文章编号: 1007-2993(2022)03-0185-05

数据仓库和数据挖掘技术在滑坡预测预报中的应用

秦文涛　郭小坤　郭军峰　郝　璐　洪碧武
（西南交通建设集团股份有限公司, 云南昆明　650000）

【摘要】 为从滑坡历史数据中获取有效的预测预报信息, 利用基于数据仓库的数据挖掘技术, 结合已有滑坡多个维度的数据集, 以巴东县新城区滑坡灾害分布作为研究对象, 建立了滑坡灾害敏感性和灾害区划模型。结果表明: 该模型对于滑坡空间分布的预测精度可达到 87.5% 左右, 对于滑坡时间尺度上的预测精度较低, 约为 65% 左右, 其准确度可以满足工程要求; 数据仓库和数据挖掘技术在地质灾害预测预报领域具有广泛的应用前景, 较传统预测预报方法更为方便和迅捷。

【关键词】 数据挖掘; 数据仓库; 滑坡预测; 地质灾害
【中图分类号】 P 642　【文献标识码】 A　doi: 10.3969/j.issn.1007-2993.2022.03.003

Application of Data Warehouse and Data Mining on Landslide Prediction

Qin Wentao　Guo Xiaokun　Guo Junfeng　Hao Lu　Hong Biwu
(Southwest Transportation Construction Group Co., Ltd., Kunming 650000, Yunnan, China)

【Abstract】 In order to obtain effective prediction from the historical landslide data, the data mining technology based on data warehouse were used. Combined with the existing landslide multi-dimensional data set, taking the landslide disaster distribution in the new urban area of Badong County as the research object, the landslide disaster sensitivity and disaster zoning model are established. The result shows that the prediction accuracy of the model for the spatial distribution of landslide can reach about 87.5%, and the prediction accuracy on the time scale of landslide is low, about 65%. Its accuracy can meet the engineering requirements. The data warehouse and data mining technology in the field of geological disaster forecast has wide application prospect, which is more convenient and rapid than traditional prediction methods.

【Key words】 data mining; data warehouse; landslide prediction; geological hazard

0　引言

数据仓库和数据挖掘技术从 20 世纪 90 年代开始在商业领域得到广泛应用, 在金融业、电子技术、图像处理等海量数据密集行业的应用尤为广泛和成熟。如王冬梅[1] 结合医院现有的 HIS 系统和数据挖掘技术实现辅助医疗诊断; 王　云等[2] 提出用多维关联规则技术分析交通事故记录, 用于识别和发现事故发生规律和起因; 陈　起[3] 将数据挖掘应用于电信客户细分中, 实现了数据分群操作; 数据仓库的设计方式与传统数据库的组织和应用方面具有很大不同, 其应用领域仍然有待继续挖掘。

在滑坡地质灾害预测预报领域, 由于滑坡相关的工程地质数据在空间和时间上具有不确定性, 数据仓库的应用非常少见, 对于数据仓库的认识也亟待建立和应用[4−6]。目前关于数据仓库和数据挖掘技术与地学的结合, 最著名的是加拿海洋深度数据仓库和美国国家水质评价数据仓库, 前者利用 Oracle 关系数据库开发出水平和垂直方向的数值地形模型; 后者的联机数据库保存约 700 万条记录, 用于监测全美 46 个州大部分河流水质数据。

本文根据数据仓库技术, 在对滑坡敏感性因子成因分析的基础上, 建立了滑坡敏感性多维数据模型, 将滑坡区的致滑因子空间数据按不同地区、不同类型储存于滑坡灾害数据仓库中, 实现了滑坡预测预报的数据快速响应, 为地区滑坡地质灾害防治提供了针对性建议。

作者简介: 秦文涛, 男, 1986 年生, 汉族, 甘肃泾川人, 硕士研究生, 高级工程师, 主要研究方向为高速公路路基、桥梁、隧道施工。
E-mail: tingtcc056@sina.com

1 滑坡数据仓库设计

滑坡灾害防治需要对各类地质资料数据进行整合与分析，这些数据包括空间上的数据如滑坡地点、岩性条件、水文地质条件和地形地貌等，以及时间尺度上的滑坡变形位移、历史滑坡等。数据从类型上可以划分为空间数据、时间数据和管理数据三类，数据存放于不同的操作数据库中，其目的是便于在预测预报时从中准确挖掘出有用数据，具有面向主体、高集成度、历时快和能够快速检索等特点。

1.1 设计原则

滑坡灾害数据仓库设计中尽量采用已有的 Oracle 关系数据库，防止研究新型数据库技术所导致的不成熟和不稳定，保证系统扩展性好、易于维护和方便快捷。在数据仓库的逻辑结构上，分为数据获取、管理和使用三步。

1.2 多维模型

数据仓库基于多维模型，该模型可以更好地理解数据分析的目的，且适用于复杂分析查询，多维模型由维和事实描述[7]。事实包括若干个相关的维，维用于描述属性来提供上下文。所有的维被组成不同的聚集层次，这样使得事实的度量能够基于不同细节进行分析。在建立滑坡灾害模型中它能够帮助使用者理清数据来源，因此形成开一个对数据的具体求解方法。

1.3 数据仓库体系

数据仓库体系主要分为三层：数据获取层、数据存储层和数据访问层[8]。三个层次分别对应着后端层、数据仓库层以及用户层。其中，后端层用于传回数据，在原始数据上传至数据仓库之前，其主要作用是对数据进行集成和转换；数据仓库层的主要作用是保存数据；用户层主要用于处理和分析数据，包括报告、统计以及数据挖掘等直接面向用户的操作。体系结构图见图 1。

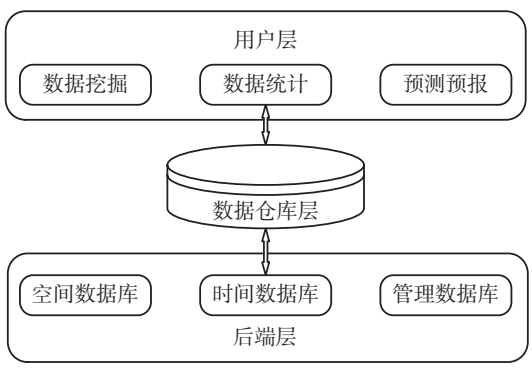

图 1 数据仓库体系结构

1.4 数据仓库设计

数据仓库的设计参照传统数据库设计进行，滑坡地质灾害数据仓库设计过程中采用“数据驱动”的设计思路，包含如下两个方面：

(1)尽量利用已有数据，在已有的基础数据库上进行建设，这是数据驱动的出发点。

(2)面向主体而非面向应用，从数据系统出发，按照地质灾害领域的要求设计数据之间的联系，来组织数据仓库中的主题。

本文所研究的基于地质环境数据仓库 ETL 的滑坡数据仓库设计内容包括数据选择、转换、清晰和加载，具体架构实现过程见图 2。

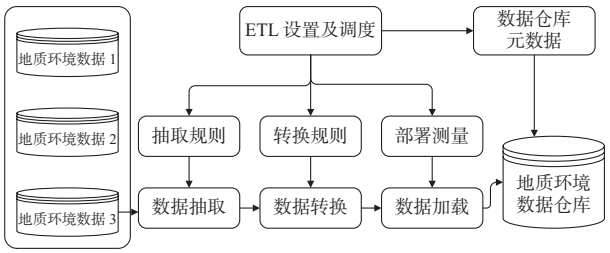

图 2 数据仓库设计流程

2 概念模型

滑坡预测预报的概念模型包括两方面：第一是模型的事实、维和划分标准；第二是多维模型的架构设计。滑坡数据仓库建立过程中，逐步对每个主题进行求解和分析。本文设计了两个主题：滑坡预测预报主题与滑坡数据监测主题，每个主题的维、度量方法和划分层次不尽相同。

2.1 滑坡预测预报

滑坡地质灾害预测预报是在数据仓库和数据挖掘的基础上，针对滑坡区域的致滑因子搜集数据并进行统计分析，然后对此区域的滑坡危险性进行评价，为地质灾害的预测预报提供可靠依据。在数据统计过程中，致滑因子的数据来源可分为如下几大类：

(1)所在区域地形地质图、地质灾害分布图、植被覆盖情况、土地规划图。

(2)区域内滑坡历史数据，包括滑坡分布地点、滑坡体外貌、坡脚特征，岩土体性质、历史滑动数据等具体数据。

(3)环境资料，包括气候条件、温湿度变化、降雨量等。

(4)经济建设与人文资料，包括研究区域人文地理，重要建筑物、道路桥梁隧道、地下管线等与人类活动相关的工程设施分布情况数据。

(5)现场监测数据，主要是危险滑坡体的变形监

测数据及防治工程监测数据。

(6)各级相关部门通讯方式和资料。

(7)国家和地区相关政策法规关于滑坡地质灾害预测预报的规定。

2.2 层次划分

在数据收集的基础上,通过对滑坡成因分析,从而将致滑因子划分为如下四个层次:

(1)区域滑坡概况,包括一个地区滑坡的地理位置、滑坡类型(土质滑坡、岩质滑坡、变形体等)、每个滑坡组成部分(滑坡体、滑动面、滑坡床、滑动带等)的特征、滑坡发生频率以及所有相关的数据。

(2)赋存环境,包括滑坡的岩土体结构构造、地层岩性、地质构造以及相关的水文条件(如水系分布等)。

(3)诱发因素,包括降雨、地震、人为活动(植被破坏、修筑建筑物、爆破等)。

(4)潜在受灾对象,滑坡周围可能因受到潜在危险而导致经济损失的地区,如人口、建筑物、生态环境等。

2.3 评价指标确定

滑坡预测预报基于滑坡敏感性区划来进行,对于滑坡敏感性的区划,需要对导致滑坡形成的所有因子的贡献进行权值划分,并对每个因子的贡献大小定量化,评价指标的目的就是通过深入分析滑坡形成因素的基础上对致滑因子定量化取值。各个致滑因子的评价指标如下:

(1)滑坡分布:对于所在区域各地点是否有滑坡,已知滑坡由数 1 表示,不存在滑坡用数值 0 表示。

(2)滑坡结构:滑坡所在地区地层岩性是控制滑坡发生的重要条件,由松散堆积体、碎石土、风化壳组成的坡体抗剪强度低易于发生滑坡,坚硬岩质坡体抗剪强度高不易发生滑坡,坡体中存在软弱滑动面的斜坡在触发因素作用下易发生滑塌;从顺向坡、顺斜坡、横向坡、逆斜坡至逆向坡,滑坡危险性逐渐降低。

(3)滑坡坡度:将滑坡坡度分为五个维度,分别为 0° ~ 15°、15° ~ 30°、30° ~ 45°、45° ~ 60°以及 > 60°。坡度越大,发生失稳破坏的可能性越高,其赋值从 1 至 0 呈 5 级递减。

(4)海拔高度:通过分析工程区滑坡分布高度发现,海拔越低地方滑坡体发生概率越小。根据海拔高度的不同统计不同类型滑坡数量,本文将海拔高度按小于 100 m, 100 ~ 200 m, 200 ~ 250 m, 250 ~ 300 m 和大于 300 m 分成 5 个级别,海拔高度较低区域发生滑坡的风险性越低。

(5)水系分布:河流对于滑坡坡脚底部侵蚀作用非常明显,冲蚀掏空直接导致滑坡底部产生临空面,从而使滑动面暴露于外部环境中;通过统计滑坡所在位置与河流水系的距离,对其危险性分级,其距离以 200 m 作为分界线,在河流流域 200 m 以内的滑坡受影响作用明显,200 m 以外的影响忽略不计。

(6)人类工程活动:人类工程活动(如修筑公路开挖坡脚)对于滑坡影响较大,因此以工程建筑物与滑坡距离作为衡量尺度进行分级。本文设定滑坡区与公路的距离来进行量化,距离公路直线距离 100 m 作为分界,100 m 以内滑坡受影响明显,100 m 以外影响忽略不计。

3 系统设计

3.1 界面设计原则

在不妨碍用户使用方便快捷的基础上,尽量保证系统的界面布局相同、操作方式一致、对信息的读取一目了然。

3.2 功能设计

(1)数据管理:数据管理主要指对数据仓库的实际操作功能的设计,对于滑坡数据中多维度、模型和 ETL 设计都必须及时同步至数据库,对于数据库用户设置不同权限便于其对数据仓库进行整理,对数据资料做好备份防止丢失。

(2)数据显示:方便用户直观地了解数据变化情况,目标区域降雨量、地震活动、人类活动等触发因素,以及所在区域岩性条件、地质构造、水文地质等环境条件等,都需要及时准确显示。

(3)数据挖掘:方便用于快速浏览和提取特征数据进行对比分析,为滑坡预测预报提供详尽资料。

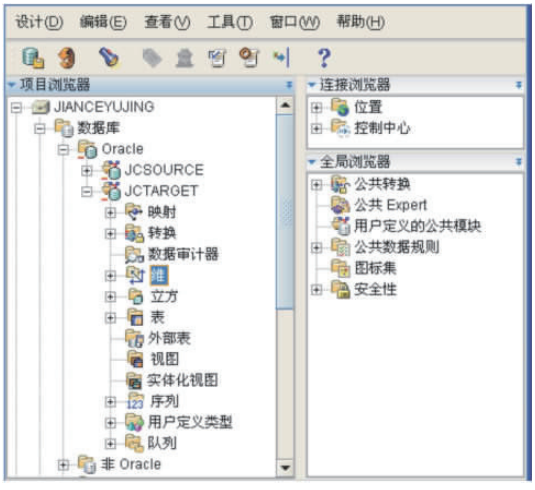

图 3 滑坡预测预报系统界面

图 3 是滑坡预测预报系统的管理员界面，通过此界面，管理可以实时对数据库进行更新操作，调用某个特定数据进行特征分析，并可以将不同地区降雨量与滑坡危险性相关联，利用实体化视图显示出来。

4 实例分析

研究区为巴东县，其城区面积约为 1 km^2，此区域属于四川盆地，地形深切割明显，冲沟发育较多，降雨量大，其详尽地质资料见文献 [4]。

4.1 预测变量

根据巴东县数据收集情况，选择如下 5 个因子的数据进行数据整理：滑坡海拔高程、坡度、滑坡结构以及与水系、公路的距离。其中，前三个因子需通过滑坡区地形地质图，结合滑坡产状、地层条件等综合分析得出，后两个因子可直接测量得到。在统计分析过程中，对巴东县滑坡危险性区划的分割选择 100 m^2 的正方形面积作为最小单位，每个单位面积统计 5 个因子权值按照给定模型计算出其危险性系数。权值分布见表 1。

表 1 影响因子权值分布

影响因子	各个因子的权值分布				
滑坡坡度/(°)	0 ~ 15	15 ~ 30	30 ~ 45	45 ~ 60	> 60
权重	4	3	2	1	0.5
海拔高度/m	< 100	200 ~ 100	250 ~ 200	300 ~ 250	> 300
权重值	5	4	3	2	1
滑坡结构	逆向坡	逆斜坡	横向坡	顺斜坡	顺向破
权重值	5	4	3	2	1
水系/m	> 200	< 200			
权重值	1	0			
公路/m	> 100	< 100			
权重值	1	0			

注：表格中水系表示滑坡点距离水系的距离，m；公路表示滑坡点距离公路的距离，m。

4.2 数据叠加模型

滑坡影响因子的叠加模型选择 Logistic 回归模型进行分析，逻辑回归模型并没有直接说明滑坡发生的可能性，而是用量化的概率进行推导，逻辑回归模型用概率计算公式进行叠加操作[9-10]：

$$P = \frac{1}{1+e^{-Y}} \tag{1}$$

式中：P 为滑坡发生概率；Y 为拟合因变量，Y 的计算公式为：

$$Y = C_0 + C_1X_1 + C_2X_2 + \cdots + C_nX_n \tag{2}$$

式中：$C_0 \sim C_n$ 为相关系数，代表贡献率大小；X 为因子的值。

通过逻辑回归模型计算出 5 个致滑因子的相关性大小分别为：坡度(0.488)，海拔高程(1.18)，河流距离(16.263)，公路距离(4.19)，结构(0.205)。滑坡敏感性区划见图 4，图中红色部分代表滑坡灾害发生概率较高，绿色部分说明地区受滑坡灾害影响较小，安全系数较高，从红色到绿色滑坡危险性逐渐降低。

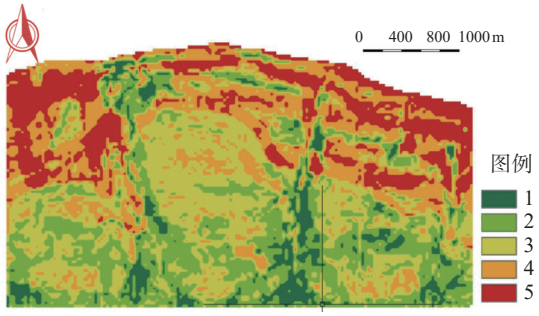

图 4 巴东县滑坡敏感性分布图

5 结论

本文将数据仓库、数据挖掘技术与地质灾害相关知识结合，以巴东县为例建立了基于数据仓库的滑坡灾害区划与预测预报模型，得出了巴东县新城区的滑坡危险性区划图，并形成如下结论：

(1)数据仓库和数据挖掘技术具有高度集成性、面向用户和快速的数据支持与决策等优势，能够对滑坡灾害数据进行分级重组，检索速度快；在滑坡预测预报过程中，可以根据分析统计在海量的数据库中挖掘有效的地质灾害信息，使得预测过程迅速、准确。

(2)在对滑坡成因深入分析的基础上，设计出影响滑坡发生五个致滑因子：坡度、海拔、结构以及与公路、水系的距离，将这五个致滑因子通过逻辑回归模型加权分析后，存放于滑坡灾害数据库中，满足滑坡空间分布预测的需求。

(3)本文所述数据库模型对滑坡致滑因子的考虑还不够精细(如未考虑降雨量、地震和人类工程活动等因素)，预测预报模型的可靠性也还需要时间检验，期待在以后的滑坡预测数据库建设与挖掘工作中进一步细化和完善。

参 考 文 献

[1] 王冬梅. 数据挖掘在医院信息系统中的研究与应用[J]. 科学技术与工程, 2007, (11): 2745-2747.

[2] 王 云, 苏 勇. 关联规则挖掘在道路交通事故分析中的应用[J]. 科学技术与工程, 2008, (7): 1824-1827.

[3] 陈 起. 数据挖掘在电信客户细分中的应用[J]. 科学

技术与工程, 2009, 9(16): 4820-4822, 4832.

[4] 杜 娟, 殷坤龙, 陈丽霞. 基于GIS的巴东县新城区滑坡灾害危险性区划[J]. 自然灾害学报, 2011, (1): 149-155.

[5] 张庭瑜, 韩 玲, 张 恒, 等. 混合分类模型在滑坡易发性分区中的适用性研究——以延安市宝塔区为例[J]. 干旱区资源与环境, 2020, 34(1): 192-201.

[6] 朱 玲. 基于不确定数据的IM-K-means算法在滑坡危险性预测的应用[D]. 赣州: 江西理工大学, 2019.

[7] 文海家, 李 洋, 薛靖元, 等. 基于大数据挖掘的山区公路沿线滑坡易发性小区划[J]. 自然灾害学报, 2018, 27(4): 159-165.

[8] 刘鹏程. 大数据分析技术在地质灾害系统中的应用研究[D]. 西安: 西安工业大学, 2018.

[9] 赵久彬, 刘元雪, 宋林波, 等. 大数据关键技术在滑坡监测预警系统中的应用[J]. 重庆理工大学学报(自然科学), 2018, 32(2): 182-190.

[10] 罗海洋. 大数据环境下关联规则挖掘算法及其应用研究[D]. 长沙: 湖南大学, 2017.

收稿日期: 2021-03-31