文章编号:0253 - 9993(2011)04 - 0619 - 04

基于决策树 ID3 改进算法的煤与瓦斯突出预测

李定启[1,2],程远平[1],王海峰[1],王　亮[1],周红星[1],孙建华[2]

(1. 中国矿业大学 煤矿瓦斯治理国家工程研究中心,江苏 徐州　221008;2. 黑龙江科技学院 安全工程学院,黑龙江 哈尔滨　150027)

摘　要: 为提高工作面突出预测指标预测的准确率,根据灰色相关理论和决策树 ID3 算法,提出了基于决策树 ID3 改进算法的煤层工作面煤与瓦斯突出预测方法。该方法以工作面的钻屑解吸指标作为主要决策属性,以地质构造、瓦斯浓度变化等现场较为直观的突出征兆作为辅助决策属性,同时根据矿井实际工作面煤与瓦斯突出数据建立预测样本数据集,把决策属性的相对灰色关联度作为决策树 ID3 改进算法的最大信息增益计算权重,建立了煤层工作面煤与瓦斯突出决策树预测模型,并采用该预测模型对 10 组煤与瓦斯突出数据进行了预测,结果表明,该模型预测的准确率显著高于采用单一钻屑指标预测的准确率。

关键词: 决策树; ID3 改进算法; 煤与瓦斯突出; 预测方法

中图分类号: TD713.2　　**文献标志码:** A

Coal and gas outburst prediction based on improved decision tree ID3 algorithm

LI Ding-qi[1,2], CHENG Yuan-ping[1], WANG Hai-feng[1], WANG Liang[1], ZHOU Hong-xing[1], SUN Jian-hua[2]

(1. *National Engineering Research Center for Coal Gas Control, China University of Mining & Technology, Xuzhou*　221008, *China*; 2. *Department of Safety Engineering and Technology, Heilongjiang Institute of Science & Technology, Harbin*　150027, *China*)

Abstract: The prediction method based on improved decision tree ID3 algorithm was proposed by gray theory to improve the accuracy of coal and gas outburst indexes prediction. Desorption index of drill cuttings was adopt as the major decision attribute, geological structure, gas concentration changes and other obvious omen of coal and gas outburst in face were adopt as the auxiliary decision attributes, and prediction data set was built. According to actual data of mine coal and gas outburst, using relative grey relation of decision attributes as the weight of maximum information gain calculating, established decision tree model of coal and gas outburst prediction. At last, this model was applied to predict 10 sets of coal and gas outburst data, and the results show that the predicting accuracy is significantly higher than predicting by a single desorption of drill cuttings.

Key words: decision tree; improved ID3 algorithm; coal and gas outburst prediction; prediction methods

为减少和预防煤与瓦斯突出事故,各国的学者对煤与瓦斯突出预测预报做了大量研究,其中主要包括各种突出敏感指标和突出预测模型的研究。常用的突出预测指标包括煤的破坏类型、煤的坚固性系数、煤样的瓦斯放散初速度、瓦斯含量、瓦斯压力、综合指标、钻屑解吸指标、钻孔瓦斯涌出初速度等,此外还有近来发展的建立在瓦斯膨胀能、微震预测、地电场预测、电磁辐射预测、声发射预测等技术上的预测指标[1-4]。突出预测模型研究领域主要集中在利用灰色理论、模糊数学、线性模型、神经网络、遗传算法、混沌时间系列等数学方法及计算机工具对突出指标进行分析和预测[5-8]。

由于煤与瓦斯突出机理较为复杂,突出预测敏感指标及其临界值的选择难度较大,导致工作面突出预测指标的预测准确率偏低。为了进一步提高煤层工作面突出预测的准确率,笔者试图采用灰色相关理论

收稿日期:2010-09-29　　责任编辑:毕永华
基金项目:国家重点基础研究发展计划(973)资助项目(2011CB201204);国家青年科学基金资助项目(50904068)
作者简介:李定启(1979—),男,江西余干人,讲师。E-mail:lidingqi1979@163.com

对决策树 ID3 算法进行改进,提出基于决策树 ID3 改进算法的煤层工作面煤与瓦斯突出预测方法。

决策树算法是一种常用的数据挖掘算法,决策树学习采用自顶向下的递归方式构造决策树,以实例为依据,从一组无序、无规则的实例数据中推理出用于决策树形成的分类规则,并可以根据分类的结果进行预测[9-12]。这类方法相对于神经网络技术和遗传算法等分类、预测方法来说更为简单有效,比较适合现场的工程应用。虽然基于决策树算法的信息处理方法在经济、金融、管理等众多领域得到了较为充分的研究和应用[13-14],但是在矿井安全控制和管理领域少有研究。

1 基于决策树 ID3 改进算法的工作面突出预测方法

钻屑解吸指标是现场工作面突出预测常用指标之一,现场单独采用其进行工作面突出预测的准确率往往偏低。为提高预测准确率,本研究采用钻屑解吸指标 Δh_2 作为工作面突出预测决策树的主要决策属性,同时选取较为简单直观的地质构造、瓦斯浓度变化、响煤炮、片帮掉渣、喷孔顶钻夹钻等因素作为辅助决策属性建立煤与瓦斯突出决策树预测模型。

1.1 预测样本数据集的建立

根据突出矿井工作面的决策属性数据建立训练样本集,其中定量数据以突出前的测定值为标准,定性数据依据数量化理论转化为二态变量,即用“0”和“1”来表示某个定性属性的“不存在”和“存在”。突出强度采用“0”和“1”表示,分别为“突出”、“不突出”。本研究收集了 100 组具有代表性的突出矿井工作面突出预测的决策树属性数据作为样本集,其中前 90 个作为训练样本,后 10 个作为验证样本,部分数据见表 1。

表 1 煤与瓦斯突出预测部分样本数据

Table 1 Part of coal and gas outburst data

钻屑解吸指标	地质构造	瓦斯变化异常	响煤炮	片帮掉渣	喷孔顶钻	是否突出
180	0	1	0	1	0	0
160	1	0	1	0	1	0
210	1	0	1	0	1	1
140	0	1	0	1	0	0
228	1	1	1	1	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
230	0	0	1	0	1	1

1.2 ID3 改进算法决策属性的选择

构造决策树的关键是要选择一个好的划分标准,以决定按哪一个属性进行节点的划分。一般用统计度量的方法来选择属性对结点进行划分,ID3 算法采用信息增益的方法进行划分。要构造尽可能小的决策树,关键在于选择合适的产生分支的属性,ID3 算法的核心是通过采用信息增益的方式来选择能够最好地将样本分类的属性[12]。

设数据集 S 有 $A_1, A_2, \cdots, A_n$,共 n 个属性。以属性 A 为根的信息增益为

$$\mathrm{Gain}(A) = I(S_1, S_2, \cdots, S_n) - E(A) = \sum_{j=1}^{v} w_j \left(\sum_{i=1}^{m} P_{ij} \log_2 P_{ij} \right) - \sum_{i=1}^{n} P_i \log_2 P_i$$

式中,$I(S_1, S_2, \cdots, S_n)$ 为样本集 S 划分之前的总熵;$E(A)$ 为使用属性 A 把数据集 S 划分 v 个子集后的总熵;$w_j = \dfrac{s_{1j} + s_{2j} + \cdots + s_{mj}}{s}$;$P_i = \dfrac{s_i}{s}$,$P_{ij} = \dfrac{s_{ij}}{s_j}$;$s_j$ 为子集 S_j 全部样本数;s_{ij} 为子集 S_j 第 i 个属性的样本数,s 为样本集全部样本数。

以上 ID3 算法虽然有效,但该算法通常偏向选择取值较多的属性,而实际中取值较多的属性往往并不是最优的,即按照信息增益最大的原则,被 ID3 算法列为应选取的属性有时对其进行测试不会提供太多的信息[15]。那么如果要改进 ID3 算法首要考虑的就是优化对属性的选择标准,可以通过对信息熵的公式加权来加强重要属性的标注,降低非重要属性的标注,本文采用相对灰色关联度来标定属性的权值。以煤与瓦斯突出危险性为根属性,钻屑解吸指标、地质构造、瓦斯浓度变化异常等属性为子属性进行灰色关联度计算,并将相对灰色关联度定义为子属性关联度与子属性平均关联度的比值:$R_{i0} = r_{i0}/r_a$,其中 r_{i0} 为各子属性与根属性之间的灰色关联度;r_a 为各子属性平均关联度。利用决策树属性相对灰色关联度对属性信息熵的计算公式加权以加强重要属性的标注,将公式改进为

$$E(A) = -\sum_{j=1}^{v} w_j R_{i0} I(S_{1j}, S_{2j}, \cdots, S_{mj})$$

$$\mathrm{Gain}(A) = I(S_1, S_2, \cdots, S_n) - E(A) = \sum_{j=1}^{v} R_{i0} w_j \left(\sum_{i=1}^{m} P_{ij} \log_2 P_{ij} \right) - \sum_{i=1}^{n} P_i \log_2 P_i$$

式中,R_{i0} 为各子属性与根属性之间的相对灰色关联度。

采用上述相对灰色关联度计算方法和 ID3 改进算法决策属性选择对训练样本集的 6 个属性进行分支计算,计算结果见表 2。

1.3 决策属性的划分

在用决策树方法进行分类时,数值型属性要离散化。设排序后 A 的属性序列为 $v_1, v_2, \cdots, v_m$ 从小到大依次取不同的分裂点,取信息增益最大的就是 A 的

最佳划分[12]。若 v_i 为最佳分裂点,则取 $v = (v_i + v_{i+1})/2$。

表 2　样本属性分支计算结果

Table 2　The branching results of sample attribute

属性	地质构造	钻屑解吸指标	瓦斯变化异常	响煤炮	片帮掉渣	喷孔顶钻
R_{i0}	1.12	1.29	0.82	0.76	0.68	0.71
Gain(A)	0.24	0.38	0.13	0.19	0.11	0.07
分支序号	2	1	4	3	5	6

根据上述方法采用最大信息增益对钻屑解吸指标值进行属性分割,考虑到《煤与瓦斯突出防治规定》规定的钻屑解吸指标 Δh_2 临界值的参考值为 200,因而只对大于 200 的钻屑解吸指标值进行分割。将钻屑解吸指标值代入信息增益公式计算得信息增益最大的分割点为 v_{41},则

$$\mathrm{Gain}(v_i) = \mathrm{Gain}(v_2)$$

$$v = \frac{v_i + v_{i+1}}{2} = \frac{v_{41} + v_{42}}{2} = 221$$

取 221 和 200 为钻屑解吸指标数值型属性分割值。利用两个分割值对钻屑解吸指标进行划分,"0"表示小于分割值 200,"1"表示介于分割值 200 与 221 之间,"2"表示大于分割值 221。分割、离散化后的样本数据集部分数据见表 3。

表 3　离散化后的样本数据集

Table 3　The sample data set after discrete

钻屑解吸指标	地质构造	瓦斯变化异常	响煤炮	片帮掉渣	喷孔顶钻	是否突出
0	0	1	0	1	0	0
0	1	0	1	0	1	0
1	1	0	1	0	1	1
0	0	1	0	1	0	0
2	1	1	1	1	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	0	0	1	0	1	1

1.4　决策树模型的构建

在煤与瓦斯突出的决策树预测模型的生成过程中,输入为煤与瓦斯突出样本数集,输出结果为决策树预测模型。决策树的每一个决策结点对应进行分类的一个决策属性,如钻屑指标、地质构造等决策属性分别对应决策树的一个决策结点。分枝对应着按该属性进一步划分的取值特征。不同叶子代表最终的分类,如"突出"、"不突出"分别对应决策树的最终分类结果。决策树构建的基本算法是贪心算法,是以自顶向下递归的各个击破的方式构建决策树[12]:① 树以代表训练样本的单个结点开始;② 如果样本都在同一个类,则该结点成为树叶,并用该类标记;③ 否则,算法选择最有分类能力的属性作为决策树的当前结点;④ 根据当前决策结点属性取值的不同,将训练样本数据集划分为若干子集,每个取值形成一个分枝,有几个取值形成几个分枝;⑤ 针对上一步得到的一个子集,重复进行先前步骤,递归形成每个划分样本上的决策树。一旦一个属性出现在一个结点上,就不必在该结点的任何后代中考虑它。

在决策树学习过程中,如果决策树过于复杂,那么运算和存储花销将大幅度增加,同时节点过多也导致支持节点的实例变少,每个叶上所支持的实例也变少,使得规则过多过细,导致用户难以理解,降低了分类器的可用性,所以需要对决策树进行相应的简化。本文采用预剪枝技术和后剪枝技术控制树的规模来化简煤与瓦斯突出预测决策树模型[13]。

1.5　决策树生成器及预测模型

根据改进决策树分类算法、决策属性划分方法、决策属性样本数据库及预剪枝、后剪枝技术,采用 Visual C++ 语言开发出煤与瓦斯突出预测决策树模型的生成软件,然后将 1.3 节中的样本数据集输入软件界面,软件运行后的输出模型即为相应的煤与瓦斯突出预测决策树模型。

2　模型预测及分析

采用决策树模型生成器生成的煤与瓦斯突出预测模型对属性数据库中的 10 组工作面突出数据进行预测,预测结果与采用单一钻屑指标预测的结果以及实际发生突出情况见表 4。

从以上预测结果分析可知,模型预测的结果与实际基本一致,总体预测准确率 90%,不发生突出的预测准确率 100%,发生突出的预测准确率 80%;钻屑指标预测误差相对较大,总体预测准确率 60%,不发生突出的预测准确率 75%,发生突出的预测准确率 50%。

上述煤与瓦斯突出预测决策树模型的预测存在一定的误差,可能有以下 3 个方面的原因:① 样本数据集的数据量较少,没有包含足够的信息量;② 工作面煤与瓦斯突出影响因素多达数十个,包括开采技术条件、人的影响因素、煤层自然条件,而以上建立的决策树预测模型仅选取了煤与瓦斯突出的煤层自然条件的部分因素;③ 决策树预测模型剪枝化简可能会造成一定信息丢失。因此,为了减少模型预测的误差,需要不断完善和更新预测样本数据集,同时根据预测的结果对决策树模型进行分析和优化。

表 4 模型预测结果

Table 4 The results of model prediction

数据序号	钻屑解吸指标	地质构造	瓦斯异常变化	响煤炮	片帮掉渣	模型预测结果	钻屑指标预测结果	实际突出情况
91	1	1	0	1	0	1	1	1
92	0	0	1	0	1	1	1	0
93	2	1	1	1	0	1	1	1
94	0	0	0	1	1	0	0	0
95	1	0	1	1	0	0	1	0
96	0	1	1	1	1	1	0	1
97	0	1	0	0	0	0	0	0
98	1	0	0	0	0	0	1	0
99	0	0	1	1	1	0	0	0
100	2	0	0	1	0	1	1	1

3 结 论

(1) 决策树预测模型的总体预测准确率 90%,不发生突出的预测准确率 100%,发生突出的预测准确率 80%;钻屑指标预测误差相对较大,总体预测准确率 60%,不发生突出的预测准确率 75%,发生突出的预测准确率 50%。

(2) 采用决策树预测模型预测的结果与实际基本一致,预测准确率明显高于采用单一钻屑指标预测准确率。

由于受到工作面煤与瓦斯突出数据来源的限制,本文对基于决策树 ID3 改进算法的煤层工作面煤与瓦斯突出预测仅限于初步的研究和探讨,模型的进一步完善和应用还有待于后续的深入研究。

参考文献:

[1] 魏风清,史广山,张铁岗. 基于瓦斯膨胀能的煤与瓦斯突出预测指标研究[J]. 煤炭学报,2010,35(S1):95-99.
Wei Fengqing,Shi Guangshan,Zhang Tiegang. Study on coal and gas outburst prediction indexes base on gas expansion energy[J]. Journal of China Coal Society,2010,35(S1):95-99.

[2] Brady B T,Rowell G A Laboratory investigation of the electrodynamics of rock fracture[J]. Nature,1986,321:488-492.

[3] 李成武,何学秋. 工作面煤与瓦斯突出危险程度预测技术研究[J]. 中国矿业大学学报,2005,34(1):72-76.
Li Chengwu,He Xueqiu. Prediction method of coal and gas outburst dangerous level in coal roadway face[J]. Journal of China University of Mining and Technology,2005,34(1):72-76.

[4] Itakura K,Nakajima I,Watanabe Y. AE activity in cross-measure derivate against outburst-prone coal seams-study on AE activity prior to gas outburst(1st report)[J]. Journal of the Mining and Metallurgical Institute of Japan,1988,104(1206):495-503.

[5] 丁 华,王 剑,王 彬. 基于灰关联分析和神经网络的煤与瓦斯突出预测[J]. 西安科技大学学报,2009,29(2):136-139.
Ding Hua,Wang Jian,Wang Bin. Coal and gas outburst forecast based on ANN and grey correlation[J]. Journal of Xi'an University of Science and Technology,2009,29(2):136-139.

[6] 由 伟,刘亚秀,李 永,等. 用人工神经网络预测煤与瓦斯突出[J]. 煤炭学报,2007,32(3):285-287.
You Wei,Liu Yaxiu,Li Yong,et al. Predicting the coal and gas outburst using artificial neural network[J]. Journal of China Coal Society,2007,32(3):285-287.

[7] 朱 玉,张 虹,苏 成. 基于免疫遗传算法的煤与瓦斯突出预测研究[J]. 中国矿业大学学报,2009,38(1):125-130.
Zhu Yu,Zhang Hong,Su Cheng. Coal and gas outburst forecasting based on immune genetic algorithm[J]. Journal of China University of Mining and Technology,2009,38(1):125-130.

[8] 施式亮,宋 译,何利文,等. 矿井掘进工作面瓦斯涌出混沌特性判别研究[J]. 煤炭学报,2009,31(6):701-705.
Shi Shiliang,Song Yi,He Liwen,et al. Coal and gas outburst forecasting based on immune genetic algorithm[J]. Journal of China Coal Society,2009,31(6):701-705.

[9] 邵峰晶,于中清. 数据挖掘原理与算法[D]. 北京:中国水利水电出版社,2003:6-12.

[10] Han J,Kamber M. Data mining:concepts and techniques[M]. Beijing:Higher Education Press,2001:2-10.

[11] David Hand,Heikki Mannida,Padhraic Smyth. Principles of data mining[M]. Beijing:Machinery Industry Press,2003:28-37.

[12] Quinlan J R. Induction of decision trees[J]. Machine Learning,1986:81-106.

[13] 程铁信,郭 涛,祁 昕. 决策树分类模型在工程项目评标风险预警中的应用[J]. 数理统计与管理,2010,29(1):122-128.
Cheng Tiexin,Guo Tao,Qi Ting. Application of decision-tree cluster model in the risk pre-warning for the tender evaluation of civil projects[J]. Journal of Applied Statistics and Management,2010,29(1):122-128.

[14] 姚靠华,蒋艳辉. 基于决策树的财务预警[J]. 系统工程,2005,23(10):102-106.
Yao Kaohua,Jiang Yanhui. Financial early-warning analysis based on decision tree[J]. Systems Engineering,2005,23(10):102-106.

[15] 曲开社,成文丽,王俊红. ID3 算法的一种改进算法[J]. 计算机工程与应用,2003,39(25):104-107.
Qu Kaishe,Cheng Wenli,Wang Junhong. Improved algorithm based on ID3[J]. Computer Engineering and Applications,2003,39(25):104-107.