doi:10.3969/j.issn.1005-3026.2015.04.016

数据挖掘技术在全断面掘进机故障诊断中的应用

张天瑞[1], 于天彪[1], 赵海峰[2], 王宛山[1]
(1.东北大学 机械工程与自动化学院, 辽宁 沈阳 110819;
2.北方重工集团有限公司 全断面掘进机国家重点实验室, 辽宁 沈阳 110141)

摘 要:分析了全断面掘进机复杂的故障机理和运行参数,研究了将粗糙集和决策树应用到数据挖掘中的方法.以全断面掘进机刀盘的一些实时数据为例,采用MATLAB 7.0对数据进行离散化处理,结合粗糙集属性约简的算法对故障样本进行冗余属性的约简;然后,利用决策树算法对约简后的故障样本集进行规则提取,利用数据挖掘工具Clementine实现了C4.5算法和改进的C4.5算法,对其结果进行了对比分析;最后,运用VB编程对全断面掘进机采集的部分数据进行测试,结果表明该融合算法是一种快速、有效、可靠的故障检测与诊断的新途径.

关 键 词:全断面掘进机;数据挖掘;粗糙集;决策树;融合算法
中图分类号:TH 17 **文献标志码**:A **文章编号**:1005-3026(2015)04-0527-06

Application of Data Mining Technology in Fault Diagnosis of Tunnel Boring Machine

ZHANG Tian-rui[1], *YU Tian-biao*[1], *ZHAO Hai-feng*[2], *WANG Wan-shan*[1]
(1. School of Mechanical Engineering & Automation, Northeastern University, Shenyang 110819, China; 2. State Key Laboratory of Tunnel Boring Machine, Northern Heavy Industries Group Co., Ltd., Shenyang 110141, China. Corresponding author: ZHANG Tian-rui, E-mail: tianjiangruixue@126.com)

Abstract: Complex fault mechanism and operation parameters of the tunnel boring machine (TBM) were analyzed, and the method of rough set and decision tree algorithm applying to data mining was studied. Take several MATLAB 7.0 dispersed data of tunnel boring machine cutter head as an example, the redundancy attribute of fault samples was reduced by the combination with the rough set attribute reduction algorithm. The rules were extracted with the decision-making tree algorithm. The C4.5 algorithm and the improved C4.5 algorithm were implemented with the data mining tool Clementine, with the results compared. The data was tested by the VB programming. The results showed that the fusion algorithm is a rapid, effective and reliable approach for fault detection and diagnosis.

Key words: tunnel boring machine; data mining; rough set; decision tree; fusion algorithm

故障诊断是现在机械设备中保证设备正常运行的重要功能[1].目前,故障诊断的研究方法和研究对象有很多,但将全断面掘进机作为研究对象,应用数据挖掘技术的诊断方法的研究还鲜有介绍[2].全断面掘进机的数据采集系统中储存了海量的隐藏着大量潜在规则的数据须被挖掘[3],因此,本文把数据挖掘技术引用到全断面掘进机的故障诊断中.

全断面掘进机体积庞大、结构复杂,仅依靠专家经验排查故障十分困难,因此及时排查故障、减少维修停机时间,将会产生明显的经济效益[4].有研究表明:全断面掘进机遇到故障时,技术人员需要约占总时间70% ~90%的时间确定故障原因和部位,故障维修工作只占约10% ~30%的时间[5];从维修成本的角度来看,预测维修成本只占事后维修成本的40%左右[6].

收稿日期:2014-04-23
基金项目:国家重点基础研究发展计划项目(2010CB736007);中央高校基本科研业务费专项资金资助项目(N110603007).
作者简介:张天瑞(1985-),男,河北深州人,东北大学博士研究生;于天彪(1968-),男,吉林榆树人,东北大学教授,博士生导师;王宛山(1946-),男,辽宁沈阳人,东北大学教授,博士生导师.

1 全断面掘进机故障分析

1.1 全断面掘进机的数据来源

全断面掘进机自带的数据采集系统可收集到包括作业数据、状态信息、工况数据、报警信息在内的大量施工数据. 利用数据库访问技术读取数据采集系统中的数据,根据全断面掘进机数据采集系统采集的历史运行状态数据,对掘进机的未来运行状态进行预测与诊断.

1.2 全断面掘进机的故障分析

全断面掘进机发生故障时,从数据采集系统及故障监视系统可判断属于电气故障、液压故障还是机械故障. 但是由于全断面掘进机的复杂结构而使得故障症状和原因呈多样性. 全断面掘进机出现的故障及成因分析如图 1 所示[7].

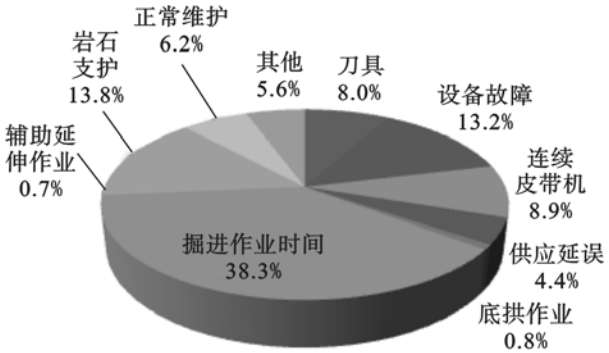

图 1 全断面掘进机施工影响因素

Fig. 1 Influence factors of TBM construction

1.3 全断面掘进机受控对象的选择

根据全断面掘进机施工中各个系统部件的重要性和对整台机器故障造成影响程度的大小,选择主轴承、主电机、主变速箱、主机液压系统作为重点监测对象.

1.4 全断面掘进机的常见施工故障

全断面掘进机常见施工故障包括:主轴承油脂润滑系统故障、胶带输送机故障、水冷器件出现温度超高情形、除尘风机过滤效果不佳等[8].

2 故障诊断中的融合诊断

2.1 基于数据挖掘算法的故障模型

融合了粗糙集和决策树两种算法,主要是利用粗糙集强大的属性约简能力[9-10]和 C4.5 算法的快速分类的优点进行故障规则的挖掘[10],步骤见图 2.

图 2 基于粗糙集的决策树诊断算法模型

Fig. 2 Diagnosis decision tree model based on rough set

首先,利用等宽算法对采集来的连续故障数据进行离散化处理,然后利用粗糙集理论对属性进行约简,获得最佳属性约简表;最后利用决策树算法建立故障决策树,获得最佳的故障诊断规则,把生成的这些规则储存到故障规则库中进行故障诊断,具体故障模型如图 3 所示.

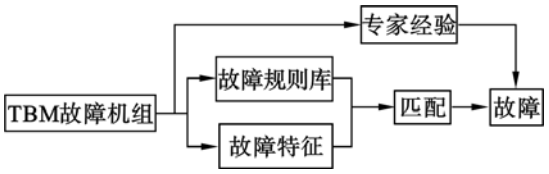

图 3 故障诊断的基本模型

Fig. 3 Basic fault diagnosis model

2.2 基于粗糙集理论的故障特征参数选择

1) 数据预处理. 数据预处理是数据挖掘过程中的重要部分[7],其流程如图 4 所示. 所谓连续属性的离散化,是指将数值的属性值划分成若干子区间,并以此区间代替原有的实值,从而使决策表范化. 通过对掘进机的部分故障样本数据分析,本文采用等距离划分算法将故障样本信息表进行离散化.

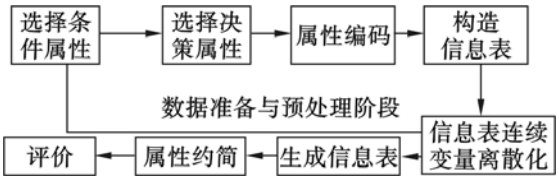

图 4 粗糙集的属性约简流程图

Fig. 4 Flow chart of rough set attributes reduction

以刀盘单元为例,在选取故障条件属性时应结合领域专家的知识,同时考虑到现场测点的布置. 全断面掘进机数据采集系统保存与刀盘有关的参数,当刀盘发生故障时某些参数就会升高或降低. 所以选取以下参数作为粗糙集的条件属性:a 为 No. 2 刀盘电流(A);b 为 No. 6 刀盘电流;c 为土砂房间压力(上)(bar);d 为土砂房间压力(下);e 为土砂房间压力(左);f 为土砂房间压力(右);g 为注浆流量(m^3/s);h 为全断面掘进机总推力(kN),刀盘打滑作为决策属性. 建立故障样本属性决策表,见表 1.

表 1　全断面掘进机故障诊断决策表(部分)
Table 1　TBM fault diagnosis decision list(portion)

故障样本	条件属性								决策属性
	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	
A	95	97	170	220	185	156	4	12 023	1
B	92	96	189	215	182	162	4	12 056	0
C	90	91	156	190	153	174	5	12 000	0
D	89	90	160	191	200	176	5	12 000	1
E	89	92	198	189	150	170	5	12 005	0
F	98	95	171	210	180	150	4	12 043	0
G	94	98	190	221	180	189	4	12 138	1
H	83	85	180	178	173	148	3	13 032	0

利用 MATLAB7.0 编程,实现对此数据的离散化处理.离散化的结果为

$$\boldsymbol{A}=\begin{vmatrix}3&3&1&3&2&0&2&0\\2&3&3&3&2&1&2&0\\1&1&0&1&0&2&3&0\\1&1&0&1&3&2&3&0\\1&1&3&1&3&2&3&0\\3&3&1&2&2&0&2&0\\2&3&3&3&2&3&2&0\\0&0&2&0&1&0&0&3\end{vmatrix}.$$

2）条件属性的约简.决策表属性约简的过程就是在保持信息系统分类能力不变的前提下,从决策表系统的条件属性中,去掉不必要的或对决策不重要的条件属性.采用基于可辨识矩阵的属性约简算法对上述离散结果进行决策属性约简.根据可辨识矩阵定义,得出故障数据的可辨识矩阵,见表 2.

根据可辨识矩阵建立相应的析取逻辑表达式:

$$L_1=a\cup c\cup f,$$
$$L_2=a\cup b\cup c\cup d\cup e\cup f\cup g,$$
$$L_3=a\cup b\cup c\cup d\cup e\cup g\cup h,$$
$$L_4=a\cup b\cup d\cup e\cup f\cup g,$$
$$L_5=a\cup c\cup d\cup f,$$
$$L_6=c,L_7=d,L_8=e,L_9=f.$$

将所有的析取逻辑表达式进行合取运算得:

$$\begin{aligned}\mathrm{RED}(C)&=L_1\cap L_2\cap L_3\cap L_4\cap L_5\cap L_6\cap L_7\cap L_8\cap L_9\\&=c\cup f\cup d\cup e.\end{aligned}$$

从约简后的决策表 3 可以看出,通过粗糙集的属性约简大大降低了决策表的复杂程度.

表 2　全断面掘进机数据可辨识矩阵
Table 2　TBM data distinguishable matrix

U	1	2	3	4	5	6	7	8
1	0	*acf*	*abcdefg*	0	*abcdg*	*d*	0	*abcdegh*
2		0	0	*abcdefg*	0	0	*f*	0
3			0	e	0	0	*abcdefg*	0
4				0	*c*	*abcdefg*	0	*abcdefgh*
5					0	0	*abdefg*	0
6						0	*acdf*	0
7							0	*abcdefgh*
8								0

表 3　TBM 故障样本训练集
Table 3　TBM fault sample training set

序号	*c*	*d*	*e*	*f*	样本输出故障决策
1	1	3	2	0	1
2	3	3	2	1	0
3	0	1	0	2	0
4	0	1	3	2	1
5	3	1	3	2	0
6	1	2	2	0	0
7	3	3	2	3	1
8	2	0	1	0	0

3　全断面掘进机故障识别方法实现

全断面掘进机故障诊断系统数据中蕴含了大量的信息,采用决策树可以提取出不同特征数据中存在的规律,并以规则的形式表现出来.利用这些故障诊断规则对故障数据进行状态预测,对未知数据样本的预测分类提供有力的决策支持.

3.1　C4.5 算法在全断面掘进机故障诊断中的应用

采用决策树算法,对表 3 的故障信息样本进行决策树分析.

由公式 $I(S_1,S_2,\cdots,S_m)=-\sum_{i=1}^{m}p_i\lg(p_i)$ 计算给定样本分类所需的期望信息,设 S_1 无故障,S_2 有故障.

$I(S_1,S_2)=I(5,3)=$

$-[(5/8)\mathrm{lb}(5/8)+(3/8)\mathrm{lb}(3/8)]=0.95.$

对子集的信息量计算:

属性 c 有 4 个取值,分别为 0,1,2,3.

$c=0$ 时,$S_{11}=1,S_{21}=1$,

$I(S_{11},S_{21})=-\frac{1}{2}\mathrm{lb}\frac{1}{2}-\frac{1}{2}\mathrm{lb}\frac{1}{2}=1$;

$c=1$ 时,$S_{12}=1,S_{22}=1$,

$I(S_{12},S_{22})=-\frac{1}{2}\mathrm{lb}\frac{1}{2}-\frac{1}{2}\mathrm{lb}\frac{1}{2}=1$;

$c=2$ 时,$S_{13}=0,S_{23}=1$,

$I(S_{13},S_{23})=0$;

$c=3$ 时,$S_{14}=2,S_{24}=1$,

$I(S_{14},S_{24})=-\frac{2}{3}\mathrm{lb}\frac{2}{3}-\frac{1}{3}\mathrm{lb}\frac{1}{3}=0.913.$

根据公式:

$E(A)=\sum_{j=1}^{n}\frac{S_{1j}+S_{2j}+\cdots+S_{mj}}{S}I\times(S_{1j},S_{2j},\cdots S_{mj})$,

得到属性 c 的信息熵:

$\therefore E(c)=\frac{2}{8}I(S_{11},S_{21})+\frac{2}{8}I(S_{12},S_{22})+\frac{3}{8}I(S_{14},S_{24})=0.84$,

$\mathrm{Gain}(c)=0.11$,

$\mathrm{Splitl}(c)=-(\frac{2}{8}\mathrm{lb}\frac{2}{8}+\frac{2}{8}\mathrm{lb}\frac{2}{8}+\frac{3}{8}\mathrm{lb}\frac{3}{8})=1.53.$

同理可求属性 d,e,f 的期望信息:

$d=0,I(S_{15},S_{25})=0$;

$d=1,I(S_{16},S_{26})=0.913$;

$d=2,I(S_{17},S_{27})=0$;

$d=3,I(S_{18},S_{28})=-0.913.$

$\therefore E(d)=\frac{3}{8}I(S_{16},S_{26})+\frac{1}{8}I(S_{17},S_{27})+\frac{3}{8}I(S_{18},S_{28})=0$,

$\mathrm{Gain}(d)=0.95$,

$\mathrm{Splitl}(d)=-(\frac{3}{8}\mathrm{lb}\frac{3}{8}+\frac{3}{8}\mathrm{lb}\frac{3}{8})=1.066.$

$e=0,I(S_{19},S_{29})=0$;

$e=1,I(S_{110},S_{210})=0$;

$e=2,I(S_{111},S_{211})=1$;

$e=3,I(S_{112},S_{212})=1.$

$\therefore E(e)=\frac{4}{8}I(S_{111},S_{211})+\frac{2}{8}I(S_{112},S_{212})=0.75$,

$\mathrm{Gain}(e)=0.2$,

$\mathrm{Splitl}(e)=-(\frac{4}{8}\mathrm{lb}\frac{4}{8}+\frac{2}{8}\mathrm{lb}\frac{2}{8})=1.$

$f=0,I(S_{113},S_{213})=0.913$;

$f=1,I(S_{114},S_{214})=0$;

$f=2,I(S_{115},S_{215})=0.913$;

$f=3,I(S_{116},S_{216})=0.$

$\therefore E(f)=\frac{3}{8}I(S_{113},S_{213})+\frac{3}{8}I(S_{115},S_{215})=0.688$,

$\mathrm{Gain}(f)=0.27$,

$\mathrm{Splitl}(f)=-(\frac{3}{8}\mathrm{lb}\frac{3}{8}+\frac{3}{8}\mathrm{lb}\frac{3}{8})=1.066$,

$\mathrm{GainRatio}(c)=0.072,\mathrm{GainRatio}(d)=0.89,\mathrm{GainRatio}(e)=0.2,\mathrm{GainRatio}(f)=0.25.$ 从上面的计算结果可以看出 d 的信息增益率最大,所以选择 d 的属性作为测试属性.

当 $d=0,d=2$ 时,出现故障;$d=3$ 时,有 2 个正例,2 个反例. 计算知 e 的信息增益率最大,用递归的方法重复计算,建立决策树,如图 5 所示.

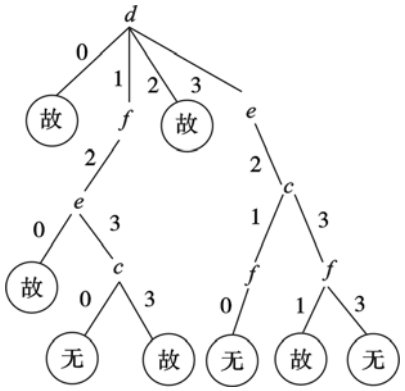

图 5 刀盘故障决策树

Fig. 5 Cutter head fault decision-making tree

3.2 改进的 C4.5 算法的应用

本文主要是研究全断面掘进机的故障诊断,监测数据规模较大. 如果直接利用这些公式生成理想的决策树规则,则是个庞大的计算过程. 因此,利用麦克劳林公式消除公式中的对数运算,从而大大节约了决策树生成的时间.

$E'(c)=\sum_{i=1}^{n}\frac{p_in_i}{p_i+n_i}=(\frac{1\times1}{1+1}+\frac{1\times1}{1+1}+\frac{0\times1}{0+1}+\frac{2\times1}{2+1})=1.67$,

$$\text{Splitl}'(c) = \sum_{i=1}^{n} \frac{p_i n_i}{(n_i + p_i)^2} =$$

$$\left[\frac{1\times1}{(1+1)^2} + \frac{1\times1}{(1+1)^2} + \frac{0\times1}{(0+1)^2} + \frac{2\times1}{(2+1)^2}\right] = 0.72,$$

$$\text{GainRatio}(c) = \frac{0.95 - 1.67}{0.72} = -1.$$

同理可得，

$E'(d) = 1.33$，$\text{Splitl}'(d) = 0.44$，

$$\text{GainRatio}'(d) = \frac{0.95 - 1.33}{0.44} = -0.86;$$

$E'(e) = 1.5$，$\text{Splitl}'(e) = 0.5$，

$$\text{GainRatio}'(e) = \frac{0.95 - 1.5}{0.5} = -1.1;$$

$E(f)' = 1.33$，$\text{Splitl}'(f) = 0.44$，

$\text{GainRatio}'(f) = -0.86$.

由递归算法可知，改进后的 C4.5 算法得出的结果是不变的，运算速度却有了很大的提高.

3.3　基于 Clementine 的决策树算法实现

采集到的全断面掘进机的故障数据庞大，因此采用数据挖掘软件 Clementine，通过建立数据流就可以完成相应的数据挖掘，省去了复杂的编程工作. 以刀盘打滑故障数据库为例，把数据库中有关全断面掘进机故障信息表导入 Clementine 中，建立模型以得到基于粗糙集与 C4.5 的决策树.

从基于粗糙集与 C4.5 的决策树中，可得到对应的 7 条规则. 将这些规则存入到故障规则库中，可利用这些规则进行全断面掘进机的故障诊断.

由此可知，运用 Clementine 建立决策树大大减少了决策树的计算时间，而且可以直接生成规则表，提高了诊断速度. 利用 Clementine 软件生成的故障诊断规则，可以直接生成网页形式供用户查阅.

根据决策树产生的故障规则判断故障类别，利用决策树 C4.5 算法建立决策树，生成故障规则，对比两种方法的不同. 同样利用 Clementine 建立故障决策树，而生成的决策树规则没有改变.

对所建立的改进的 C4.5 算法与单独运用决策树 C4.5 算法进行对比，仿真结果表明：改进后算法的建树时间缩短了近 20%，详见表 4.

3.4　基于属性约简的决策树预测诊断

为验证决策树生成规则的准确性，运用 VB 设计一个故障测试程序来进行测试.

通过采集此故障的实时数据，输入本故障测试系统，然后进入测试界面. 在窗口输入相应参数，利用决策树生成故障规则得到诊断结果. 利用生成的故障诊断规则，输入 4 组待测故障征兆样本对生成的规则进行测试，测试结果与生成的规则相匹配，可得出正确率达到 98.5%.

表 4　两种算法仿真结果的对比

Table 4　Comparison of two algorithms simulation results

算法模型	维数	建树时间/s	正确率/%
C4.5 算法	8	0.37	98.5
改进的 C4.5 算法	4	0.30	98.5

4　结　　论

1）以全断面掘进机刀盘的一些实时数据为例，分别采用 C4.5 算法和改进的 C4.5 算法对故障信息样本进行决策树分析，得到的结果是一致的，说明改进的 C4.5 算法是正确的.

2）利用数据挖掘工具 Clementine 实现了 C4.5 数据挖掘，快速提取了故障诊断的规则；Clementine 仿真结果表明，改进的 C4.5 算法减少了故障特征获取和诊断决策树构建工作量，提高了诊断速度.

3）与其他算法相比，数据挖掘技术能从大量的故障案例数据中得到故障分类的规则，并将这些规则保存于故障规则库中，便于故障匹配，实现故障诊断的任务，能够大幅度提高诊断精度.

参考文献：

[1] Dai Y Y, Zhao J S. Fault diagnosis of batch chemical processes using a dynamic time warping based artificial immune system [J]. *Industrial & Engineering Chemistry Research*, 2011, 50: 4534 – 4535.

[2] Lee S W, Chang S H, Park K H, et al. TBM performance and development state in Korea[J]. *Procedia Engineering*, 2011, 14: 3170 – 3175.

[3] 张天瑞，代沅兴，武继将，等. 基于虚拟仪器的 TBM 状态监测系统仿真研究[J]. 系统仿真学报，2013，25(8)：1716 – 1723.
(Zhang Tian-rui, Dai Yuan-xing, Wu Ji-jiang, et al. Research on simulation for TBM monitoring system based on virtual instrument[J]. *Journal of System Simulation*, 2013, 25(8): 1716 – 1723.)

[4] 黄克，赵炯，周奇才，等. 基于多变量统计过程监控的盾构机故障诊断[J]. 中国工程机械学报，2012，10(2)：222 – 227.
(Huang Ke, Zhao Jiong, Zhou Qi-cai, et al. Fault diagnosis on shield machines based on multivariable statistical process monitoring[J]. *Chinese Journal of Construction Machinery*, 2012, 10(2): 222 – 227.)

(下转第 541 页)