

基于粗糙不相容系统的膨胀土分类规则提取

丁加明, 王永和, 丁力行

(中南大学 土木建筑学院, 湖南 长沙, 410075)

摘要: 分析膨胀土分类的粗糙性, 指出膨胀土分类是一个基于粗糙集的信息不相容决策系统。针对常规方法容易引起规则失真的不足, 提出将贝叶斯理论和不相容系统决策挖掘相结合来提取膨胀土分类规则: 以膨胀土分类决策系统的可信度为先验概率, 膨胀土试验数据的支持度为后验概率, 计算膨胀土分类规则的条件概率; 提取条件概率大于某一阈值的规则; 通过逻辑合取与析取归并膨胀土分类规则。实例计算和应用分析结果表明: 采用贝叶斯理论和基于粗糙集的不相容系统决策挖掘相结合的方法有利于基于粗糙集的不相容系统的数据挖掘, 而且为膨胀土分类规则的提取提供了一种切实可行的算法。

关键词: 粗糙集; 膨胀土; 不相容系统; 规则提取

中图分类号: TU443 文献标识码: A 文章编号: 1672-7207(2006)02-0396-05

Generating expansive soil classification rule based on rough sets for inconsistent decision system

DING Jia-ming, WANG Yong-he, DING Li-xing

(School of Civil and Architectural Engineering, Central South University, Changsha 410075, China)

Abstract: Expansive soil classification is an inconsistent system based on rough sets after analyzing its roughness. To avoid rule distortion resulting from routine methods, expansive soil classification rule was generated by utilizing the Bayes theory and data mining based on rough sets for inconsistent decision system. The reliability of expansive soil classification decision system was regarded as the priori probability, and the support degree of expansive soil data from experiment was regarded as posterior probability, and then the conditional probability was calculated. The rules were preserved that conditional probability was bigger than a given threshold value. The rule of expansive soil classification was generated through logic conjunction and disjunction to all of the preserved rules. The results show that the method not only perfects data mining based on rough sets for inconsistent decision system, but also provides an efficient and feasible algorithm to classify expansive soil.

Key words: rough sets; expansive soil; inconsistent system; generating rule

膨胀土是颗粒高度分散, 成分以粘粒矿物为主, 对环境湿热变化敏感的高液限黏土, 具有吸水膨胀、失水收缩及其反复变形的特点, 易给工程结构物造成危害, 如路面开裂、路堤坍塌以及建筑物

隆起或沉陷等。为了合理而有效地利用膨胀土, 首先要正确地评判膨胀土的胀缩等级^[1]。目前, 在系统研究反映膨胀土基本性质的众多指标及其组合规律的基础上, 出现了若干评判膨胀土胀缩等级的综

收稿日期: 2005-08-10

基金项目: 国家教育部博士点基金资助项目(20030533043)

作者简介: 丁加明(1973-), 男, 湖南岳阳人, 讲师, 博士研究生, 从事膨胀土研究

论文联系人: 丁加明, 男, 讲师, 博士研究生; 电话: 0731-2655512(H), 13507408054(手机); E-mail: dj_ming@126.com

合指标体系^[2]。由于膨胀土胀缩等级评判指标相互之间关系复杂,所需试验资料不全,研究者研究问题的侧重点各不相同,根据现行的膨胀土分级指标体系,对同一种土的评判极有可能得到截然不同的结论。在此,作者用粗糙集中数据挖掘方法提取膨胀土分类规则。通过计算每一规则的可信度和支持度,用贝叶斯方法计算条件概率。条件概率大于某一阈值的规则成立并保留,通过逻辑合取与析取的计算后,得到膨胀土的分类规则。

1 基于粗糙集的规则提取

1.1 膨胀土的分类指标和方法

可作为判别膨胀土的指标很多,现有判别指标大体可归纳为两大类^[3]: 其一是土的物质组成(黏土矿物组成、粒度组成等); 其二是土粒与水相互作用所呈现的水理性质指标(塑性指数、液限、自由膨胀率、膨胀力等)。目前,判定膨胀土的胀缩等级基本上是采用实测指标与条例规定的界限值进行对比的经验性方法。由于影响膨胀土胀缩的指标较多,按单一指标对膨胀土进行分类时,由于不能全面反映土性有可能出现误判。而采用多指标时,又经常会遇到各项指标的实测值不是落在同一级别的界限中,从而会出现相互矛盾的结论。

在系统研究反映膨胀土基本性质的众多指标及其组合规律的基础上,有若干评判膨胀土胀缩等级的方法^[4,5],如最大胀缩性指标分类法、风干含水量分类法、塑性图分类法、灰色聚类法、灰色关联分析法、模糊数学方法以及神经网络法^[5]等。

采用传统的模糊数学方法及灰色聚类法时由于都要确定权数,有的还要确定隶属度,因此,不可避免地给建模带来了人为因素。而神经网络为了提高网络泛化能力和预测精度,在预测之前要尽量增大和扩充训练样本的组数。同时,尽管神经网络预测功能强大,但是它无法解释预测过程以及预测结果。因此,如何利用这些指标进行合理评判,目前还没有一种切实可行的方法。

1.2 膨胀土分类的粗糙性

膨胀土的胀缩具有中介过渡性或亦此亦彼性,是一个不精确的概念,不具备明确定义或标准。Z. Pawlak^[6]提出的粗糙集理论是研究不完整数据及不精确知识的表达、学习、归纳的理论。该理论认为知识与分类是密不可分的^[7],知识就是基于对对象分类的能力。分类就是将相差不大的对象归为一

类,它们的关系是不可分辨关系,也称等价关系。通过对观察和测量所得数据进行分析、近似分类、推理数据间的关系,从中发现隐含的知识,揭示潜在的规律,是知识发现和获取的一种重要研究方法^[8]。

给定膨胀土试验数据库 $K = (U, R)$, 对于每个试验样本 $X \in U$ 和 1 个试验指标 $R \in \text{ind}(K)$, 其中 $U \neq \emptyset$ 是所有试样组成的非空有限集合, 称为论域。 R 是 U 上的一族等价关系, U/R 为 R 的所有等价类族, 粗糙集通过定义上近似和下近似 2 个精确集来表示不精确概念。上近似为 $R^+(X) = \{X \in U \mid [x]_R \cap X \neq \emptyset\}$, 下近似为 $R^-(X) = \{x \in U \mid [x]_R \subseteq X\}$ 。设 $P \subseteq R, Q \subseteq R$, 若 $\text{ind}(P) \subseteq \text{ind}(Q)$, 则称试验指标 Q 可由试验指标 P 完全推导出来。显然, 在膨胀土分类的过程中, 根据某些特定的试验指标并不能准确地对膨胀土的胀缩等级进行分类。因此, 不是所有的试验指标都可以完全被推导。为此, 引入 Q 的 P 正域, 记为 $\text{pos}_P(Q) = \bigcup_{x \in U/Q} R^-(X)$, 表示论域 U 中所有根据分类 U/P 的信息可以准确地划分到关系 Q 的等价类中的对象集合。

当试验指标 $r \in P$, 且 $\text{pos}_P(Q) = \text{pos}_{(P-r)}(Q)$ 时, 称 r 为试验指标集 P 中 Q 不必要的, 否则 r 为 P 中 Q 必要的。若 P 中每个 r 都为 Q 必要的, 则称 P 为 Q 独立的。设 $S \subseteq P$, S 是 P 的独立子族, 且 $\text{pos}_S(Q) = \text{pos}_P(Q)$, 则称 S 为 P 的 Q 约简。 P 中所有 Q 必要的原始关系构成的集合称为 P 的 Q 核。

1.3 相容系统规则提取

决策规则的提取是粗糙集理论的具体应用。在大量的原始数据信息中, 隐藏着许多包含知识规律等有用信息, 从大量的原始数据信息中发现有用的规律信息叫做知识获取。基于粗糙集理论的知识获取主要是借助信息表这样一种有效的知识表达方式, 对数据进行分类、约简、知识获取直至形成规则。

设所有膨胀土试验数据 $K = (U, A, V, f)$ 是一个决策表。其中试样 $U \neq \emptyset$ 为论域; A 为试验指标及分类的非空有限集合, $A = C \cup D, C \cap D = \emptyset$, C 为试验指标集, D 为膨胀土分类集; $V = \bigcup_{a \in A} V_a$, V_a 为属性 a 的值域; $f: U \times A \rightarrow V$ 是一个信息函数, 它为每个对象的每个属性赋予一个信息值, 即 $\forall a \in A, x \in U, f(x, a) \in V_a$ 。令 X_i 和 Y_j 分别代表 U/C 与 U/D 中的各个等价类, $\text{des}(X_i)$ 表示对同一分类 X_i 的描述, 即同一分类 X_i 对于各试验指标的特定取值; $\text{dex}(Y_j)$ 表示对分类 Y_j 的描述,

膨胀缩等级 Y_j 对于各试验指标值的特定取值。

决策规则定义如下: $r_{ij}: des(X_i) \rightarrow des(Y_j)$, $Y_j \cap X_i = \emptyset$, 规则的可信度 $\mu(X_i, Y_j) = |Y_j \cap X_i| / |X_i|$, $0 < \mu(X_i, Y_j) \leq 1$ 。当系统中有一个 $0 < \mu(X_i, Y_j) < 1$ 时, r_{ij} 是不确定的, 所有试验数据是不相容系统。否则, 所有数据为相容系统。

粗糙集理论应用于膨胀土分类试验规则挖掘的一个不足之处是它对噪声反映十分敏感。例如按试验指标等价类 $des(X_i)$ 有 2 个分类, 即 $des(Y_j) (j = 1, 2)$, 则规则的可信度为 0.5。假设决策规则 $des(X_i) \rightarrow des(Y_1)$ 包含 99 个试样, 而 $des(X_i) \rightarrow des(Y_2)$ 只有 1 个试样, 尽管这 2 类规则的可信度均为 0.5, 但从统计的角度看, 规则 $des(X_i) \rightarrow des(Y_2)$ 中的那个“与众不同”的试样很可能是个特例, 或许是由操作不当引起的。按上述理论计算, 规则 $des(X_i) \rightarrow des(Y_1)$ 不能归入近似集中, 从而失去形成确定规则的机会。为此, 应该将包含试样数小的规则去掉, 减少噪声参数对规则提取的影响。在膨胀土试验中, 可能由于一些人为误差, 使试验结果出现偏差, 这些规则应该去掉。设 $W \subseteq U$, 对于分类 U/a , 定义的下近似为 $W^{(U/a)^-} = \cup_{V \in U/a, V \subseteq W} V$, 也用 $S_a(W)$ 来表示。子集 $S_a(W)$ 称为 W 关于属性 a 的支持子集, $spt_a(W) = |S_a(W)| / |U|$ 称为 W 关于属性 a 的支持度。

2 膨胀土分类规则的提取

2.1 膨胀土分类规则相容性分析

采用文献[5]的试验资料和胀缩等级指标标准值提取膨胀土胀缩等价的评判规则。粗糙集只能对离散数据进行处理, 当输出数据是连续数据时, 在应用粗集时必须对数据进行离散化处理。离散化本质上可归结为利用选取的断点对条件属性构成的空间进行划分。把条件属性划分成有限个区域, 使得每个区域中对象的决策值相同。

数据离散化预处理按文献[5]中的指标胀缩等级进行, 评判结果 $D = \{\text{膨胀土胀缩等级} | 0, 1, 2\}$ 。其中, “0”表示弱胀缩, “1”表示中胀缩, “2”表示强胀缩。剔除离散化以后相同的数据行或数据列, 进行规则约简和属性约简, 得到如表 1 所示决策表。样本数为膨胀土试验过程中试件试验结果大致相同、被划分到同一子集中的试验样本数。子集的样本数为 1, 总数为 10。从表 1 可以看出, 子集 1 和 8, 4 和 5, 以及 9 与 10 和 11 为不相容。

表 1 试验指标及评判结果离散化数值

Table 1 Discretization values of test indexes and classification results

U	$F_s / \%$	P_P / kPa	$ep_{50} / \%$	$ep_s / \%$	D
1	0	1	1	1	1
2	2	0	0	2	1
3	1	1	1	1	1
4	1	1	2	2	1
5	1	1	2	2	2
6	1	2	2	2	2
7	2	2	2	2	2
8	0	1	1	1	0
9	0	2	2	2	0
10	0	2	2	2	1
11	0	2	2	2	2
12	1	0	1	2	1

注: U 为论域; F_s 为自由膨胀率; P_P 为膨胀力; ep_{50} 是压力为 50 kPa 时的膨胀率; ep_s 为总胀缩率; D 为评判结果。

2.2 不相容系统规则的贝叶斯提取

对于不相容系统的规则提取, 常用的办法是将不相容的规则删除, 只计算相容的规则。显然, 这样会使原始数据和要提取的评判规则失真, 使规则提取丧失意义。还有的算法是计算每一个规则的可信度和支持度, 设定 2 个阈值^[9]。当这个规则的可信度和支持度均大于其对应的阈值时, 规则成立。这对可信度大而支持度小或者可信度小而支持度大的规则而言, 将不可避免地丧失有意义的规则。因此, 这种双阈值的规则提取也不合适。

采用贝叶斯方法计算条件概率可以解决这个问题。贝叶斯方法^[10]是概率论中的一种计算方法。把膨胀土分类规则的不确定性(可信度)看作系统本身具有的先验概率, 获得每一条规则的确定性即可信度为 $P(N_j)$, 共有 n 种。这种验前概率分布充分表达了膨胀土分类系统本身固有的信息。把支持度看作膨胀土试验过程中样本可信度为 $P(N_j)$ 时的后验概率 $P(Z_j | N_j)$, $j = 1, 2, \dots, n$, 应用贝叶斯公式和全概率公式, 得出规则 r_j 在可信度为 N_j , 支持度为 Z_j 时条件概率 $P(N_j | Z_j)$ 为:

$$P(N_j | Z_j) = \frac{P(Z_j | N_j) \cdot P(N_j)}{\sum_{j=1}^n P(Z_j | N_j) \cdot P(N_j)} \quad (1)$$

其中: $j = 1, 2, \dots, n$ 。设定某一阈值, 大于这一阈

值的规则将被提取。这既考虑了膨胀土分类决策表的信息, 又考虑了试验过程中样本反映的实际情况。并且由于提取规则的阈值在整个系统中是同一数值, 因而膨胀土分类规则的提取更加合理。

3 算法及其应用

3.1 算法步骤

膨胀土分类规则提取的算法步骤为:

a. 数据准备。如上所述, 粗糙集只能处理离散的数据, 对于连续的数据需要离散化。

b. 数据约简。膨胀土的试验指标体系很庞杂。许多反映土壤膨胀性的指标可能互相关联, 经过数据离散化后, 在同一条规则里许多信息可能是一致的, 因此, 这些不必要的属性应该被约简, 这样更有利于规则的提取。对于一个系统可能存在多个约简的情况, 应以能反映膨胀土的本质、指标的测定简单便捷和指标数据可靠、重现性好为原则对几个约简进行选择。

c. 计算约简后膨胀土每一条规则的可信度和支持度。计算时, 应计算每一条规则的每一个属性及其与其他属性组合的可信度与支持度。

d. 用贝叶斯方法计算上述属性组合的条件概率, 设定阈值进行规则提取。

e. 逻辑合取与析取计算。被提取的规则还比较凌乱, 有些规则互相包含。为此, 必须通过逻辑合取与析取计算提炼规则。

3.2 实例计算

以表 1 所示不相容系统的膨胀土分类评判为例进行规则提取。经计算, 表 1 中数据不能被约简, 因此, 可直接进行规则提取。对于每一条规则有 4 个属性, 考虑其属性的组合有 $C_4^1 + C_4^2 + C_4^3 + C_4^4 = 15$ 个, 这里仅列出规则 1 的条件概率计算分析过程。表 1 中第一条规则的条件属性等价类为 $\text{des}(F_s) = \{1, 8, 9, 10, 11\}$, $\text{des}(P_P) = \{1, 3, 4, 5, 8\}$, $\text{des}(e_{P50}) = \{1, 3, 8, 12\}$, $\text{des}(e_{Ps}) = \{1, 3, 8\}$, 决策属性的等价类 $\text{des}(D) = \{1, 2, 3, 4, 10, 12\}$, 如表 2 所示。

经计算, 满足条件概率大于 0.1 的规则被提取, 即 $P_P \rightarrow D$ 和 $e_{P50} \rightarrow D$ 。该规则的意义为: 当膨胀力大于 45 kPa, 并小于 85 kPa, 或者压力为 50 kPa 时的膨胀率大于 0.2 而小于 0.7 时, 可以推出膨胀土胀缩等级为中等, 该规则与实际情况相吻合。该规则连同表 1 中其他 11 条规则通过逻辑合取和析取,

表 2 膨胀土分类规则提取计算结果

Table 2 Generating rule calculation results of expansive soil classification

序号	规则	条件属性等价类	与决策规则等价类	可信度	支持度	条件概率
1	F_s	{1, 8, 9, 10, 11}	$\cap D = \{1, 10\}$	0.40	0.2	0.053
2	P_P	{1, 3, 4, 5, 8}	$\cap D = \{1, 3, 4\}$	0.60	0.3	0.120
3	e_{P50}	{1, 3, 8, 12}	$\cap D = \{1, 3, 12\}$	0.75	0.3	0.150
4	e_{Ps}	{1, 3, 8}	$\cap D = \{1, 3\}$	0.67	0.2	0.089
5	$F_s P_P$	{1, 8}	$\cap D = \{1\}$	0.50	0.1	0.033
6	$F_s e_{P50}$	{1, 8}	$\cap D = \{1\}$	0.50	0.1	0.033
7	$F_s e_{Ps}$	{1, 8}	$\cap D = \{1\}$	0.50	0.1	0.033
8	$P_P e_{P50}$	{1, 3, 8}	$\cap D = \{1, 3\}$	0.67	0.2	0.089
9	$P_P e_{Ps}$	{1, 3, 8}	$\cap D = \{1, 3\}$	0.67	0.2	0.089
10	$e_{P50} e_{Ps}$	{1, 3, 8}	$\cap D = \{1, 3\}$	0.67	0.2	0.089
11	$F_s P_P e_{P50}$	{1, 8}	$\cap D = \{1\}$	0.50	0.1	0.033
12	$F_s P_P e_{Ps}$	{1, 8}	$\cap D = \{1\}$	0.50	0.1	0.033
13	$F_s e_{P50} e_{Ps}$	{1, 8}	$\cap D = \{1\}$	0.50	0.1	0.033
14	$P_P e_{P50} e_{Ps}$	{1, 3, 8}	$\cap D = \{1, 3\}$	0.67	0.2	0.089
15	$F_s P_P e_{P50} e_{Ps}$	{1, 8}	$\cap D = \{1\}$	0.50	0.1	0.033

最终形成膨胀土胀缩等级的评判规则。

3.3 应用分析

现行的膨胀土胀缩等级评判指标和方法很多,对原始数据进行离散化时可能出现不同的结果。而粗糙集只适合离散数据的处理。因此,基于粗糙集理论的膨胀土分类应同时选取不同的评判标准离散试验数据,对计算结果进行对比研究,最后确定合理的数据离散化方法和原始样本集。

膨胀土试验数据样本必须具有一定的普遍性和代表性^[11]。若所选取的原始样本数据非常少且不具有代表性,则随着计算样本的增加,新添加的数据会对计算结果产生影响^[12]。这是粗糙集的优点之一,因为它能够随着数据的增加而更加准确地提取膨胀土的分类规则。

基于粗糙集理论不相容系统的膨胀土分类规则计算量较大,但这里提到的规则提取计算思路清晰,过程明确,对每一条规则的条件属性提取计算都遍历,在计算过程中不需要人为干涉和判断,易于编制计算机程序。

4 结 论

a. 针对传统膨胀土分类的方法处理不相容信息引起的知识失真等不足,提出了用贝叶斯理论和基于粗糙集的不相容系统决策挖掘相结合来提取膨胀土分类规则。

b. 以膨胀土分类决策系统的可信度为先验概率,膨胀土试验数据的支持度为后验概率,计算膨胀土分类规则的条件概率,提取条件概率大于某一阈值的规则。最后通过逻辑合取与析取归并膨胀土分类规则,既避免了对可信度和支持度设定双阈值带来的规则损失,又避免了可信度和支持度相矛盾时的规则丢失。

c. 采用将贝叶斯理论和基于粗糙集的不相容系统决策挖掘相结合的方法不仅有利于基于粗糙集的不相容系统的数据挖掘,而且为膨胀土分类规则的提取提供了一种切实可行的算法。

参考文献:

[1] 廖世文. 膨胀土与铁路工程[M]. 北京: 中国铁道出版社, 1984.
LIAO Shi-wen. Expansive soil and railway engineering [M]. Beijing: China Railway Press, 1984.

[2] 杨 涛, 纪桂霞. 路基膨胀土胀缩等级综合评判[J]. 华东公路, 2002(3): 45 - 47.
YANG Tao, JI Gui-xia. Synthesis judgment about the grade of shrink and expansion for the expansive soil in

the subgrade [J]. East China Highway, 2002(3): 45 - 47.

[3] 黄 卫, 钟 理, 钱振东. 路基膨胀土胀缩等级的模糊评判[J]. 岩土工程学报, 1999, 21(4): 408 - 413.
HUANG Wei, ZHONG Li, QIAN Zhen-dong. Fuzzy judgment about the grade of shrink and expansion for the expansive soil in the subgrade [J]. Chinese Journal of Geotechnical Engineering, 1999, 21(4): 408 - 413.

[4] 郭昱葵, 熊友山, 姚海林, 等. 模糊数学在当宜高速公路膨胀土判别和分类中的应用[J]. 岩土力学, 1999, 20(3): 61 - 65.
GUO Yu-kui, XIONG You-shan, YAO Hai-lin, et al. Application of fuzzy mathematics in classification of expansive soils for express way from Dangyang to Yichang [J]. Rock and Soil Mechanics, 1999, 20(3): 61 - 65.

[5] 傅鹤林, 范臻辉, 刘宝琛. 利用人工神经网络模型判定膨胀土等级[J]. 中国铁道科学, 2002, 23(5): 118 - 120.
FU He-lin, FAN Zhen-hui, LIU Bao-chen. Determination of expanded clay with artificial neural network model [J]. China Railway Science, 2002, 23(5): 118 - 120.

[6] Pawlak Z. Rough sets [J]. International Journal of Information and Computer Science, 1982 (11): 341 - 356.

[7] Pawlak Z. Rough sets approach to knowledge based decision support [J]. European Journal of Operation Research, 1997(99): 48 - 57.

[8] 张文修, 吴伟志, 梁吉业, 等. 粗糙集理论与方法[M]. 北京: 科技出版社, 2001.
ZHANG Wen-xiu, WU Wei-zhi, LIANG Ji-ye, et al. Theory and method of rough sets [M]. Beijing: Science and Technology Press, 2001.

[9] 马 力, 焦李成. 一种基于粗集理论的知识发现系统的研究与设计[J]. 微电子学与计算机, 2003, 33(3): 8 - 12.
MA Li, JIAO Li-cheng. A system for knowledge discovery in database on rough set theory: Research and design [J]. Microelectronics & Computer, 2003, 33(3): 8 - 12.

[10] 刘柏刚, 丁加明. 贝叶斯决策在确定复合标底报价中的应用[J]. 铁路工程造价管理, 2002(3): 5 - 7.
LIU Bai-gang, DING Jia-ming. The use of Bayes strategy in off-making of compound bid [J]. Railway Engineering Cost Management, 2002(3): 5 - 7.

[11] 石 骁, 赵庆飞. 基于粗糙集理论的模糊综合评判权值确定[J]. 西南石油学院学报, 2001, 23(3): 16 - 18.
SHI Xiao, ZHAO Qing-fei. Determination of fuzzy comprehensive estimation weighting coefficient based on rough set theory [J]. Journal of Southwest Petroleum Institute, 2001, 23(3): 16 - 18.

[12] 熊樱菲, 何文权, 王 兵. 粗糙集理论(RST)及其在古陶瓷分类上应用的初探[J]. 文物保护与考古科学, 2002, 14(增刊): 298 - 308.
XIONG Ying-fei, HE Wen-quan, WANG Bing. Introduction on rough set theory and its application to ancient ceramics classification [J]. Sciences of Conservation and Archaeology, 2002, 14(Suppl): 298 - 308.