

粗糙集属性约简在尾矿坝浸润线预测模型中的应用^{*}

王云海 李春民 谢旭阳
(中国安全生产科学研究院)

摘 要 在建立尾矿坝浸润线支持向量回归机(SVR)模型的过程中,预测精度低、计算时间长等问题较难解决,并且严重制约SVR模型的推广应用。为了解决以上问题,尝试引入粗糙集(RS)算法对训练样本的输入属性进行约简,同SVR算法共同建立浸润线预测模型。实例证明,RS-SVR模型有效降低了SVR模型在迭代时的计算难度,并使浸润线的预测精度得到了提高。由此可得,RS-SVR结合无论在理论上还是在实例应用中都具有可行性。

关键词 浸润线 粗糙集 约简 支持向量回归机

Application of Rough Set Attribute Reduction in the Prediction Model of Infiltration Route in Tailing Dam

Wang Yunhai Li Chunmin Xie Xuyang
(China Academy of Science and Technology)

Abstract During the establishment of the support vector regression(SVR) prediction model of infiltration route data in tailing dam, the prediction accuracy and the computational time are two difficult problems to control, which severely constrain the extensive application of the SVR model. To solve these two problems, the rough set(RS) algorithm is adopted to reduce the attribute for training samples, then the prediction model of infiltration route is established with the SVR algorithm. The cases proved that RS-SVR model effectively lowered the difficulty of SVR model iteration, and made the prediction accuracy improved. It can be seen that the RS-SVR is feasible not only in theory but in application.

Keywords Infiltration route, Rough set, Reduction, Support vector regression

日益发展的数据挖掘技术的在矿业领域得到广泛应用^[1-2],其已应用到尾矿库监测预警领域中。众所周知,国家安监部门对尾矿库的安全运营十分重视,目前已经建立了一套预测精度较高的尾矿库在线监测预警系统^[3]。整个系统能够较为精确地预测出监测指标,比如浸润线^[1]埋深,是其中至关重要的一步。因为有了精确的预测值,才能够对危险进行较为准确的预警。

而在监测系统中常用的SVR浸润线预测模型有很大的局限性,比如由于影响属性较多造成的数据计算量较大导致计算时间过长,浸润线某些属性数据冗余对回归模型的训练造成干扰。因此,亟待能够对监测的实时属性数据进行约简,这将对建立更有效的监测系统有着深远的意义。本研究将探讨粗糙集属性约简算法和支持向量回归机结合应用,利用粗糙集算法首先对监测到的属性数据进行约简,理论上可以大大降低计算的难度,并且能够提高预测的精度,在理论上具有可行性。

1 方法介绍

1.1 数据离散化

应用粗糙集对浸润线属性进行约简,首先要对属性数据进行离散化^[4-6],这是至关重要的一步。数据离散化分类的方法,如等频法,布尔推理法等,这些方法对于大量的连续数据进行分类,尤其对于那些孤立的、离某类聚类中心较偏的点的效果较差。为此,可采用模糊聚类方法来解决大量连续数据离散化的问题。Matlab作为一种科学计算软件,目前在数值处理与计算中应用较多。处理大量连续数据离散化问题,可采用Matlab模糊工具箱中的模糊SOFM聚类函数。

SOFM网聚类的原理为^[7-8]:SOFM网络两层之间的各神经元实现双向全连接,而且网络中没有隐含层。在SOFM中,输出结点与其邻域其它结点广

^{*} “十一五”国家科技支撑计划项目(编号:2006BAK04B04)。

王云海(1965—),男,中国安全生产科学研究院,教授级高级工程师,100029 北京市朝阳区北苑路32号甲1号楼安全大厦。

泛相连,并相互激励。每一个输入单元都与二维阵列的输出神经元相连,每一个输出单元都与其相邻的输出单元相连。因此 SOFM 网络中包含两类权值:

(1)输入神经元与输出神经元之间的连接权值。

(2)输出神经元间的侧向连接权值。

在 SOFM 网络训练过程中,竞争获胜神经元的邻近效应的邻域划分并不是一成不变的,而是随着迭代次数的增加而逐步减小。如图 1 所示。

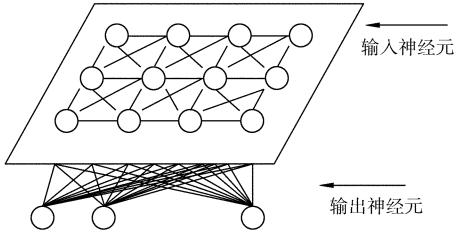


图 1 SOFM 网示意

应用该方法对连续属性进行离散处理,关键是正确选取聚类数目。聚类数目过少,可能会得出不相容的决策系统,导致实际应用时根据判断条件无法作出决策;聚类数目多,会出现过离散的情况,极端情况下,离散处理后决策系统中对象的条件部分互不相同,各自形成独立的规则。

1.2 粗糙集属性约简

粗糙集(rough sets)是波兰数学家 Pawlak 提出的一种新型的处理模糊和不确定知识的数学工具。它是对传统集合论的扩展,使其能够适应信息表示、信息推理等方面的需要。粗糙集理论认为知识都是有力度的,有些概念之所以不能得到很好的表示,是因为知识的力度超过了这些概念的边界,从而不能准确地表达概念。

粗糙集约简的概念如下。

(1)信息系统。一个信息系统 I 是一个四元组,即 $I = \langle U, \Omega, V, f \rangle$ 。其中, U 为论域,是全体对象的有限集合,设有 n 个对象,则其可以表示成 $U = \{x_1, x_2, \dots, x_n\}$; Ω 为有限个属性的集合,设有 m 个属性,则其可以表示为 $\Omega = \{A_1, A_2, \dots, A_m\}$; V 为属性的值域集合, $V = \{V_1, V_2, \dots, V_m\}$, V_i 是属性 A_i 的值域; f 为信息函数, $f: U \times \Omega \rightarrow V, f(x_i, A_j) \in V_j$ 。

设有对象子集 $V \subseteq U$, 属性子集 $P \subseteq \Omega$, 则定义不可区分关系为

$$IND(V, P) = \left\{ (x, x') \in V \times V = \right.$$

$$\left. \mid \forall a \in P, a(x) = a(x') \right\}.$$

不可区分关系是一个等假关系。

(2)约简。约简可以区分所有对象的最小属性子集。

设有两个属性集 C 和 $B, B \subset C$, 如果 $IND(B) = IND(C)$, 且对于任意的 $B' \subset B, IND(B') \neq IND(C)$ 都不成立, 则称属性集 B 为 C 的一个约简, 记做 $RED(C)$ 。

(3)核。属性 C 的所有约简的交集称为核, 记作 $CORE(C) = \cap RED(C)$ 。核表示 C 中对于所有约简都不可缺少的属性集合。核一定是唯一的。

在对数据的分析研究中,原始的条件属性并不是同等重要的,甚至其中某些条件属性是冗余的。冗余属性的存在,一方面是对资源的浪费(需要存储空间和处理时间);另一方面,也干扰作出正确而简洁的决策。

本研究采用基于可辨识矩阵的约简算法,算法介绍如下。

应用可辨识理论可以得到全部的可能约简组合,也就是说能够获得理论上的最优约简。但是这样计算的时间复杂度和空间几何度都成几何级数增长,在实际应用中,数量巨大时是难以实现的。

令 M 是决策表 T 的可辨识矩阵, $A = \{a_1, a_2, \dots, a_n\}$ 是 T 中所有条件属性的集合, S 是 M 中所有条件属性组合的集合,且 S 中不包含重复项。令 S 中总共有 s 个条件属性组合,并且每个条件属性组合表示为 B_i , 可用符号描述为 $B_i \in S, B_j \in S$ 且 $B_i \neq B_j (i, j = 1, 2, \dots, s)$ 。令 $card(C_{i0}) = m$, 则 B_j 中每个条件属性表示为 $b_{i,k} \in B_i (k = 1, 2, \dots, m)$ 。根据可辨识矩阵的定义,容易得知:矩阵中条件属性组合数为 1 的元素项(即 $card(B_i) = 1$)表明,除该属性外其余条件属性无法将决策不同的两条记录区分出来,即该属性必须保留,与决策表中核属性的概念一致。因此矩阵中所有条件属性组合数为 1 的属性均为决策表的核属性(核属性可能为空)。令 C_0 是 M 中核属性的集合, C_{i0} 为其中元素,则有 $C_0 \subseteq A, C_{i0} \in C_0$ 且 $card(C_{i0}) = 1$ 。

除核属性外,另外需要建立的概念是:可辨识矩阵中凡是条件属性组合中包含核属性的矩阵元素,都可以仅用核属性就把决策不同的记录区分出来,也就是说属性组合中凡是包含有核属性的其它条件属性都是多余的。

如果可辨识矩阵中某些元素未包含核属性,说明该决策表中存在一些无法由核属性判断决策的记录,所有这些不包含核属性的属性组合当中,每个组合必然都至少有 1 个元素应当成为约简后的 1 个条件属性,否则决策表中的某些记录将无法识别。由此可以构造一个不包含核属性的合理范式。最终选取一个合理范式(该范式每一项由合取范式表示的属性组合连同核属性一起表达)作为最终的约简后果。

2 实例分析

2.1 粗糙集属性约简

针对某尾矿库主坝建立浸润线预测模型。首先,建立指标体系^[9]:确定目标属性为浸润线;确定影响、决定浸润线埋深的其它重要且相对容易获得的指标为输入属性。因为单一尾矿本身的沉积规律基本固定不变,所以不作为样本的输入属性的考虑范围^[7]。结合某尾矿坝监测数据,确定输入属性为:坝顶高程、滩顶高程、滩前 100 m 高程、干滩长度、库水位,如表 1。

表 1 某尾矿坝主坝部分数据 m						
坝顶 高程	滩顶 高程	滩前 100 m 高程	干滩 长度	库水位	监测点 浸润 线值	监测日期
90.94	87.47	85.31	>220	84.38	9.2	2007-05-25
90.94	87.47	85.31	>220	84.48	9.15	2007-06-25
90.94	87.47	85.37	>220	84.5	9.2	2007-07-25
90.94	87.47	85.6	>220	84.57	9.25	2007-08-17
90.94	87.47	85.6	>220	84.97	9.1	2007-09-07
90.94	87.47	85.86	210	84.44	9.2	2007-10-24
90.94	87.55	86.48	220	85.17	9.2	2007-11-25
90.94	87.86	86.59	220	85.36	9.1	2007-12-19
90.74	88.06	86.93	240	84.95	9.7	2008-02-18
90.74	89.13	87.93	280	84.07	9.6	2008-03-25
90.74	89.04	87.93	280	85.13	9.5	2008-04-25
90.74	89.04	87.93	280	84.51	9.8	2008-05-19
90.74	89.28	87.93	280	86.21	9.8	2008-06-19

应用 SOFM 对属性数据进行离散化处理。本研究中的数据分别为坝顶高程、滩顶高程、滩前 100 m 高程、干滩长度、库水位;决策属性为浸润线埋深。本研究分别对指标以 3,4,5 类进行试验,结果表明,当 SOFM 网络将每个属性离散为 3 类,决策属性离散为 2 类时,运用粗糙集理论对指标属性约简取得较好效果。对监测数据应用 SOFM 网络,在 Matlab

软件中将数据按分为 3 类的情况对其进行离散化处理,结果如表 2。

表 2 某尾矿坝主坝部分数据 m					
坝顶 高程	滩顶 高程	滩前 100 m 高程	干滩 长度	库水位	监测点 浸润线值
1	1	3	3	1	1
1	1	3	3	1	1
1	1	3	3	1	1
1	1	3	3	1	1
1	1	3	3	2	1
1	1	3	3	1	1
1	1	2	3	3	1
1	2	2	2	3	1
3	2	2	1	2	2
3	3	1	1	1	2
3	3	1	1	3	2
3	3	1	1	1	2
3	3	1	1	3	2

运用 ROSETTA 软件对决策表进行约简^[8],经过计算,得出约简结果为滩顶高程和库水位。

2.2 回归模型的样本建立

首先建立指标体系,确定目标属性为浸润线,确定输入属性为滩顶高程、库水位^[9-10]。建立测试样本和训练样本如表 3 和表 4。

表 3 回归模型测试样本 m		
目标属性:浸润线	滩顶高程	库水位
9.7	88.06	84.95
9.6	89.13	84.07
9.5	89.04	85.13
9.8	89.04	84.51
9.8	89.28	86.21

表 4 回归模型训练样本		
目标属性:浸润线	滩顶高程	库水位
9.2	87.47	84.38
9.15	87.47	84.48
9.2	87.47	84.50
9.25	87.47	84.57
9.1	87.47	84.97
9.2	87.47	84.44
9.2	87.55	85.17
9.1	87.86	85.36

3 结果与讨论

将训练样本经过归一化处理,运用支持向量回归机算法进行训练,建立预测模型,并运用测试样本进行测试^[10-11]。

为了便于比较,本研究采用未经粗糙集约简的训练样本进行训练并建立模型并预测,将预测结果与粗糙集约简后的模型预测结果进行对比,如表 5。

表 5 测试结果				
监测值	SVM 预测值	误 差 /%	RS - SVM 预测值	误 差 /%
9.7	9.320 87	4.06	9.166 28	5.82
9.6	8.986 13	6.83	9.165 78	4.73
9.5	8.978 85	5.80	9.166 28	3.64
9.8	8.990 97	9.00	9.163 38	6.91
9.8	8.951 55	9.47	9.166 28	6.91

模型的均方差 (MSE) 为 $MSE (RS - SVM) = 0.277\ 508$, $MSE (SVM) = 0.433\ 314$, 平均误差 ($RS - SVM$) = 5.6%, 平均误差 (SVM) = 7.1%。可以得出,经粗糙集约简后的 SVM 模型的预测结果优于属性数据直接训练所得模型,这是由 SVM 算法的局限性造成的。SVM 算法是一种机器学习算法,其需要对数据进行反复的迭代运算,而不会考虑属性的干扰性和冗余性,这些严重制约着预测结果的精度。而粗糙集的约简算法以粗糙集核属性等的相关理论为基础,根据数据之间的内在联系进行计算,剔除冗余属性,得出对决策属性影响最大的属性,作为建立模型的输入属性,实例证明,RS - SVM 浸润线预测模型不仅减少了预测的精度,还大大减少了模型计算所需的时间。

4 结 论

将粗糙集属性约简算法加入到建立浸润线预测模型过程中,运用属性约简可以降低数据计算的复杂度和数据冗余度,理论上可提高数据的预测精度。

(上接第 19 页)

困难和顶板安全性差等问题,在该矿开采中采用尽量不留或少留矿石矿柱的分条推进的回采方法,回采幅度控制在 1.25 m,利用间断或连续条柱支撑顶板岩层,最大限度地提高了矿石回收率,减少了贫化,保证了回采工作的安全。回采实践表明,连续分条回采、工作面临时支护、废石充填采矿方案,解决了薄矿脉开采采切工程量大、生产能力小的弊端,使采矿贫化率降至 10% ~ 11%,损失率控制在 15% 以内,采场生产能力达到 100 t/d,该回采方法既降低了贫化损失率,又可以少向井外排废石,节约运输费用和废石堆放用地,有利于保护地面环境、维持地表生态平衡,采矿成本仅为 6.59 美元/t,经济效益明显。因此,该采矿方法具有广阔的市场前景和良好

实例证明,RS - SVM 模型的预测精度相对于 SVM 模型的预测精度有所提高,但精度提高并不是很大,但对于尾矿库在线监测系统而言,数据精度有所提高对模型会产生重大的影响,因而证明 RS - SVM 模型是值得进一步研究并推广的。

参 考 文 献

[1] 聂兴信,刘书香. 数字化矿山构建过程中的数据挖掘模型研究[J]. 金属矿山,2007(6):19-21.

[2] 马立强,温国锋. 数据挖掘技术在矿山勘探开采中的应用研究[J]. 金属矿山,2007(6):90-92.

[3] 李全明,王云海,张兴凯,等. 尾矿库溃坝灾害因素分析及风险指标体系研究[J]. 中国安全生产科学技术,2008(3):50-53.

[4] 张瑞玲,张银丽. 基于变精度粗糙集的地下硐室稳定性研究[J]. 矿山机械,2007,35(12):38-40.

[5] 云庆夏,李 刚. 粗糙集理论在矿井调度知识挖掘中的应用[J]. 金属矿山,2004(4):1-4.

[6] 付 华. 煤矿瓦斯灾害特征提取与信息融合技术研究[D]. 阜新:辽宁工程技术大学,2006.

[7] 章 兢,张小刚. 数据挖掘算法及其工程应用[M]. 北京:机械工业出版社,2006.

[8] 徐 裘,刘玉波,范学鑫. 基于模糊工具箱和 ROSETTA 的粗糙集数据挖掘[J]. 微计算机信息,2007,23(3):174-178.

[9] 李 娟,李翠平,李仲学,等. 支持向量回归机在尾矿坝浸润线中的应用[J]. 中国安全生产科学技术,2009(1):76.

[10] 王云海,李 娟,李春民. 尾矿坝浸润线数据挖掘预测模型的样本选取研究[J]. 中国安全生产科学技术,2009,5(5):9-12.

[11] 谢旭阳,江田汉,王云海,等. 基于支持向量机的尾矿库灾害区域预警[J]. 中国安全生产科学技术,2008(4):17-21.

(收稿日期 2010-08-12)

的社会效益,可供国内同类型矿山借鉴。

参 考 文 献

[1] Blackfriars Court Investment Holdings Limited. Mining Geotechnical Feasibility Study of Buffelsfontein Chrome Mine Report No. 303129/Geotech[R]. [S. 1]:Blackfriars Court Investment Holdings Limited,2003.

[2] 《采矿设计手册》编写组. 采矿设计手册[M]. 北京:中国建筑工业出版社,1986.

[3] 兰州有色冶金设计研究院. 酒钢集团公司南非铬铁工程可行性研究报告[R]. 兰州:兰州有色冶金设计研究院,2003.

[4] Blackfriars Cout Investment Holdings Limited. Buffelsfontein Chrome Mine Mining Feasibility Study report 303129/Mining[R]. [S. 1.]: Blackfriars Court Investment Holdings Limited, 2003.

(收稿日期 2010-08-16)