	二级影响因素	编码
岩性	稳定岩	11	高程/m	＜1 769	21
	偶滑岩	12		1 769～2 393	22
	较易滑岩	13		2 393～2 990	23
	易滑岩	14		2 990～3 669	24
	/	/		＞3 669	25
坡度/(°)	＜9.08	31	地表起伏度/m	＜62	41
	9.08～16.53	32		62～140	42
	16.53～23.98	33		140～234	43
	23.98～32.09	34		234～717	44
	＞165	35		＞717	45
地形地貌	平原	51	地形湿度指数	＜3.09	61
	小起伏山地	52		3.09～6.28	62
	中起伏山地	53		6.28～9.47	63
	大起伏山地	54		9.47～11.91	64
	极大起伏山地	55		＞11.91	65
年平均降雨量/mm	＜767	71	与地质构造线距离/m	＜500	81
	767～870	72		500～1 000	82
	870～956	73		1 000～1 500	83
	956～1 047	74		1 500～2 000	84
	＞1 047	75		＞2 000	85

之间的置信度。基于组合中各个二级因素的置信度计算平均值表征组合置信度，置信度1表示在存在前一个二级因素的滑坡中，该因素组合存在的概率；置信度2表示在存在后一个二级因素的滑坡中，该因素组合存在的概率。置信度越大，说明该组合中二级因素在滑坡区域的关联越紧密。结果见表3。

表3　基于滑坡个数的二级因素组合及其置信度

Table 3　Secondary factor combinations based on landslide frequency and their confidence

编码组合	二级因素组合	置信度1	置信度2	置信度
31, 41	坡度＜9.08°、地表起伏度＜62 m	0.771	0.556	0.669
12, 54	偶滑岩、大起伏山地	0.717	0.557	0.637
54, 85	大起伏山地地形、与地质构造线距离＞2 000 m	0.601	0.644	0.623

根据表3，以滑坡个数为挖掘依据，二级因素组合(31, 41)、(12, 54)和(54, 85)被挖掘，说明具有该3个二级因素组合的区域更有可能发生滑坡灾害。

3.3.2　基于滑坡面积挖掘滑坡频繁因素组合

同样叠加分析影响因素图层与历史滑坡图层并导入二级因素编码，以带有二级因素编码的滑坡面积作为挖掘数据，利用优化FP-Growth算法挖掘滑坡频繁二级因素组合。置信度1表示存在该因素组合的滑坡面积与存在前一个二级因素的滑坡面积之比；置信度2表示存在该因素组合的滑坡面积与存在后一个二级因素的滑坡面积之比。结果见表4。

表4　基于滑坡面积的二级因素组合及其置信度

Table 4　Secondary factor combinations based on landslide area and their confidence

编码组合	二级因素组合	置信度1	置信度2	置信度
12, 54	偶滑岩、大起伏山地	0.707	0.550	0.629
22, 54	高程1 769～2 393 m、大起伏山地	0.668	0.394	0.531
42, 54	地表起伏度62～140 m、大起伏山地	0.660	0.513	0.587
12, 42	偶滑岩、地表起伏度62～140 m	0.494	0.495	0.495

根据表4，以滑坡面积为挖掘依据，二级因素组合(12, 54)、(22, 54)、(42, 54)和(12, 42)被挖掘，说明具有该4个二级因素组合的区域更有可能发生滑坡灾害。

3.3.3　基于滑坡密度挖掘滑坡频繁因素组合

以带有影响因素编码的滑坡密度作为挖掘数据，利用优化FP-Growth算法挖掘滑坡频繁二级因素组合。置信度1表示存在该因素组合的滑坡密度与存在前一个二级因素的滑坡密度之比；置信度2表示存在该因素组合的滑坡密度与存在后一个二级因素的滑坡密度之比。结果见表5。

根据表5，以滑坡密度为挖掘依据，二级因素组合(12, 22)、(21, 54)和(21, 42)被挖掘，说明具有该3个二级因素组合的区域更有可能发生滑坡灾害。

表5　基于滑坡密度的二级因素组合及其置信度

Table 5　Secondary factor combinations based on landslide density and their confidences

编码组合	二级因素组合	置信度1	置信度2	置信度
12，22	偶滑岩、高程1 769～2 393 m	0.421	0.570	0.496
21，54	高程＜1 769m、大起伏山地	0.540	0.540	0.540
21，42	高程＜1 769m、地表起伏度62～140 m	0.393	0.449	0.421

4　挖掘结果检验

4.1　卡方检验

利用卡方检验进一步检验二级因素组合与滑坡的关联，根据卡方检验的原理，分别统计三种方法挖掘的二级因素组合所在区域发生滑坡与未发生滑坡的面积，根据式7计算各个二级因素组合的卡方检验值。研究区域的总面积为59 486.88 km^2，其中滑坡灾害面积为77.34 km^2，因此研究区域的滑坡密度为0.130%。计算各个卡方检验值与滑坡密度，结果见表6。

表6　二级因素组合卡方检验值与滑坡密度

Table 6　Chi-squared test values and landslide density of secondary factor combinations

挖掘依据	二级因素组合	卡方检验值	滑坡密度/%
历史滑坡个数	31，41	0.817	0.098
	12，54	4.754	0.185
	54，85	0.549	0.118
历史滑坡面积	12，54	4.754	0.185
	22，54	6.872	0.221
	42，54	0.000	0.130
	12，42	1.448	0.165
历史滑坡密度	12，22	3.886	0.202
	21，54	15.748	0.322
	21，42	11.874	0.341

卡方检验值越大说明二级因素组合与滑坡灾害的关联越显著，根据表6，三种挖掘方法中，基于滑坡密度挖掘的二级因素组合的卡方检验值最大，其与滑坡灾害的关联最显著，且其滑坡密度均远远大于研究区域的滑坡密度，说明以滑坡密度为挖掘依据的挖掘方法更适用于滑坡频繁二级因素组合的挖掘研究。

4.2　频率比

本文根据式(8)计算各个滑坡频繁二级因素组合的频率比以证明其与滑坡的相关性，频率比计算结果见表7。频率比越大，二级因素组合所在区域发生滑坡的可能性越大，反之，发生滑坡的可能性越小。

表7　二级因素组合频率比

Table 7　Frequency ratio of secondary factor combinations

挖掘依据	因素组合编码	区域滑坡面积/km^2	区域面积/km^2	频率比
历史滑坡个数	31，41	8.405	8 615.328	0.749
	12，54	28.305	15 325.065	1.419
	54，85	30.525	25 963.366	0.903
历史滑坡面积	12，54	28.305	15 325.065	1.419
	22，54	20.277	9 191.384	1.695
	42，54	26.382	20 316.207	0.998
	12，42	19.790	11 961.409	1.271
历史滑坡密度	12，22	16.153	7 997.188	1.552
	21，54	16.312	5 061.870	2.475
	21，42	11.188	3 284.795	2.616

根据表7，以滑坡密度为挖掘依据挖掘的大部分因素组合频率比均大于1，说明该挖掘方法更适用于滑坡频繁二级因素组合的挖掘。同时，可以看出滑坡发生最频繁的二级因素组合是“高程＜1 769 m、地表起伏度62～140 m”，具有该因素组合的区域需要对滑坡灾害重点关注与防治。

4.3　算法对比分析

由卡方检验临界值表可知，在利用原始FP-Growth算法挖掘的因素组合中，只有因素组合(12，54)的卡方检验值大于临界值3.841，且其滑坡密度大于研究区域滑坡密度，频率比大于1，说明只有该组合所在区域频繁发生滑坡。而在利用优化FP-Growth算法挖掘的因素组合中，只有以历史滑坡面积为挖掘依据的因素组合(42，54)所在区域没有频繁发生滑坡，以历史滑坡密度为挖掘依据的三个因素组合所在区域均频繁发生滑坡。对比检验结果可以发现，基于优化FP-Growth算法的挖掘结果准确性更高，优化FP-Growth算法比原始算法更适用于滑坡频繁二级因素组合的挖掘，尤其是以历史滑坡密度为挖掘依据的优化算法。

5 结 论

本文提出了一种基于优化关联规则算法的滑坡频繁因素组合挖掘方法，以四川凉山彝族自治州为例，基于海量滑坡灾害与滑坡影响因素数据，对比分析原始关联规则算法与优化关联规则算法，对滑坡灾害的影响因素进行关联分析，得到以下结论：

(1)针对原始关联规则算法只能通过计数滑坡的发生次数进行关联分析的局限性，引入历史滑坡数据作为连续型变量优化FP-Growth算法，在原有的数据集{TID：itemset}中引入特征参数，形成新的数据集{TID：itemset，characteristic}，以历史滑坡面积与历史滑坡密度作为挖掘依据，实现了对需要考虑滑坡范围的研究分析，能为滑坡防灾减灾决策提供初步参考。

(2)分别以历史滑坡个数、历史滑坡面积和历史滑坡密度挖掘滑坡频繁因素组合，基于历史滑坡个数挖掘的滑坡频繁因素组合编码为(31，41)、(12，54)和(54，85)；基于历史滑坡面积挖掘的滑坡频繁因素组合编码为(12，54)、(22，54)、(42，54)、(12，42)；基于历史滑坡密度挖掘的滑坡频繁因素组合编码为(12，22)、(21，54)、(21，42)。

(3)利用卡方检验与频率比对三种挖掘方法的结果进行检验，优化FP-Growth算法比原始FP-Growth算法更适用于滑坡频繁二级因素组合的挖掘，以历史滑坡密度为挖掘依据的结果与滑坡相关性更高，基于历史滑坡密度的优化FP-Growth算法更适用于滑坡频繁因素组合的挖掘研究。研究区域中二级因素组合“高程＜1 769 m、地表起伏度62～140 m”的区域滑坡最频繁，该区域需要对滑坡灾害重点关注与防治。

参考文献：

[1] 张泽方，钱志宽，魏勇，等.考虑最优影响因素组合的滑坡易发性评价：以水城区为例[J].科学技术与工程，2023，23(10)：4091-4099.
Zhang Z F，Qian Z K，Wei Y，et al. Landslide susceptibility evaluation considering optimal combination of influencing factors：a case study of shuicheng district[J]. Science Technology and Engineering，2023，23(10)：4091-4099.(in Chinese)

[2] Sun D L，Wen H J，Wang D Z，et al. A random forest model of landslide susceptibility mapping based on hyperparameter optimization using Bayes algorithm[J]. Geomorphology，2020，362：107201.

[3] 秦文涛，郭小坤，郭军峰，等.数据仓库和数据挖掘技术在滑坡预测预报中的应用[J].岩土工程技术，2022，36(3)：185-189.
Qin W T，Guo X K，Guo J F，et al. Application of data warehouse and data mining on landslide prediction[J]. Geotechnical Engineering Technique，2022，36(3)：185-189.(in Chinese)

[4] 朱鸿鹄，王佳，李厚芝，等.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，2022，30(5)：1517-1527.
Zhu H H，Wang J，Li H Z，et al. Association rule analysis for giant landslide deformation of the Three Gorges Reservoir region based on data mining[J]. Journal of Engineering Geology，2022，30(5)：1517-1527.(in Chinese)

[5] 毛正君，张瑾鸽，仲佳鑫，等.基于确定性系数法的梯田型黄土滑坡隐患影响因素分析[J].水土保持通报，2023，43(2)：183-192，340.
Mao Z J，Zhang J G，Zhong J X，et al. Sensitivity analysis on factors influencing loess terrace landslide potential using certainty factor method[J]. Bulletin of Soil and Water Conservation，2023，43(2)：183-192，340.(in Chinese)

[6] Dun J，Feng W，Yi X，et al. Detection and mapping of active landslides before impoundment in the baihetan reservoir area (China) based on the Time-Series InSAR Method[J]. Remote Sensing，2021，13(16)：3213.

[7] 刘伟淇，张家铭.竹溪县滑坡灾害易发性分区评价[J].自然灾害学报，2024，33(1)：175-185.
Liu W Q，Zhang J M. Zoning evaluation of landslide hazard susceptibility in Zhuxi County[J]. Journal of Natural Disasters，2024，33(1)：175-185.(in Chinese)

[8] 黄发明，刘科技，曾子强，等.环境因子筛选及组合方法对滑坡易发性预测的影响规律[J].应用基础与工程科学学报，2024，32(1)：49-71.
Huang F M，Liu K J，Zeng Z Q，et al. The impact of environmental factor screening and combination methods on landslide susceptibility prediction[J]. Journal of Basic Science and Engineering，2024，32(1)：49-71.(in Chinese)

[9] 郭俊辉，李侠，叶晨男，等.输电线路区域滑坡易发性分析方法与应用[J].山西建筑，2023，49(20)：58-62.
Guo J H，Li X，Ye C N，et al. Analysis method and application of landslide susceptibility in transmission line areas[J]. Shanxi Architecture，2023，49(20)：58-62.(in Chinese)

[10] Sameen M I，Sarkar R，Pradhan B，et al. Landslide spatial modelling using unsupervised factor optimisation and regularised greedy forests[J]. Computers & Geosciences，2020，134：104336.

[11] 崔成涛，李丽敏，符振涛，等.基于博弈论赋权信息量模型的滑坡易发性评价[J].人民珠江，2024，45(2)：9-17.
Cui C T，Li L M，Fu Z T，et al. Landslide susceptibility evaluation based on empowerment information quantity model of game[J]. Theory Pearl River，2024，45(2)：9-17.(in Chinese)

[12] 李阳，张建军，魏广阔，等.晋西黄土区极端降雨后浅层滑坡调查及影响因素分析[J].水土保持学报，2022，36(5)：44-50.
Li Y，Zhang J J，Wei G K，et al. Investigation of shallow landslide after extreme rainfall and analysis of its influencing factors in the West Shanxi Loess Region[J]. Journal of Soil and Water Conservation，2022，36(5)：44-50.(in Chinese)

[13] 高明，贺可强，刘洪华，等.基于变权重的水库滑坡稳定性模糊综合评价[J].科学技术与工程，2022，22(10)：3885-3891.
Gao M，He K Q，Liu H H，et al. Fuzzy comprehensive evaluation of reservoir landslide stability based on variable weight[J]. Science Technology and Engineering，2022，22(10)：3885-3891.(in Chinese)

[14] Zhang H，Yin C，Wang S P，et al. Landslide susceptibility mapping based on landslide classification and improved convolutional neural networks[J]. Natural Hazards，2023，116(2)：1931-1971.

[15] Fister I J，Fister I，Fister D，et al. A comprehensive review of visualization methods for association rule mining：Taxonomy，challenges，open problems and future ideas[J]. Expert Systems with Applications，2023，233：120901.

[16] Yang Y，Tian N，Wang Y P，et al. A parallel FP-Growth mining algorithm with load balancing constraints for traffic crash data[J]. International Journal of Computers Communications & Control，2022，17(4)：4806.

[17] Alsaeedi H A，Alhegami A S. An incremental interesting maximal frequent itemset mining based on FP-Growth algorithm[J]. Complexity，2022，2022：1942517.

[18] Jang H J，Yang Y，Park J S，et al. FP-Growth algorithm for discovering region-based association rule in the IoT environment[J]. Electronics，2021，10(24)：3091.

[19] Li J Y，Wang W D，Han Z，et al. Analysis of secondary-factor combinations of landslides using improved association rule algorithms：a case study of Kitakyushu in Japan[J]. Geomatics Natural Hazards & Risk，2021，12(1)：1885-1904.

[20] 李佳临，邬阳，魏奇，等.改进关联规则算法在自然资源云中的应用研究[J].时空信息学报，2024，31(1)：140-147.
Li J L，Wu Y，Wei Q，et al. Research on the application of improved association rule algorithm in natural resource cloud[J]. Journal of Spatio-Temporal Information，2024，31(1)：140-147.(in Chinese)

[21] Fernandez-Basso C，Ruiz M D，Martin-Bautista M J. New Spark solutions for distributed frequent itemset and association rule mining algorithms[J]. Cluster Computing the Journal of Networks Software Tools and Applications，2024，27(2)：1217-1234.

[22] Li H S，Sheu P C Y. A scalable association rule learning and recommendation algorithm for large-scale microarray datasets[J]. Journal of Big Data，2022，9(1)：35.

[23] 乔阳阳，王丽娟.数据点位置并行FP-Growth挖掘算法仿真[J].计算机仿真，2023，40(5)：501-505.
Qiao Y Y，Wang L J. Simulation of parallel FP-Growth mining algorithm for data point location[J]. Computer Simulation，2023，40(5)：501-505.(in Chinese)

[24] 魏坤，王芳，黄树成.改进的频繁模式挖掘算法[J].计算机与数字工程，2021，49(11)：2175-2179.
Wei K，Wang F，Huang S C. Improved frequent pattern mining algorithm[J]. Computer & Digital Engineering，2021，49(11)：2175-2179.(in Chinese)

[25] 毛伊敏，吴斌，许春冬，等.基于Spark的并行频繁项集挖掘算法[J].计算机集成制造系统，2023，29(4)：1267-1283.
Mao Y M，Wu B，Xu C D，et al. Parallel algorithm for mining frequent item sets based on spark[J]. Computer Integrated Manufacturing Systems，2023，29(4)：1267-1283.(in Chinese)

[26] Doğan O，Taşpınar S，Bera A K. A Bayesian robust chi-squared test for testing simple hypotheses[J]. Journal of Econometrics，2021，222(2)：933-958.

（本文编辑：周小潭）