基于数据挖掘技术的黄土湿陷性评价

井彦林[1,2], 仵彦卿[1,3], 杨丽娜[1], 侯晓涛[4]
(1 西安理工大学 岩土所, 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 环境科学与工程学院, 上海 200240; 4 中国建筑西北设计研究院, 陕西 西安 710003)

[摘 要] 为了运用数据挖掘技术进行黄土湿陷性评价, 根据实际工程资料建立了黄土物理力学数据库, 用主成分分析法对原数据进行压缩, 用压缩后的新变量依据人工神经网络理论建立了预测模型, 用BP 算法进行了模型的校正及预测。工程实例分析表明, 预测湿陷系数与试验值所得湿陷系数的湿陷量计算值相比, 准确率可达96%以上, 说明这种智能化评价方法具有可行性和实用性。

[关键词] 黄土湿陷性; 数据挖掘技术; 主成分分析; BP 神经网络
[中图分类号] TU 444; TB11 [文献标识码] A [文章编号] 1671-9387(2006)04-0130-05

黄土在我国西北和华北等地分布普遍。由于大量工程勘察和科研工作的开展, 在这些地区积累了丰富的黄土物理试验和固结试验数据, 这些数据是宝贵的信息资源, 有必要对其加以充分利用。黄土的湿陷性是黄土最主要的工程性质, 其影响因素较多。如果利用已积累的数据, 依据黄土的密度、含水率、液限等指标建立湿陷性指标的预测模型, 间接预测黄土的湿陷性, 显然可减少湿陷性试验数量, 降低投资, 加快工程进度。目前针对此问题的研究成果较多, 主要可归结为两个方面: 一是基于概率统计理论的一元线性和多元线性回归及相关分析方法的湿陷性研究, 采用的参数主要为黄土固结试验和三轴固结试验数据, 指标为不扰动土的干密度、含水率、孔隙比、显微结构特征以及重塑土的结构性指标等, 得到的模型主要是回归方程[1-4]; 二是运用非线性理论如神经网络、模糊数学等研究黄土的湿陷性, 采用多元分析模型, 这种方法对湿陷性的影响因素考虑得较为全面, 同时参数也较多, 可同时考虑孔隙比、密度、含水率以及土的显微结构特征等[2,5]。目前研究中存在的问题有: 若考虑的因素少, 则模型简单, 且易于分析, 但其精度较差; 若考虑的因素多, 参数自然就多, 各参数间则有可能产生信息重复和冗余, 这样不但增加了分析问题的难度, 还会使模型变得复杂, 同时也会降低模型精度。此外, 由于已知数据的覆盖面小, 有些参数不易测得, 因此从应用的角度来说, 参数应尽量简单、易测。本文应用数据挖掘的方法进行了湿陷性研究, 通过实际工程数据的收集建立黄土物理力学数据库, 用数据挖掘的预测技术消除信息冗余, 依据数值稳定、易于测定的物理指标对湿陷系数进行预测, 以达到减少试验数量、降低工程投资的目的。

1 数据挖掘简介

数据挖掘(Data Mining, DM)是20世纪90年代中期发展起来的一门高新技术。它是从大量的不完整、有噪声、模糊、随机实际数据中提取隐含在其中的、不为人们所知但又有潜在应用价值的信息和知识的过程, 它是知识发现(Kowledge Discovery in Database, KDD)中的关键环节。数据挖掘的主要挖掘对象可以是结构化的关系型数据(仓)库, 也可以是半结构化数据或网络数据; 任务主要包括特征规则挖掘、辨识规则挖掘、关联(分为布尔关联和量化关联)规则挖掘、分类规则挖掘、数据聚类、预测、趋势性规则挖掘、公式发现、可视化及数据纠正等; 算法有归纳学习法、决策树、遗传算法、神径网络、统计分析、模糊数学、关联分析、粗糙集等; 数据挖掘的结果输出有图形、报告、逻辑公式等形式。数据挖掘的过程包括数据预处理、方法选择、建立模型、挖掘实施、规则抽取、知识解释等[6-8]。

针对本研究的目的, 本次数据挖掘的任务是对

[收稿日期] 2005-08-16
[基金项目] 国家自然科学基金项目(10572090)
[作者简介] 井彦林(1963-), 男, 陕西蒲城人, 高级工程师, 在读博士, 主要从事黄土力学与工程、环境岩土工程研究。E-mail: jingyanlin@sina.com

湿陷系数进行预测, 挖掘的对象是黄土物理力学数据库, 算法采用人工神经网络模型, 挖掘过程即是建立预测模型的过程。具体思路是首先根据实际工程的试验数据建立数据库, 用主成分分析法对原数据进行压缩、变换等预处理, 然后建立预测模型; 选用实际工程对挖掘出的预测模型进行验证, 对模型的性能作出评价。

2　数据库建立

数据库是数据挖掘的基础, 本文使用采样数据建立数据库。数据的采样方法直接关系着挖掘数据的质量, 数据挖掘中的采样方法有随机采样、分层采样和窗口采样, 本文采用分层采样, 即根据分布的不均衡控制采样的频率。当数据分布密度较大时, 采样频率低, 反之采样频率高。采样范围选定在陕西关中中部地区, 采样原则一是样本应具有广泛性和代表性, 二是样本应尽量覆盖多种地貌。要求样本全部为探井人工采取, 均来自实际工程。这些数据是工程设计中已采用过的数据, 均已经过严格分析甚至测量不确定度分析, 所以质量相当可靠。以此原则, 共收集黄土固结试验数据1 257 组, 每组试验数据包括取土深度、含水率、密度、液限、塑限、压缩系数、湿陷系数等 14 种指标, 指标来源及场地特性见表 1。

表 1　训练样本场地特征

Table 1　Site characteristic of training samples

工程名称 Engineering items	所处地貌 Geomorphology	湿陷性土层厚度/m Depth to collapsible loess	土样数量 No. of samples
景观 360 Landscape 360	渭河三级阶地 Weihe third terrace	15.0	163
白水煤矿铁路 Railway of Baishui coal mine	黄土塬 Loess highland	8.0	43
西安矿院宿舍楼 Residence of XTSU	渭河三级阶地 Weihe third terrace	12.0	18
王家岭爆破库 Explorsive bank	洪积扇 Pluvial fan	8.5	25
铁一院科研楼 Science & research Buiding of FDIR	黄土塬 Loess highland	10.0	37
蒲白马村矿浴室 Bath rooms of Macun mine	黄土塬 Loess highland	12.0	30
蒲白电厂 Power plant of Pubai	黄土塬 Loess highland	15.0	137
联合油气站 Petrol & gas station	黄土塬 Loess highland	20.0	273
黄陵总机厂 Overall workshop of Huangling	沮水河一级阶地 Jushihe of first terrace	4.0	31
黄陵一号井水池 Water pool of Huangling No. 1 mine	黄土塬 Loess highland	12.0	32
长武商贸大厦 Comercial & trade masion of Changwu	渭河二级阶地 Weihe second terrace	18.0	37
长安县三水厂 No. 3 water plant in Chang'an county	黄土斜坡 Loess highland	15.0	59
彩虹主厂房 Main building of Caihong	渭河一级阶地 Weihe first terrace	7.0	74
彩虹配料厂房 Material supply building	渭河一级阶地 Weihe first terrace	8.0	14
彬县下沟煤矿 Xiagou Mine in Binxian	泾河一级阶地 Jinghe first terrace	11.0	179
彩虹住宅三期 No. 3 stage residence of Caihong	渭河二级阶地 Weihe second terrace	10.0	105

3　主成分分析及模型建立

3.1　主成分分析

结合文献[1]、[2]对黄土湿陷性影响因素的分析, 预测变量选定为取土深度、含水率、密度、液限、塑限、压缩系数$a_{1\text{-}2}$及自重应力等7 个指标。自重应力不是实测指标, 需对数据库增加这一新的属性, 其数值由取土深度及其上覆土的密度计算得出。由于数据库中有些数据是相关的, 如含水率和压缩系数等, 而很多算法(如神经网络、回归计算等) 要求输入量之间相互独立; 同时输入量过多会使模型变得复杂, 且影响计算精度、增加运算时间, 因此必须进行数据压缩, 以消除重复和冗余数据, 并对数据降维。降维后的数据集将转化为正交集, 每个记录的属性也将减少[9]。

主成分分析(Principal Component Analysis, PCA) 是一种基于多元数理统计的数据压缩技术, 在数据挖掘、模式识别等方面应用广泛。它用降维的思想, 将相关集转化为正交集, 把多指标转化为少数几个综合指标。综合指标代表了原指标的大部分信息, 彼此不相关, 便于独立分析输入量和目标变量之间的关系。同时由于其剔掉了冗余属性, 可使模型变得简单, 且可提高模型的精度。计算步骤如下[10-12]:

设有n 个观测量, 每个观测量有p 个属性, 则构成 p 维随机向量$x = (x_1, x_2, \cdots\cdots, x_p)^T$, 令$y = u^T x$, 其中$u$ 为正交阵, y 的各分量互不相关。

(1) 将原始变量标准化

因各指标量纲往往不同, 分析时不同的量纲和数量级会引出新的问题, 各变量的不同量纲不能直

接计算,需进行标准化,使标准化的变量平均值为0,方差为1。

$$x_{ij}^{*} = \frac{x_{ij} - \overline{x_j}}{\sqrt{\mathrm{var}(x_j)}},\ i = 1, 2, \cdots\cdots, n;\ j = 1, 2, \cdots\cdots, p \tag{1}$$

式中,$\overline{x_j}$ 及$\sqrt{\mathrm{var}(x_j)}$分别为第 j 个属性的平均值及标准差;x_{ij}^{*} 为x_{ij}的标准化值。

(2) 计算x^{*}的相关矩阵R

标准化后的变量构成$n \times p$ 矩阵。利用(2)式计算得到$p \times p$ 矩阵,即相关矩阵R:

$$r_{ij} = \frac{1}{n}\sum_{t=1}^{n} x_{ti}^{*} x_{tj}^{*},\quad i, j = 1, 2, \cdots\cdots, p \tag{2}$$

式中,r_{ij}为相关矩阵R 的元素;x_{ti}^{*} 及x_{tj}^{*} 分别为标准化变量中第 t 行第 i 列与第 j 列的元素。

(3) 求相关矩阵R 的特征值及特征向量

计算出各特征值$\lambda_1 > \lambda_2 > \cdots > \lambda_p$ 和各特征向量$u_1, u_2, \cdots, u_p$。

(4) 选择主成分

计算各主成分的贡献率及累计贡献率时,主成分的个数视具体问题而定,一般取累计贡献率为80%～85%时所对应的m 个变量代替原变量($m < p$)。

(5) 根据主成分表达式计算各主成分得分,得到新变量。

$$y = u^{T}x \tag{3}$$

按上述步骤,对数据库中的取土深度、含水率、密度、液限、塑限、压缩系数$a_{1\text{-}2}$及自重应力7种指标进行分析。标准化后,按(2)式计算相关矩阵。前已述及,预测变量即输入量为7个,所以相关矩阵为7×7矩阵;然后计算相关矩阵的7个特征值$\lambda_1, \lambda_2, \cdots, \lambda_7$以及对应的特征向量$u_1, u_2, \cdots, u_7$,再按(3)式计算相应的7个主成分即新的综合变量。每一个特征值与特征值总和之比为相应主成分的贡献率。各特征值、贡献率及其累计贡献率见表2。为便于分析,表2还列出了各主成分与湿陷系数(δ)的相关系数。

表2 主成分分析结果

Table 2 Results of principal component analysis

主成分 Components	特征值 Eigenvalues	贡献率/% % of variance	累计贡献率/% Cumulative	与δ的相关系数 Correlative coefficient with δ
第一主成分 Component 1	2.561	36.59	36.59	0.437
第二主成分 Component 2	1.926	27.52	64.11	0.152
第三主成分 Component 3	1.211	17.30	81.41	-0.245
第四主成分 Component 4	0.844	12.06	93.47	-0.066
第五主成分 Component 5	0.294	4.20	97.67	0.179
第六主成分 Component 6	0.159	2.27	99.94	-0.011
第七主成分 Component 7	0.004	0.06	100.00	0.170

主成分的贡献率越大,说明该主成分所占的信息量越大、越重要。一般情况下,少数几个主成分已占总信息量的绝大部分,所以可选择占总信息量绝大部分的前若干个主成分作为新变量建模,而舍弃后面的主成分。通常选择总信息量为80%～85%时所对应的前若干个主成分,也有学者赞成选择总信息量90%甚至95%时所对应的主成分,但普遍的观点是视具体问题而定。本文所研究的问题对预测精度要求较高,如湿陷系数为0.014和0.015则具有完全不同的工程意义,因此应尽量减少信息损失。但也应舍弃对贡献率很小且与湿陷系数相关性较弱的主成分。分析结果表明,前3个主成分的累计贡献率达81%,即前3个主成分代表原数据集的绝大部分信息。主成分分析在数据压缩过程中是抛开预测变量(即湿陷系数)进行的,因此,对所得到的主成分应与预测变量一起进行相关分析,结合相关分析结果对其主成分进行选择。由表2可以看出,第五、七主成分的贡献率小,但二者与δ的相关系数分别为0.179和0.170,相关性强于信息含量较大的第二主成分,故第五、七主成分均应作为新变量予以保留。第四、六主成分所占信息量小,与湿陷系数的相关性也弱,可舍弃,则所选主成分的信息量占总信息量的85.67%。这样输入变量就由原来的7个变为5个,且这5个新变量相互独立,不具相关性。

3.2 建立预测模型

神经网络最初起源于机器学习,其主要用途是预测,包括连续型数据和离散型数据的预测。随着对神经网络研究的不断深入,其应用领域也在不断地拓展,现已成为重要的数据挖掘方法。神经网络在数据挖掘中的用途非常广泛,尤其适合处理描述性和预测性数据的挖掘,还可用于特征规则提取和关联规则提取。

本文研究的问题是挖掘湿陷系数的预测模型, 而湿陷系数是一个连续型数值, 其预测实际上是一个回归问题。建模时传递函数选用Tansig函数, 采用的3层网络分别为输入层、隐含层及输出层。以Tansig为传递函数的BP神经网络的主要计算公式如下:

$$I = \sum_{i=1}^{p} w_{1i} y_i + b_1 \tag{4}$$

$$O = \frac{e^I - e^{-I}}{e^I + e^{-I}} \tag{5}$$

$$S = \sum_{i=1}^{q} w_{2i} y_i + b_2 \tag{6}$$

$$\delta = \frac{e^S - e^{-S}}{e^S + e^{-S}} \tag{7}$$

$$err = \frac{1}{2} \sum_{i=1}^{n} (t_i - \delta_i)^2 \tag{8}$$

式中, I 及 O 分别为隐含层的输入值及输出值; S 为输出层的输入值; δ 为网络输出结果, 即湿陷系数的预测值; err 为能量函数即网络允许误差; y 为经PCA后的新变量即输入量; p 及 q 分别为输入层及隐含层节点数; n 为样本数量; w_1, w_2 分别为隐含层与输入层以及输出层与隐含层神经元的连接权; b_1, b_2 分别为隐含层和输出层神经元的阈值; t 为湿陷系数的测试值。(5)式及(7)式为Tansig型传递函数。

本文中输入层为5个节点, 隐含层为20个节点, 输出层为1个节点, 即湿陷系数; 网络误差为0.01, 学习率为0.02, 训练样本数为1 257; BP网络的训练过程是不断调整权值和阈值的过程, 当网络输出误差小于允许误差时终止训练。网络经训练后得到预测模型。

4　模型检验

本文以陕西澄城董东煤矿工业场地的湿陷性评价结果对所建预测模型进行检验。该场地位于陕西省澄城县城东北4.3 km处, 距离训练样本所在场地最近者为蒲白矿务局矸石电厂, 两者相距70余km。董东煤矿地貌为渭北黄土塬, 其场地地层表层为耕土, 其下为第四纪全新世残积黑垆土(Q_4^{el})、早更新世风积黄土(Q_3^{eol})、残积古土壤(Q_3^{el})及中更新世风积黄土(Q_2^{eol}), 共6层黄土, 2层古土壤。常规压力下, 湿陷性土层厚近20 m。自重湿陷量的计算值为386～940 mm, 场地为自重湿陷类型。

用所建立的预测模型对该场地8个探井土样的湿陷系数进行预测, 然后用预测的湿陷系数进行黄土的湿陷量计算, 计算结果见表3。与黄土固结试验湿陷系数所得湿陷量相比较, 未经主成分分析(即用含水率、密度等指标直接作为网络输入)所得试验值的湿陷量计算值为842～1 082 mm, 预测湿陷系数所得湿陷量计算值为788～961 mm, 预测相对误差为-20.9%～-2.6%, 预测准确率平均为89.0%, 预测计算的湿陷等级与试验值一致, 均为Ⅳ级。经采用主成分分析法进行数据压缩后, 预测相对误差为-5.2%～-0.5%, 预测准确率平均达到96.4%, 预测计算的湿陷等级与试验值同样完全一致, 均为Ⅳ级, 但预测相对误差明显减小, 精度提高。同时输入层网络权值由20×7矩阵缩小为20×5阵, 模型参数减少, 模型的复杂度减小。

表3　预测值与试验值结果比较

Table 3　Comparison of results on test and prediction

建筑物名称 Items	探井编号 Drill number	湿陷土层起始深度 Depth to collapsible loess	试验值 Test values		未经PCA的预测值 Initial data			经PCA后的预测值 Predicted values by using PCA		
			湿陷层/mm Collapse account	湿陷等级 Collapsibility grade	湿陷层/mm Collapse account	相对误差/% Relative errors	湿陷等级 Collapsibility grade	湿陷层/mm Collapse account	相对误差/% Relative errors	湿陷等级 Collapsibility grade
变电所 Substation	1	2.49～16.50	920	Ⅳ	788	-14.2	Ⅳ	915	-0.5	Ⅳ
	4	2.18～16.50	842	Ⅳ	830	-2.6	Ⅳ	817	-2.9	Ⅳ
日用消防水池 Priefighting water pool	5	1.84～16.50	960	Ⅳ	856	-10.8	Ⅳ	910	-5.2	Ⅳ
单身宿舍 Single's dormitory	10	2.09～16.50	984	Ⅳ	864	-12.2	Ⅳ	941	-4.4	Ⅳ
	12	2.02～16.50	901	Ⅳ	827	-8.2	Ⅳ	861	-4.4	Ⅳ
	16	1.50～16.50	1 082	Ⅳ	961	-11.2	Ⅳ	1 044	-3.7	Ⅳ
休息室 Sitting room	21	2.70～16.50	1 008	Ⅳ	797	-20.9	Ⅳ	978	-3.0	Ⅳ
消防材料库 Priefighting materials store	34	2.06～16.50	930	Ⅳ	853	-8.3	Ⅳ	884	-4.9	Ⅳ

5 结 论

1) 本研究运用数据挖掘技术对黄土的湿陷性评价进行了研究, 建立了黄土湿陷系数的预测模型。验证结果表明, 模型具有较高的预测精度, 在工程上具有实用性。

2) 利用预测模型, 可预测湿陷系数以评价黄土的湿陷等级, 有减少试验数量、降低成本的作用。评价结果适用于工程建设的规划和初步设计等阶段。

3) 经采用主成分分析法进行数据压缩后, BP网络输入变量及其权值减少, 使模型得到了简化, 精度有所提高, 网络训练速度加快。

4) 本文所采用的数据主要来自陕西关中盆地的黄土塬、渭河阶地等地貌, 下一步有必要增加样本数量, 针对更多的地貌单元进行研究。

[参考文献]

[1] 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.

[2] 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.

[3] 胡再强, 沈珠江, 谢定义. 非饱和黄土的显微结构与湿陷性[J]. 水利水运科学研究, 2000(2): 68-71.

[4] 齐吉琳. 土的结构性及其定量化参数的研究[D]. 西安: 西安理工大学, 1999.

[5] 李晓军. 黄土的特性分析及黄土湿陷性的预测[D]. 西安: 西安科技大学, 1996.

[6] 刘同明. 数据挖掘技术及其应用[M]. 北京: 国防工业出版社, 2001.

[7] Sun Jun Lee, Keng Siau. A review of data mining techniques[J]. Industrial Management & Data Systems, 2001(1): 41-46.

[8] 陈富赞, 寇继淞, 王以直. 数据挖掘方法的研究[J]. 系统工程与电子技术, 2000, 22(8): 78-81.

[9] 魏广芬, 唐祯安, 余 隽. 基于主成分分析和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.

[10] 赵希男. 主成分分析法评价功能浅析[J]. 系统工程, 1995, 13(2): 24-27.

[11] 徐 伟, 王 波. 主成分分析方法在多因素经济分析评价中的应用[J]. 北京邮电大学学报, 2000, 2(2): 34-39.

[12] 夏清东, 李乃元. 主成分分析法评价施工效果的研究[J]. 青岛建筑工程学院学报, 2000, 21(3): 10-14.

Assessment of loess collapsibility based on data mining

JING Yan-lin[1,2], WU Yan-qing[1,3], YANG Li-na[1], HOU Xiao-tao[4]

(1 *Institute of Geotechnical, Xi'an University of Technology, Xi'an, Shaanxi* 710048, *China*;
2 *Xi'an Design & Research Institute of Coal Industry, Xi'an, Shaanxi* 710054, *China*;
3 *School of Environment Science and Technology, Shanghai Jiao Tong University, Shanghai* 200240, *China*;
4 *School of Environment Science and Technology, China Northwest Building Design & Research Institute, Xi'an, Shaanxi* 710003, *China*)

Abstract: This paper presents a method for assessment of loess collapsibility based on the dada mining technology. Loess collapsibility is predicted by using the function of data mining. The database should be created based on practical engineering. Data in the database are compressed with principal component analysis (PCA). Prediction model is built with BP neural network. Variables processed through PCA are used as input part of prediction model. Loess collapsibility is predicted by the model. The predicted loess collapse settlement is compared with measured loess collapse settlement. The result shows that prediction precision of collapse settlement is up to 96% by a specific project example, indicating that the intelligent method of evaluating loess collapsibility is very useful in engineering.

Key words: loess collapsibility; data mining; principal component analysis; BP neural network