

黄土力学数据挖掘系统研究

井彦林^{1,2}, 作彦卿^{1,3}

(1. 西安理工大学, 陕西 西安 710048; 2. 煤炭工业西安设计研究院, 陕西 西安 710054; 3. 上海交通大学, 上海 200240)

摘要: 岩土工程智能化系统研究是岩土工程领域前沿问题。本文基于信息技术中最先进的数据挖掘技术, 研制了黄土力学数据挖掘系统, 包括有预处理模块、挖掘操作模块、知识库管理模块, 自重湿陷与湿陷性判定决策树模型、判定规则和评价模型的精度以及其他应用软件接口等。可实现数据的归约、聚类分析、分类、预测等。通过工程实例分析, 证明了该系统的有效性和实用性。采用 59 项工程的 2766 组黄土试验数据对决策树的每条规则进行测试, 结果表明自重湿陷性判定规则的准确率为 87.3%, 湿陷性判定规则的准确率为 92.5%。利用挖掘所得到的判定规则和决策树模型, 可减少试验工作、降低成本; 提出的模型还可用于确定湿陷底界与自重湿陷的底界, 从而为地基处理深度的确定提供依据。

关键词: 黄土力学; 数据挖掘系统; 决策树模型

中图分类号: TU 444

文献标识码: A

文章编号: 1000-4548(2005)10-1154-05

作者简介: 井彦林 (1963-), 男, 西安理工大学岩土工程学科在职博士生, 主要从事黄土力学与工程、环境岩土工程研究。

Data mining system of loess mechanics

JING Yan-lin^{1,2}, WU Yan-qing^{1,3}

(1. Xi'an University of Technology, Xi'an 710048, China; 2. Xi'an Design & Research Institute of Coal Industry, Xi'an 710054, China; 3. Shanghai Jiaotong University, Shanghai 200240, China)

Abstract: The intelligent system of geotechnical engineering is an important research interest in geotechnical engineering field. A data mining system of loess mechanics was developed based on advanced data mining system in information technology. The data mining system of loess mechanics included the preprocessing module, mining operation module, knowledge base management module, and software interface of geotechnical application. The system could be applied to data reduction, cluster, classification, and prediction for loess mechanics. Through a lot of engineering applications, the results indicated that the system was effect and practical for loess collapse. Through the assessment of each rule in decision trees model by using 2766 groups of loess testing data in 59 projects, the results show that the precision discriminating self weight collapse loess is about 87.3%, and the precision discriminating collapse loess is about 92.5%.

Key words: loess mechanics; data mining system; decision trees model

0 前言

数据挖掘 (Data Mining, 简称为 DM) 是上世纪 90 年代中期迅速发展起来的一门交叉学科, 它涉及到数据库技术、统计学、人工智能与机器学习等多个领域。数据挖掘是从大量的、不完全的、有噪声的、模糊的、随机的实际数据中提取隐含在其中的、不为人们所知的、但又有潜在应用价值的信息和知识的过程。它是知识发现 (Knowledge Discovery in Database, KDD) 中的关键环节。数据挖掘的主要对象是数据 (仓库); 任务包括特征规则、辨识规则、关联规则、分类规则挖掘, 数据聚类、分类、预测、趋势性规则挖掘, 公式发现、可视化、数据纠正等; 算法有归纳

学习法、决策树、遗传算法、神经网络、统计分析、模糊数学、关联分析、粗糙集等; 一个完整的数据挖掘系统包括多种算法和多种功能, 但它不是算法的堆积, 而是多种信息技术系统地、综合地应用。数据挖掘的结果输出有规则、网络权值、图形、报告、逻辑公式等形式。数据挖掘和传统的数据分析有着根本的区别。传统的数据分析是对过去事实的描述; 而数据挖掘则是预测未来将要发生的事情, 并解释过去发生的原因; 传统数据分析中的数据是根据研究的具体需要通过专项试验获得的, 数据挖掘研究中使用的数据

基金项目: 煤炭工业西安设计研究院科研基金资助项目

收稿日期: 2005-01-04

则不是为了挖掘研究而产生和收集的,而是在日常事务中出于其它某种目的而产生和积累的。

近年来,在数据挖掘方面的研究特点是研究算法的多,进行应用研究的少;在应用方面,运用单一算法进行研究的,而进行系统研究的少。国内外已有一些数据挖掘系统,分专用型与通用型,通用型如 SAS、DBMiner、IntelligenMiner、Golden-Eye^[1,2]等,专用型如 RSCHMiner^[3]等。每一种挖掘系统都是针对某种特定的目的开发的,所以要将流行的挖掘系统应用到岩土工程领域,都必须花费相当的精力进行二次开发;数据挖掘的目的在于将挖掘到的知识(结果)进行应用。如果能将数据挖掘系统与岩土工程的应用软件集成化、系统化,则更利于数据共享、功能的互相调用、充分发挥信息技术在岩土工程领域的作用。但将岩土工程软件同现有的通用挖掘系统进行集成是较困难的。基于此,结合岩土工程技术的特点,尤其是湿陷性黄土物理力学指标的分布特点开发了本系统,以便于高效地进行(黄土)岩土工程的智能化分析和系统分析,使挖掘系统与岩土工程应用软件如黄土岩土工程评价分析系统^[4]进行集成,挖掘结果可通过应用软件的调用直接转化为工程所需的技术报告,为设计和施工服务。显然,本系统为专用型挖掘系统。

1 黄土力学数据挖掘系统的结构

图 1 为本系统的结构。系统包括预处理模块、挖掘操作模块、数据库管理模块、知识管理模块等^[2,3,5]。与一般挖掘系统所不同的是,该系统的挖掘结果(知识)可通过知识管理模块被(现有的)岩土工程应用软件直接调用,而岩土工程应用软件又可将其原始数据或计算结果追加至数据库。

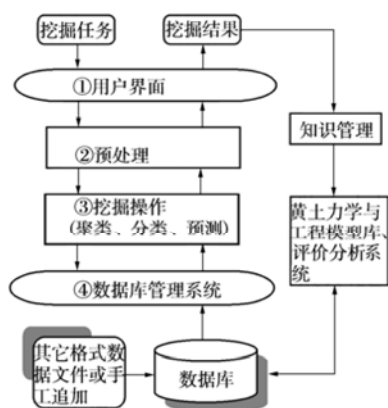


图 1 黄土力学数据挖掘系统结构流程图

Fig. 1 Data mining system for loess mechanics

1.1 预处理模块

源数据库中常含有不完整的、含噪声的、不一致的数据,这些数据能使挖掘过程陷入混乱,导致不可

靠的输出。预处理是对这些数据进行处理、清洗和同化的过程。它属于数据准备,是数据挖掘中最关键的步骤之一,占全部数据挖掘的 40%~90%的时间。数据预处理技术在数据挖掘之前使用,它为挖掘操作提供有效而可靠的输入,从而提高数据挖掘模式的质量,降低实际挖掘所需要的时间。

1.2 数据库管理模块

该模块的主要功能之一是建立源数据库。可通过岩土工程应用软件将以往工程中的大量数据转入数据库,可将其它格式的数据文件中的数据转入数据库,也可以用手工输入对数据库进行追加;对于高级用户,可直接使用标准 SQL 语言对源数据库进行各种查询,包括智能查询。它可将某一阶段的挖掘结果进行存储,作为下一步挖掘的输入。挖掘系统通过 ODBC 使执行程序调用数据库,这样使系统对数据库的类型具有广泛的适用性。该模块起到了外部文件、数据源、挖掘操作和挖掘结果的桥梁作用。

1.3 知识管理模块

包括模型评价、规则抽取、知识解释。可将挖掘到的模型、模式进行评价存储、使用等,如使用模型进行预测;在数据库中产生新的属性,如使用湿陷性指标的预测模型预测湿陷系数;将新的属性保存为指定的数据格式或建立新的数据库,便于岩土工程应用软件的调用,从而进行评价分析、曲线及图件绘制,为设计和施工提供技术报告。同时可实现知识库与源数据库的互连,知识库、挖掘系统以及岩土力学与工程应用软件的互连。

1.4 数据挖掘操作模块

该模块根据要求、用经过预处理后的数据完成各种挖掘任务,如进行分类、聚类、预测、关联分析等挖掘操作。每种挖掘技术构成一个子模块,它们在功能上是相互独立的。挖掘结束后将挖掘到的知识写入知识库。每种挖掘技术包含一些不同的具体算法。

系统主界面及粒子群优化最小二乘支持向量机(PSO LS-SVM)预测模块界面见图 2。

2 黄土力学数据挖掘系统功能

2.1 数据准备

该阶段主要是进行数据的预处理。系统通过 ODBC 将数据源中的数据库调入,对数据进行清洗、数据归约、格式转换、孤立点分析等。

(1) 数据归约:有些数据集很大,可能相关集,即数据间信息重复,这会导致挖掘速度太慢。可通过数据归约的方法消除数据间的相关性,形成正交集。它比原数据集要小得多,减少了出现在发现模式上的



图 2 系统主界面及粒子群支持向量机预测模块界面

Fig. 2 Interfaces of system and PSO LS-SVM prediction module

属性数目,但却能够产生几乎完全相同的分析结果,可使挖掘结果更易于理解,使挖掘出的模型得到简化。本系统提供基于主成份分析技术的数据归约算法^[6,7]。

(2) 孤立点分析:有的数据集中存在野值,它具有同其它数据明显不同的特性,它常会对挖掘结果产生巨大的负面影响,需要对其进行识别并剔除。系统提供基于聚类的孤立点分析等技术。

(3) 数据变换:不同的挖掘模型需要不同格式的输入数据。如 BP 神经网络等,要求将数据按比例映射到一个特定区间[0.0, 1.0],即要求进行数据的归一化;有的算法不能处理连续型数据或非数值型数据,这就需要将其离散化、数字化。数据库中的数据不一定全部参与挖掘,可通过 SQL 语言进行属性选择的方法实现特征提取。

2.2 分类

分类也称有指导的学习,是要得到一个分类模型或分类函数(也称分类器),该模型可将给定输入映射到给定类别中的某一个。分类也可用于预测。由于分类只能用于离散数据,所以分类法用于预测时需将连续数据离散化。本系统的分类算法包括 BP 神经网络、径向基神经网络、概率神经网络、学习矢量量化神经网络、CART 决策树算法^[8~13]等。

2.3 聚类

聚类也称无指导的学习,它在数据挖掘研究中占有很重要的位置。所谓聚类,是从数据集中找出相似的数据并划分为不同的组,也就是不同的类。本系统的聚类分析算法包括自组织竞争神经网络、自组织特征映射神经网络等^[14~16]。

2.4 预测

根据训练样本,用给定的变量值对预测的目标变量建立预测模型,同时可对给定的输入对其相应的输出做出预测。预测和分类很类似,但也有所不同。分类是用于离散值,预测用于连续值。本系统中的预测算法包括 BP 神经网络、径向基神经网络、概率神经

网络、CART 决策树算法、粒子群最小二乘支持向量机算法等。在实际应用上, CART 决策树与非线性回归常结合使用。

3 应用实例

黄土的湿陷系数及自重湿陷系数是黄土的重要力学指标,在实际工作中其测试、试验工作量大,费用高。如果能用较为稳定的、常用的物理指标描述其湿陷性分布规律,判断其湿陷性,降低实际工作中的试验费用,无疑具有很高的实用价值;针对此问题,用数据挖掘技术展开研究工作。挖掘研究中的关键问题是原始数据的收集,建立源数据库。通过对众多科研单位多年来试验资料的分析、筛选,共收集到 498 项工程的 14008 组黄土试样数据,数据来源于陕西关中及其周边地区。指标类型包括含水量、密度、液限、塑限、压缩系数、静探指标、湿陷性指标等。数据库管理系统采用 MS SQL Server 2000 开发;研究的目的是挖掘湿陷性与非湿陷性、自重湿陷性与非自重湿陷性的划分规则。因数据库中的数据来自科研单位,在深度大于 10 m 时各单位所采用的测定湿陷系数的压力标准并不统一,所以数据库中只包括取土深度在 10 m 以内的土性指标;测定湿陷系数的试验压力为 200 kPa,测定自重湿陷系数的试验压力为饱和自重压力。

3.1 指标选择与数据压缩

根据大量的研究成果,国内外都是从黄土的紧密度和含水性两方面来评价黄土的湿陷性,同时湿陷性与深度也有关系^[17,18];对于自重湿陷系数,试验压力为饱和自重压力,所以它与深度的关系就更为密切。因此本次挖掘选用的输入变量为取土深度及其它常规物理力学指标如含水率、干密度、液限、塑限、压缩系数 a_{1-2} 。由于数据中存在冗余信息,如含水率与压缩性关系较为密切,所以首先应进行数据压缩,以消除各指标间的关联性。压缩方法采用系统中提供的主成份分析法(PCA)。主成份分析法(详见文献[6]、[7])

计算步骤如下

(1) 将原始变量 x 进行标准化

$$x_{ij}^* = \frac{x_{ij} - \bar{x}}{\sqrt{\text{var}(x_j)}}$$

$$i = 1, 2, \dots, n; j = 1, 2, \dots, p$$

式中 $\bar{x}$ 为平均值; $\sqrt{\text{var}(x_j)}$ 为标准差; n 、 p 分别为试样数据组数及每组中指标个数。

(2) 计算 x^* 的相关矩阵 R

$$r_{ij} = \frac{1}{n} \sum_{t=1}^n x_{ti}^* x_{tj}^*$$

(3) 选择主成份, 计算各主成份得分

首先求相关矩阵 R 的特征值及特征向量 u , 然后计算各主成份的贡献率及累计贡献率。主成份的个数视具体问题而定, 一般取累计贡献率为 80%~90%所对应的 m 个变量作为主成份 ($m < p$)。本例中, 前 3 个主成份所占的信息量已达到总信息量的 81.5%, 所以在此选择前 3 个主成份 ($y_1 \sim y_3$) 代替原变量。 y_1 、 y_2 、 y_3 为经标准化后的原变量的线性组合, 组合系数为相应的特征向量, 其值如表 1 所示。主成份得分即新变量按公式 $y = u^T x^*$ 计算。

表 1 主成份分析法得到的黄土物理力学指标特征向量值

Table 1 The eigenvectors of physical and mechanical indicators of loess given by principal component analysis

变量名称	取土深度	含水率	干密度	液限	塑限	压缩系数 a_{1-2}
y_1	-0.1512	-0.4659	0.1541	-0.6152	-0.5971	-0.0356
y_2	-0.4505	0.0668	-0.5573	-0.0453	-0.0763	0.6886
y_3	-0.6819	0.3819	0.6227	0.0296	0.0036	0.0232

3.2 决策树模型

自重湿陷系数和湿陷系数以 0.015 为界, 均可将黄土分为两类。针对此分类问题, 选用 CART 决策树模型进行挖掘。基于 CART (Classification and Regression Trees) 算法的决策树技术是一种非参数、非线性的预测分析方法。其特点是在计算过程中使用二叉树结构, 即根节点包含所有样本, 在一定的分割规则下被分为两个子节点, 此过程在子节点一直重复

进行, 直至不可再分为止。所构建的预测准则形式易于理解、使用、说明和解释, 易于形成 if-then 规则。构建决策树首先确定最佳分割阈值及最佳分组变量。在数据挖掘中, 采用杂质函数测度目标变量值在每个结点内的差异程度。杂质函数值越小, 节点内样本的类别越少; CART 算法中最常使用的是以基尼系数 $Gini(T)$ 表示杂质函数。CART 算法(详见文献[11~13])的建模步骤如下。

(1) 设训练集中有 N 类样本, 样本容量为 n , 根节点的基尼系数为

$$Gini(T) = 1 - \sum_{j=1}^N P_j^2$$

(P 为每一类在总体中的概率)。

(2) 指定分组变量的某个取值为分割阈值, 计算各组的基尼系数, 再加权平均基尼系数

$$Gini_{split(i)}(T) = \frac{1}{n} \sum_{j=1}^N n_{ij} Gini_i(T)$$

(3) 计算根节点与其下一层基尼系数加权值的差值

$$\Delta = Gini(T) - Gini_{split(i)}(T)$$

取基尼系数差值最大的 Δ_{max} 所对应的分割阈值为最佳分割阈值。对 3 个输入变量重复以上过程, 取基尼系数差值最大的分组变量为最佳分组变量。按此原理对每个叶节点进行分割, 直到不可再分或叶节点样本数小于某个数为止。然后再通过交叉验证法对原决策树进行剪枝, 得到最终的决策树。自重与非自重湿陷、湿陷性与非湿陷性黄土划分的决策树见图 3、图 4, 均为经剪枝后的决策树。图 3 终端节点上的 1, 2 分别代表具、不具自重湿陷性, 图 4 终端节点上的 1, 2 分别代表具、不具湿陷性。由图 3 及图 4 可以得出分类规则, 规则内容见表 2。

4 结果分析与讨论

通过上例的挖掘研究, 取得了黄土的自重湿陷与非自重湿陷、湿陷性与非湿陷性划分的决策树模型及其相应规则。为了评价模型的精度, 采用 59 项工程的

表 2 判定规则及测试结果

Table 2 Discrimination rules and validation results

规则类型	规则编号及内容	测试样本数	分类准确率	平均准确率
自重湿陷性判定	①if $y_3 < -0.950$ then 具有自重湿陷性	114	83.3%	87.3%
	②if $y_2 < -0.497$ and ($y_3 \geq -0.950$ and $y_3 < -0.279$) then 不具自重湿陷性	81	66.7%	
	③if $y_2 \geq -0.497$ and ($y_3 \geq -0.950$ and $y_3 < -0.279$) then 具自重湿陷性	173	68.8%	92.5%
	④if $y_3 \geq -0.279$ then 不具自重湿陷	2122	89.9%	
湿陷性判定	⑤if $y_2 < -1.658$ then 不具湿陷性	1017	96.3%	92.5%
	⑥if $y_2 \geq -1.648$ then 具湿陷性	1473	89.9%	

2766 组黄土试验数据对决策树的每条规则进行测试, 结果表明自重湿陷性判定规则的准确率为 87.3%, 湿陷性判定规则的准确率为 92.5% (见表 2)。利用挖掘所得到的判定规则或决策树模型, 可根据含水率、干密度等 6 项基本指标先判断土样是否具湿陷或自重湿陷性; 若不具湿陷性或自重湿陷性, 则可考虑不进行湿陷性或自重湿陷性试验, 这样起到了减少试验工作、降低成本的作用; 对于进行湿陷性试验的土试样用此模型进行判定, 可通过实测值与预测值的对比对试验结果进行辅助检验; 此模型还可用于确定湿陷的底界与自重湿陷的底界, 从而为地基处理深度的确定提供依据。

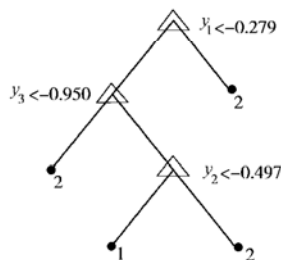


图 3 黄土的自重湿陷性判定决策树

Fig. 3 Discriminating decision trees of self-weight collapsing property of loess

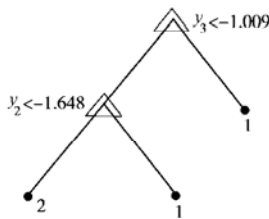


图 4 黄土的湿陷性判定决策树

Fig. 4 Discriminating decision trees of collapsing property of loess

5 结 论

(1) 系统成功地融合了数据挖掘的多种技术与岩土工程应用软件。

(2) 通过对黄土力学数据库的挖掘, 提出了具有一定应用价值的黄土的自重湿陷性与湿陷性判定规则, 体现了本系统的有效性和实用性。

(3) 系统中所采用的算法是通用的, 系统是针对黄土地区的特点开发的, 但在应用中, 对数据略加变换, 则可用于其它方面的数据挖掘。

(4) 本文实例分析中所使用的黄土力学数据库还有较大的挖掘潜力, 可通过预测性挖掘直接预测出黄土的自重湿陷系数、湿陷系数、湿陷起始压力等湿陷性指标, 这将是下一步应研究的问题。

(5) 系统中的有些算法在对大数据库进行挖掘时, 计算速度较慢, 所以在算法和程序结构优化方面还有待改进。

参考文献:

- [1] 张雪英. 国外先进数据挖掘工具的比较分析[J]. 计算工程, 2003, 129(16): 1 - 3.
- [2] 钱卫宁, 魏 黎, 王 焱, 钱海雷, 周傲英. 一个面向大规模数据库的数据挖掘系统[J]. 软件学报, 2002, 13(8): 1540 - 1545.
- [3] 刘同明. 数据挖掘技术及其应用[M]. 北京: 国防工业出版社, 2001: 223 - 225.
- [4] 井彦林, 穆晓云, 陈剑康, 李锁全. 黄土工程地质计算的微机化[J]. 煤矿设计, 1993, (11): 20 - 23.
- [5] 汤宇松, 刘相峰, 黄亚楼, 卢桂章. 数据挖掘系统设计[J]. 系统工程理论与实践, 2000, (9): 56 - 63.
- [6] 魏广芬, 唐祯安, 余 隽. 基于主成份分析和 BP 神经网络的气体识别方法研究[J]. 传感技术学报, 2001, (4): 292 - 297.
- [7] 何晓群. 现代统计分析方法[M]. 北京: 中国人民大学出版社, 1998: 281 - 314.
- [8] 闻 新, 周 露, 王丹力, 熊晓英. MATLAB 神经网络应用设计[M]. 北京: 科学出版社, 2000: 207 - 301.
- [9] 飞思科技研发中心. MATLAB6.5 辅助神经网络分析与设计[M]. 北京: 电子工业出版社, 2003.
- [10] 胡 雄, 陈兆能, 佟德纯, 朱继梅, 沈一飞. 基于自组织神经网络的实时统计特征抽取方法[J]. 振动工程学报, 1998, 11(1): 96 - 101.
- [11] 张松林. CART - 分类与回归树方法介绍[J]. 火山地质与矿产, 1997, 18(1): 67-75.
- [12] Schettini R, Brambilla C, Ciocca G, Valsasna A, Deponi M. A hierarchical classification strategy for digital documents[J]. Pattern Recognition, 2002, 35: 1759 - 1769.
- [13] 张立彬, 张其前, 胥 芳, 杜奖胜. 其于分类回归树(CART)方法统计解析模型的应用与研究[J]. 浙江工业大学学报, 2002, 30(4): 315 - 318.
- [14] 马立新, 王仁峰. 基于自组织特征映射神经网络的矢量量化及其在图像压缩中的应用[J]. 上海电力学院学报, 2003, 19(4): 9 - 13.
- [15] 李 戈, 邵峰晶, 朱本浩. 基于神经网络聚类研究[J]. 青岛大学学报, 2001, 16(4): 21 - 24.
- [16] 行小帅, 焦李成. 数据挖掘的聚类算法[J]. 电路与系统学报, 2003, 8(1): 59 - 67.
- [17] 关文章. 湿陷性黄土工程性能新篇[M]. 西安: 西安交通大学出版社, 1992: 67 - 135.
- [18] 刘祖典. 影响黄土湿陷系数因素的分析[J]. 工程勘察, 1994, (5): 6 - 10.