

基于Apriori算法的滑坡监测数据关联规则挖掘研究

吕梅洁¹, 李明亮^{1,2*}, 宗佳泰¹, 侯梦媛¹

¹河北地质大学信息工程学院, 河北 石家庄

²智能传感物联网技术河北省工程研究中心, 河北 石家庄

收稿日期: 2023年10月17日; 录用日期: 2023年11月16日; 发布日期: 2023年11月24日

摘 要

滑坡是一种常见的自然灾害, 对人类生命和财产安全造成严重威胁, 为了得到海量滑坡监测数据中隐藏的有价值信息, 基于关联规则的数据挖掘技术在滑坡监测领域得到了广泛应用。本文以2016~2019年藏东南扎木弄巴滑坡观测数据作为研究对象, 通过挖掘滑坡监测数据中的关联规则, 识别出不同变量之间的关联性, 并利用这些关联规则进行滑坡的监测和预测。首先介绍了Apriori算法的原理, 然后使用K-means算法对滑坡监测数据中几类数据进行聚类处理, 并将Apriori算法应用在研究数据集上, 进行规则挖掘及分析, 发现降雨量和地下水位的变化与滑坡的相关性, 最终获得有意义的规则信息, 了解滑坡发生的机理和规律, 帮助我们更好地理解 and 应对滑坡灾害。

关键词

关联规则, 滑坡监测, Apriori

Research on Mining Association Rules for Landslide Monitoring Data Based on Apriori Algorithm

Meijie Lv¹, Mingliang Li^{1,2*}, Jiatai Zong¹, Mengyuan Hou¹

¹College of Information Engineering, Hebei GEO University, Shijiazhuang Hebei

²Intelligent Sensor Network Engineering Research Center of Hebei Province, Shijiazhuang Hebei

Received: Oct. 17th, 2023; accepted: Nov. 16th, 2023; published: Nov. 24th, 2023

*通讯作者。

文章引用: 吕梅洁, 李明亮, 宗佳泰, 侯梦媛. 基于 Apriori 算法的滑坡监测数据关联规则挖掘研究[J]. 计算机科学与应用, 2023, 13(11): 2109-2115. DOI: 10.12677/csa.2023.1311210

Abstract

Landslides are a common natural disaster that poses a serious threat to human life and property safety. In order to obtain valuable information hidden in massive landslide monitoring data, data mining technology based on association rules has been widely applied in the field of landslide monitoring. This article takes the observation data of the Zhamulongba landslide in southeastern Tibet from 2016 to 2019 as the research object. By mining association rules in landslide monitoring data, the correlation between different variables is identified, and these association rules are used for landslide monitoring and prediction. Firstly, the principle of Apriori algorithm was introduced. Then, K-means algorithm was used to cluster several types of landslide monitoring data, and Apriori algorithm was applied to the research dataset for rule mining and analysis. The correlation between rainfall and groundwater level changes and landslides was found, and meaningful rule information was obtained to understand the mechanism and laws of landslide occurrence, helping us better understand and respond to landslide disasters.

Keywords

Association Rules, Landslide Monitoring, Apriori

Copyright © 2023 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

我国是一个频繁遭受地质灾害威胁的国家，其中尤以滑坡灾害最为频发。近些年随着监测技术和大数据分析挖掘、机器学习等技术逐渐发展，这些技术已在各行各业进行广泛的应用。大数据作为一场里程碑式的重大技术和思维革命，正逐步深入到地质灾害监测领域。结合大数据分析挖掘技术的滑坡地质灾害监测，可实现数据的快速收集、合理准确地分析挖掘及高效的预警评判，能快速、有效的方式向有关部门提供准确的预警信息。滑坡灾害的影响因素主要分为内在因素和诱发因素。其中内在因素包括坡度、植被覆盖、地层岩性、地质构造、地形等，诱发因素包括降雨、降雪、地震及人类工程活动等[1]。

由于地理大数据的独特性，近年来国内外学者已经在空间数据挖掘的基础上，发展了一系列针对性的地理大数据挖掘方法。其中关联规则挖掘是地理大数据关联关系定性挖掘的代表性方法，旨在发现在相邻时空区域内频繁出现的地理对象关系。陈锐[2]等应用 Apriori 算法对三峡库区白水河滑坡进行关联规则挖掘，赵久彬[3]等提出一种前后部项约束关联规则并行化 FRPFP 算法，对三峡库区奉节至江津库段滑坡进行挖掘，朱鸿鹄[4]等以长江三峡库区新铺滑坡为例提出了滑坡变形行为的关联规则分析挖掘方法。引入了多因素进行关联分析能够更大限度的挖掘滑坡监测数据的价值，为滑坡监测和预警提供科学的指导[2]。

2. 关联规则算法原理

2.1. 关联规则挖掘

关联规则挖掘是进行大数据分析最常用的研究方法之一，它的目的在于从庞大数据集中找出各项之

间的关联,而这种关联不会在数据中表现出来,需要进行关联分析,分析多个变量之间的联系。关联分析多被分为3类简单关联、时序关联、因果关联[5][6]。在关联规则算法中,大于等于最小支持度的项集称为频繁项集,其中不被其它频繁项集包含的频繁项集称为最大频繁项集。其2个重要参数是最小支持度、最小可信度,参数取值会直接影响最后得到的关联结果。

支持度(Support),包含项集的事务数量与全部事务数量的百分比,其计算公式如公式1所示。

$$\text{Support}(X \rightarrow Y) = \frac{(\text{Support_count}(X \cap Y))}{(\text{Total_count}(N))} \quad (1)$$

其中, N 为所有事务的个数。 $X \cap Y$ 表示事务同时发生。

置信度(Confidence),同时包含数据项 X 和数据项 Y 的事务数量与事务 X (或事务 Y) 出现的次数之比,其计算公式如公式2所示。

$$\text{Confidence}(X \rightarrow Y) = \frac{(\text{Support_count}(X \cap Y))}{(\text{Total_count}(X))} \quad (2)$$

2.2. 两步聚类法

两步聚类法(Two-Step Cluster Analysis)是一种聚类分析方法,它主要基于模型的选择和信息度量的方法,可以根据分类指标将样本数据分为不同的簇,从而发现簇内和簇间的关系[7]。

其原理大致如下:

- 1) 初始阶段,通过对数据进行随机抽样或者 K-means 等聚类方法初步聚类,确定初始聚类中心。
- 2) 在每次计算过程中,通过最近邻算法选择最合适的初始类别。
- 3) 在计算过程中,根据簇内距离和簇间距离的度量方法(如对数似然比、信息准则等),不断迭代调整样本所属簇类别,直到簇内距离最小、簇间距离最大。
- 4) 最终根据聚类结果,确定最佳的分类指标和分类数,用于解释和预测数据。

由于样本数据既包括数值型变量又包括分类型变量,两步聚类算法通常采用对数似然函数,若聚成 j 类,则其定义为:

$$l = \sum_{j=1}^J \sum_{i \in I_j} \ln p(X_i | \theta_i)$$

其中, p 为对数似然函数, I_j 为第 j 类的样本集合, θ_i 为第 i 类的参数向量, J 为聚类数目。针对全部样本,其对数似然聚类是各类对数似然聚类之和。两步聚类法相较于传统的聚类方法来说,更加适用于大样本和大维度的数据集,同时还可以控制聚类的数量和稳定性。

2.3. Apriori 算法

Apriori 算法是通过多次对数据库的遍历来寻找频繁集,其 k 项频繁集的产生过程如下:

- 1) 第一步遍历整个事务数据库,得到一项集,将其与最小支持进行比较得到频繁项集 L_1 。
- 2) 判断频繁一项集之间是否可以相互连接,将连接后的候选集通过与最小支持度比较找到符合条件的频繁项集 L_2 。
- 3) 将频繁集相互连接得到候补集,并进行剪枝,剪掉多项候选集子集不在 L_2 中的集合。再次遍历数据库并计数,得到多项频繁集。按此逐层迭代,由频繁 k 项集来生成频繁 $k+1$ 项集,直到无法生成更高维度的项集时循环结束。算法流程图如图1所示。

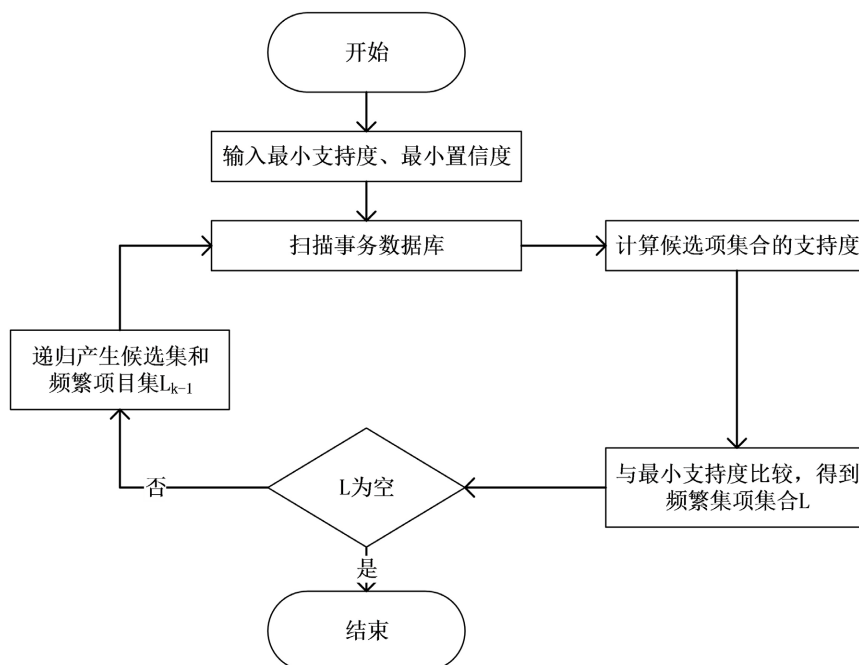


Figure 1. Apriori algorithm flowchart
图 1. Apriori 算法流程图

3. 滑坡监测数据处理

3.1. 数据对象的选取

本文研究的数据对象是 2016~2019 年藏东南扎木弄巴滑坡观测数据, 数据来自青藏高原冰冻圈观测研究站。数据由布设现场的翻斗式降雨量传感器、土壤温度传感器、孔隙水压力传感器、含水量传感器、水势传感器、振动传感器测得, 采集频率 10 分钟/条, 由 CR1000 数据采集器取得原始数据, 土壤温度、孔隙水压力、含水量、水势、振动这些参数采用日平均数据并整理, 降雨数据采用日累计数据并整理[8]。

3.2. 数据对象的处理

针对滑坡监测数据集中数据的特点做如下处理: 1) 对于数据的缺失, 采取均值法、最佳可能法、前后关联法三种措施对数据进行补充。2) 对于数据的异常, 直接采用删除数据的方式进行处理。3) 对于大量的连续型数据, 因为关联规则算法只能针对关系表中的布尔型数据进行挖掘, 所以需要对连续型数据进行离散化和布尔型转换。在地质灾害监测数据中连续型数据均为数值型数据, 可以对连续型数据进行等级划分, 将数据归为多个类, 从而达到离散化的目的。

3.3. 数据对象的聚类

根据库区滑坡等地质灾害相关研究的成果和经验[9] [10], 结合降雨、土壤温度、孔隙水压力、含水量、水势、振动这些参数采用日平均数据并整理, 从中选取 3 种降雨因素, 即日降雨量、月累计降雨量、月最大连续降雨量和 3 种土壤因素, 即土壤温度、孔隙水压力、含水量, 共 6 种影响因子作为关联准则的前项, 监测数据中的测孔位移作为关联准则的后项, 如表 1 所示。

首先应用 K-means 算法对 3 种降雨因素和 3 种土壤因素进行离散化处理并划分为高和低两个等级。对于关联规则后项测孔位移同样根据 K-means 聚类将位移变化划分为两个阶段。

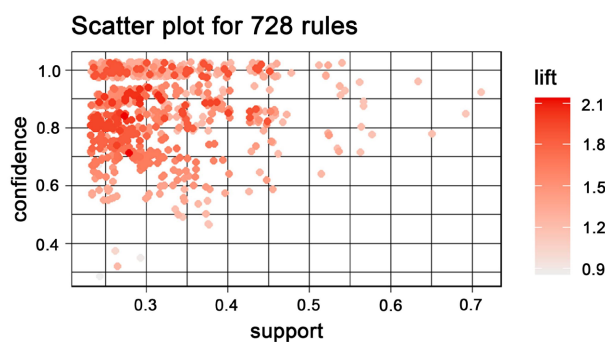
Table 1. List of types of influencing factors**表 1.** 影响因子种类表

降雨因素	标记	单位	土壤因素	标记	单位
日降雨量	RD	mm	土壤温度	T	℃
月累计降雨量	RC	mm	孔隙水压力	P	Kpa
月最大连续降雨量	RM	mm	含水量	W	m ³ /m ³

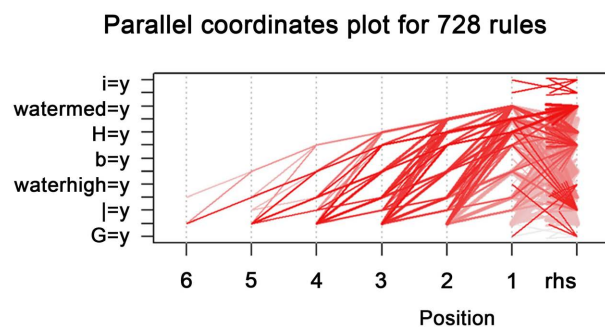
4. 滑坡监测数据关联规则挖掘

4.1. 关联规则挖掘

将 6 种影响因子作为关联前项, 滑坡位移量作为关联后项。设置最小支持度为 0.4, 使用 Apriori 算法共产生有效的关联规则 728 条, 并生成关联规则散点图如图 2 所示, 其中横坐标为支持度(support), 置信度(confidence)为纵坐标, 各点颜色的深浅表示提升度(lift)的大小。

**Figure 2.** Association rule scatter chart**图 2.** 关联规则散点图

将 728 条关联规则利用平行坐标图将关联规则可视化如图 3 所示, 其中 rhs 表示关联规则的后继项。

**Figure 3.** Association rule parallel coordinate graph**图 3.** 关联规则平行坐标图

由于关联规则数目过多, 部分存在明显错误, 因此对所生成的 728 条关联规则进行经验筛选, 剩余符合要求的有效关联规则 189 条, 选择其中置信度较高的 5 条关联规则, 整理得到滑坡位移影响因子的关联规则挖掘结果如表 2 所示, 其中后项为测孔位移。

Table 2. Association rule table
表 2. 关联规则表

序号	前项	后项	支持度	置信度
1	RD = 高 & W = 高	高	0.94	1
2	RM = 高 & RD = 高 & P = 高	高	0.66	1
3	RC = 高 & P = 高 & W = 高	高	0.63	1
4	P = 高 & W = 高	高	0.57	1
5	RD = 高	高	0.55	0.72

4.2. 关联规则分析

根据表 2 所示的五条关联规则挖掘结果可以看出, 测孔位移受降雨因素影响较大。由关联规则序号 1 可以看出, 当日降雨量高且土壤含水量大的时候, 测孔位移量大, 发生滑坡的概率增加, 支持度为 0.94。由关联规则序号 2 可以看出, 当日降雨量高、月最大连续降雨量高且孔隙水压力高的时候, 也容易发生较大位移。综合所有关联规则结果来看, 日降雨量和土壤含水量以及孔隙水压力三个影响因素对测孔位移的影响最为明显, 当监测到三类影响因子数据过高时, 应注意可能带来的滑坡隐患。

5. 总结与展望

本文提出了一种基于关联规则分析滑坡影响因素的方法, 通过 K-means 算法进行聚类, 使用 Apriori 算法进行关联规则挖掘, 分析了藏东南扎木弄巴滑坡观测数据间的关联规则。通过对历史滑坡数据的分析, 得出了滑坡发生的关键影响因素。该方法具有较高的准确性和可解释性, 在滑坡预测、预警以及灾害管理中具有潜在应用价值。未来的研究可以进一步提升模型的性能和稳定性, 拓展该方法在其他地区和领域的适用性。同时, 还可以结合其他机器学习算法, 进行更深入的滑坡影响因素研究。

致 谢

感谢青藏高原冰冻圈观测研究站提供的 2016~2019 年藏东南扎木弄巴滑坡观测数据。

基金项目

- ① 河北省重点研发计划项目“基于时空大数据及深度学习的地质灾害风险识别关键技术研究”(项目编号: 22375415D)。
- ② 2023 年河北省硕在读研究生创新能力培养资助项目“基于时空大数据的滑坡灾害监测预警模型研究”(项目编号: CXZZSS2023131)。

参考文献

- [1] 王涛, 刘甲美, 栗泽桐, 辛鹏, 石菊松, 吴树仁. 中国地震滑坡危险性评估及其对国土空间规划的影响研究[J]. 中国地质, 2021, 48(1): 21-39.
- [2] 陈锐, 范小光, 吴益平. 基于数据挖掘技术的白水河滑坡多场信息关联准则分析[J]. 中国地质灾害与防治学报, 2021, 32(6): 1-8.
- [3] 赵久彬, 刘元雪, 刘娜, 胡明. FRPFP 模型滑坡监测预警关联规则挖掘分析——以三峡库区江津到奉节段为例[J]. 岩土工程学报, 2019, 41(3): 492-500.
- [4] 朱鸿鹄, 王佳, 李厚芝, 叶霄, 施斌, 张勤. 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.
- [5] 金劭南. 基于大数据分析挖掘的铁路滑坡灾害监测预警模型研究[D]: [硕士学位论文]. 成都: 电子科技大学,

- 2021.
- [6] 刘文雅, 徐永能. 基于改进 Apriori 算法的地铁故障关联规则挖掘[J]. 兵器装备工程学报, 2021, 42(12): 210-215.
 - [7] Guo, L., Miao, F., Zhao, F., *et al.* (2022) Data Mining Technology for the Identification and Threshold of Governing Factors of Landslide in the Three Gorges Reservoir Area. *Stochastic Environmental Research and Risk Assessment*, **36**, 3997-4012. <https://doi.org/10.1007/s00477-022-02247-6>
 - [8] 陈宁生, 丁海涛, 邓明枫. 2016-2019 年藏东南扎木弄巴滑坡观测数据[J/OL]. 国家冰川冻土沙漠科学数据中心. <https://doi.org/10.12072/ncdc.BoMi.db0001.2020>, 2020-12-24.
 - [9] 刘文雅, 徐永能. 基于改进 Apriori 算法的地铁故障关联规则挖掘[J]. 兵器装备工程学报, 2021, 42(12): 210-215.
 - [10] 罗海洋. 基于 MapReduce 的关联规则挖掘算法的研究及应用[J]. 信息通信, 2019(2): 142-144.