决策树算法的研究与应用

杨 静[1],张楠男[2],李 建[1],刘延明[1],梁美红[1]
(1.西南石油大学,四川 成都 610500;
2.西南油气田分公司信息中心,四川 成都 610500)

摘 要:主要研究了数据挖掘中决策树算法的基本思想和算法。针对目前钻井过程故障诊断的需求,结合决策树算法的特点,提出了一种基于决策树的钻井过程故障诊断专家系统模型。分析了钻井系统事故状态下的相关特征参数,并对基于决策树的钻井过程状态和知识获取进行了详细的论述。通过实例运用ID3算法实现了决策树的建立,为钻井过程故障诊断奠定了坚实的基础。最后提出了对算法的改进,综合对实际数据的处理结果表明,基于数据挖掘的决策树算法可以很好地识别钻井过程中的不同状态,能够实现故障诊断。

关键词:数据挖掘;决策树;ID3;钻井;故障诊断模型
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2010)02-0114-03

Research and Application of Decision Tree Algorithm

YANG Jing[1], ZHANG Nan-nan[2], LI Jian[1], LIU Yan-ming[1], LIANG Mei-hong[1]
(1.Southwest Petroleum University, Chengdu 610500,China;
2.Information Center of Southwest Oil and Gas Field Company, Chengdu 610500,China)

Abstract:Mainly researches the basic method and algorithm of decision tree in data mining. In view of the requirement of the fault diagnosis in drilling project, combining with the characters of decision tree, proposes the drilling fault diagnosis expert system model based on the decision tree. It analyzes the characters of the drilling fault states, and it makes a detail discourse about the state of drilling process and the knowledge acquired, through the example with ID3,it implements the establishment of decision tree, which fixes the solid foundation for the expert system. At last proposes a method to improve the ID3 algorithm, and combining with the data processing result show that the decision tree algorithm based on data mining can recognize the different drilling states very well and implement the fault diagnosis.

Key words:data mining; decision tree; ID3;drilling; fault diagnosis model

0 引 言

决策树算法是以实例为基础的归纳学习算法,以其易于提取显示规则、计算量相对较小、可以显示重要决策属性和较高的分类准确率等优点而得到广泛的应用。

决策树是一种常用于预测模型的算法,它通过将大量数据有目的地分类,从中找到一些有价值的信息供决策者作出正确的决策[1]。所以,研究决策树生成算法就显得尤为重要。而目前在钻井过程中,存在着大量复杂和不确定的影响因素,很难用精确建模的方式建立适用于实际钻井过程的数学模型,数据挖掘与人工智能理论的发展允许人们可以利用钻井系统实际输入输出数据和专家的丰富知识经验建立不严重依赖于钻井系统内在机理的模型[2]。结合钻井过程状态,笔者在研究了决策树的算法后提出了将决策树应用于钻井工程设计和工艺软件中钻井过程故障诊断的解决方案,以实现故障诊断知识的自动获取与表示,提高故障诊断的效率。

1 决策树的基本思想

顾名思义,决策树的结构,就像是一棵树。它利用树的结构将数据记录进行分类,树的一个叶节点就代表某个条件下的一个记录集,根据记录字段的不同取值建立树的分支;在每个分支子集中重复建立下层节点和分支,便可生成一棵决策树[3]。对生成的决策树进行修剪,很容易得到具有商业价值的信息,供决策者参考。如图1所示。

收稿日期:2009-06-12;**修回日期**:2009-09-05
基金项目:国家重大专项项目(2008ZX05021-006)
作者简介:杨 静(1985-),女,四川人,硕士研究生,研究方向为数据库应用与开发;李 建,教授,硕士生导师,研究方向为数据仓库、数据挖掘和建模仿真等。

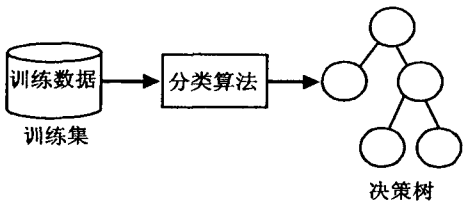

图 1　决策树基本思想

ID3 是引用率较高的决策树算法之一,是 Quinlan 提出的一个著名决策树生成方法。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。而 ID3 算法的核心正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。

设 $E = D_1 \times D_2 \times \cdots \times D_n$ 是 n 维有穷向量空间,其中 D_j 是有穷离散符号集,E 中的元素 $e = < v_1, v_2, \cdots, v_n >$ 叫做例子,其中 $v_j \in D_j, j = 1,2,3,\cdots,n$。设 $s_1, s_2, \cdots, s_m$ 是 E 的 m 个例子集。假设向量空间 E 中的这 m 个例子集的大小为 S_i,ID3 基于下列 2 个假设[4~6]:

1) 在向量空间 E 上的一棵正确决策树对任意例子的分类概率同 E 中这 m 个例子的概率一致。

2) 一棵决策树能对一个例子做出类别判断所需的熵为:

$$\text{Entropy}(s_1, s_2, \cdots, s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

其中 p_i 用 s_i/s 来估算。

如果以属性 A 作为决策树的根,A 具有 v 个值,它将 E 分成 v 个子集 $\{E_1, E_2, \cdots, E_v\}$,假设 E_i 中含有 $S_i (i = 1,2,\cdots,m)$,那么子集 E_i 所需的期望信息是 $E(A)$。

$$\text{Entropy}(A) = -\sum_{j=1}^{v} (s_{1j} + s_{2j} + \cdots + s_{mj})/s * \text{Entropy}(s_{1j}, s_{2j}, \cdots, s_{mj})$$

因此,以属性 A 为根的信息增益是:

$$\text{Gain}(A) = \text{Entropy}(A)(s_1, s_2, \cdots, s_n) - \text{Entropy}(A)$$

ID3 选择使 $\text{Gain}(A)$ 最大的属性 A^* 作为根节点,对 A^* 的不用取值对应的 E 的 v 个子集 E_i 递归调用上述过程生成 A^* 的子节点,从而生成一棵树。

2　基于决策树的钻井故障诊断应用

2.1　决策树与故障诊断

针对文中提出的决策树的基本思想及算法 ID3,以钻井工程设计和工艺为原型,设计了基于决策树的钻井过程故障诊断模型。主要包括数据收集与预处理、告警数据库、知识库、决策树运行、训练样本库、推理机等模块。文中研究的重点在于决策树的建立这一模块。如图 2 中虚线所示。

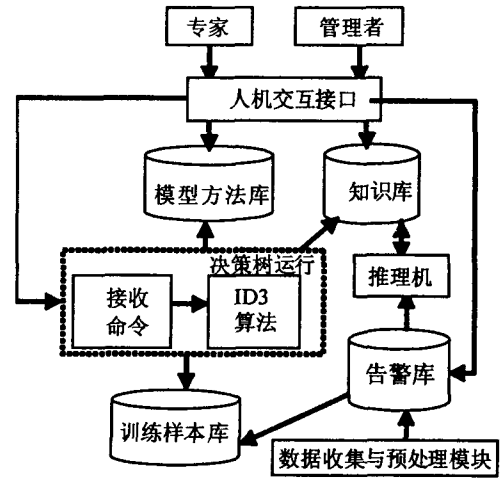

图 2　基于决策树的钻井过程故障诊断专家系统

2.2　钻井过程状态分析

在建立决策树之前,首先要考虑的关键因素是选取建树参数。那么现在的主要环节是分析钻井过程中的主要影响参数。

钻进过程中的系统状态类型众多,依据钻具是否仍然保持连接或活动状态,可以把钻井过程的复杂问题分为井下复杂情况与井下事故两种情形,文中以常见的井下事故为例进行研究。钻井过程的复杂情况及事故产生之前各种工艺参数可能产生缓增、缓降、急增、急降等变化[2],通过分析各种钻井事故中可能产生的参数变化趋势,总结出几种主要的钻井事故与钻井参数表,如表 1。

表 1　井下事故状态与钻井参数

钻井故障	钻压	泵压	泵量	转速	钻速	扭矩
卡钻			……			
烧钻			……			
井塌/埋钻			……			
断钻			……			

表 1 说明钻井过程发生上述 4 种主要事故类型时钻压、泵压、泵流量、转速、钻速、扭矩参数是其主要特征参数[4]。那么选取这 6 个主要特征参数进行分析建树,可以实现钻井专家对钻井过程的异常状态判断。

2.3　基于决策树方法的知识获取

为了验证构造决策树方法在系统知识获取上的有效性,以钻进过程中的常见事故卡钻、烧钻、埋钻和断钻为例进行研究。按上述分析,钻井过程的事故诊断可以选择钻压、泵压、泵量、转速、钻速、扭矩作为决策树的输入参数。那么将这 6 种属性组成故障识别参数集 $A\{A_1, A_2, A_3, A_4, A_5, A_6\}$,其中 A_1 代表钻压;A_2 代表泵压;A_3 代表泵量;A_4 代表转速;A_5 代表钻速;A_6 代表扭矩,共 30 个样本实例来建立故障决策树。选取的值均是反映该参数对应的曲线走向,图中显示为参数名,选取的样本值如图 3 所示。

钻压	泵压	泵量	转速	钻速	扭矩	样本数	钻井故障
水平	持续上升	水平	下降	平缓	下降	2	正常
持续上升	忽高忽低	水平	下降	平缓	水平	2	正常
持续上升	忽高忽低	缓慢上升	下降	平缓	下降	3	卡钻
下降	忽高忽低	缓慢上升	下降	小幅尖峰	上升	7	卡钻
下降	持续上升	突然上升	水平	大幅尖峰	下降	2	烧钻
下降	持续上升	突然上升	上升	大幅尖峰	水平	2	井塌/埋钻
下降	持续上升	缓慢上升	上升	小幅尖峰	下降	3	烧钻
持续上升	持续下降	小幅振动	下降	平缓	水平	4	卡钻
持续上升	持续下降	缓慢上升	水平	大幅尖峰	下降	3	烧钻
水平	忽高忽低	缓慢上升	下降	小幅尖峰	水平	2	断钻

图 3 训练样本集与期望值

最终需要分类的属性为钻井故障，它有 5 个不同的值，分别为正常、卡钻、烧钻、井塌/埋钻和断钻，分别对应的样本数为 4、14、8、2、2。那么首先要计算的是每个属性的信息熵，利用文中 ID3 算法中熵的计算公式得出：

$$\text{Entropy}(S_1,S_2,S_3,S_4,S_5) = -\frac{4}{30}\log_2\frac{4}{30} - \frac{14}{30}\log_2\frac{14}{30} - \frac{8}{30}\log_2\frac{8}{30} - \frac{2}{30}\log_2\frac{2}{30} - \frac{2}{30}\log_2\frac{2}{30} = 1.8947$$

接下来计算每个属性的熵。首先从 A_1 属性开始，观察 A_1 的每个样本值分布，得出：

$$\text{Entropy}(A_1) = \frac{4}{30}\left(\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4}\right) + \frac{12}{30}\left(\frac{2}{12}\log_2\frac{2}{12} + \frac{7}{12}\log_2\frac{7}{12} + \frac{3}{12}\log_2\frac{3}{12}\right) + \frac{14}{30}\left(\frac{7}{14}\log_2\frac{7}{14} + \frac{5}{14}\log_2\frac{5}{14} + \frac{2}{14}\log_2\frac{2}{14}\right) = 1.2877$$

类似地，其他属性的值如图 4 所示，根据 ID3 算法的原则，选取信息增益最大的值作为树的根。对比图 4 可得，应将 A_3（泵流量）作为树的根节点，而“泵流量”的值包括 4 部分，所以该树将产生 4 个分支，最后在 4 个分支的基础上以相同的属性选择算法递归构造各自的子节点以及最终的叶节点，根据此原理得出的树如图 5 所示。

迭代次数	测试属性	信息熵	条件熵	信息增益
1	钻压	1.8947	1.2877	0.6070
1	泵压	1.8947	1.4531	0.4416
1	泵流量	1.8947	0.9825	0.9122
1	转速	1.8947	1.1210	0.7737
1	钻速	1.8947	1.2509	0.6438
1	扭矩	1.8947	1.4735	0.4212
2	钻压	1.1033	0.2687	0.8346
2	泵压	1.1033	0.4585	0.6448
2	泵流量	1.1033	0.4585	0.6448
2	转速	1.1033	0.8258	0.2775
2	钻速	1.1033	0.3673	0.7360
…	…	…	…	…

图 4 测试特征计算结果表

从图 5 的树根开始遍历决策树，可以得到以下的分类规则：

1. IF 泵量曲线成水平状态，THEN 钻井正常；

2. IF 泵量曲线程缓慢上升，钻压曲线上升，钻速曲线平缓，THEN 钻井卡钻故障；

3. IF 泵量曲线程缓慢上升，钻压曲线上升，钻速呈大尖峰状，THEN 钻井烧钻；

4. IF 泵量曲线程缓慢上升，钻压曲线下降，扭矩曲线上升，THEN 钻井卡钻；

5. IF 泵量曲线程缓慢上升，钻压曲线下降，扭矩曲线下降，THEN 钻井烧钻；

6. IF 泵量曲线程缓慢上升，钻压曲线水平，THEN 钻井断钻故障；

7. IF 泵量曲线程突然上升，转速曲线下降，THEN 卡钻；

8. IF 泵量曲线程突然上升，转速曲线上升，THEN 井塌/埋钻；

9. IF 泵量曲线小幅震动，THEN 卡钻。

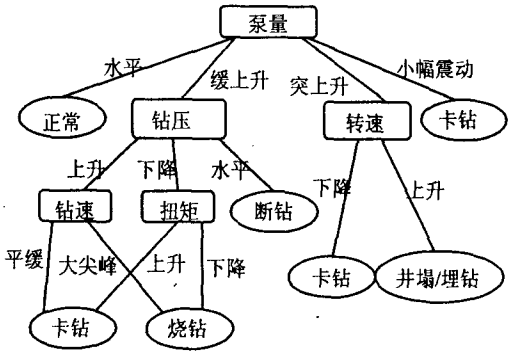

图 5 故障决策树

以上这些规则体现了故障出现的参数特征自合。最终把这些规则存入知识库，利用它对故障分类提供决策依据，给出故障原因。

2.4 ID3 算法的改进

上述算法尽管有效，但该算法往往偏向于选择取值较多的属性，而实际中取值较多的属性有时候并不是最优的。即按照使熵值最小和信息增益最大的原则，被 ID3 算法列为应选取的属性，对其进行测试不会提供太多的信息[7]。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，通过对信息熵的公式加权来增加各个属性重要程度，以加强属性的标注，降低非重要属性的标注。那么利用属性重要性程度，将公式改进为：

$$\text{Entropy}(s_1,s_2,\cdots,s_m) = -\sum_{i=1}^{m} p_i\log_2(p_i)h_i(x)$$

$h_i(x)$ 分别是各自属性的重要性的加权，其取值范围为[0,1]，其大小由训练数据集数据计算出。那么在以上实际项目中，首先用 ID3 算法构造决策树，若出现取值少的重要属性比取值多的非重要属性离根节点

（下转第 120 页）

为,当发送 I 帧时,发送的帧数据增加,即分组数增加,导致分组发送时延和排队时延增加。当发送 P 帧或 B 帧时,分组数据量减少,时延也随之减小。

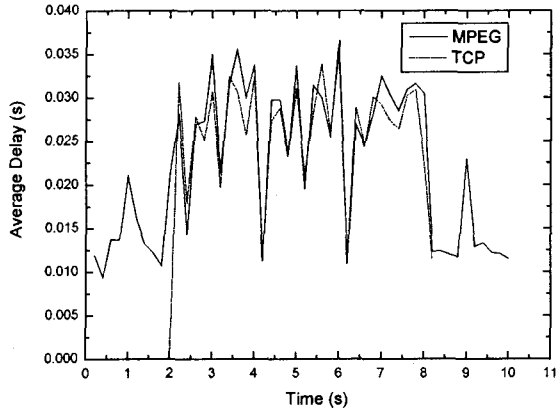

图 6 MPEG 流量与 TCP 流量的平均延时

对文中提出的视频帧优先级传输机制进行了仿真实验,根据仿真实验结果复原传输后的视频帧流,并进行视频回放。由视频回放结果可知,优先级视频传输质量明显优于一致丢弃机制的传输视频质量,尤其表现在眼、口等活动频繁的面部图像元素的视频质量。这是因为,优先级传输机制对 I、P、B 帧标记为优先级从高到低的次序,当网络拥塞时,优先丢弃重要性较低的 B 帧和 P 帧,使重要性高的 I 帧尽可能不被丢弃,从而获得较优的视频传输质量。

5 结束语

文中以视频流的仿真方法为研究背景,对 NS-2 网络仿真器结构和特性进行了深入分析,在此基础上,以视频流仿真为例,对 NS-2 中的流量产生器模块及视频仿真接口进行了扩展,并通过仿真实例,证明了该方法的可行性和有效性。

参考文献:

[1] Ke C H, Shieh C, Hwang W, et al. An evaluation framework for more realistic simulations of MPEG video transmission[J]. Journal of Information Science and Engineering, 2008, 24(2): 425-440.

[2] Klaue J, Rathke B, Wolisz A. EvalVid - A framework for video transmission and quality evaluation[C]//In proc. of the 13th International Conference on Modelling Techniques and Tools for Computer Performance Evaluation. Urbana, Illinois, USA: [s.n.], 2003: 255-272.

[3] Van G, David P, Reisslein M. Traffic characteristics of H.264/AVC variable bit rate video[J]. IEEE Communications Magazine, 2008, 46(11): 164-174.

[4] Van G, David P T, Reisslein M. Traffic and Quality Characterization of Single-Layer Video Streams Encoded with the H.264/MPEG-4 Advanced Video Coding Standard and Scalable Video Coding Extension [J]. IEEE Transactions on Broadcasting, 2008, 54(3): 698-718.

[5] Patrick S, Frank H P, Martin R. Video Traces for Network Performance Evaluation[M]. [s.l.]: Springer Press, 2006.

[6] McCanne S, Floyd S. The LBNL network simulator, ns-2 [EB/OL]. 2008. http://www.isi.edu/nsnam/ns.

[7] 张 铭,窦赫蕾. OPNET Modeler 与网络仿真[M]. 北京:人民邮电出版社,2007.

[8] Welch B B, Jones K, Hobbs J. Practical Programming in Tcl and Tk[M]. fourth edition. [s.l.]: Prentice Hall, 2003.

(上接第 116 页)

的距离远,则用改进的 ID3 算法重新构造决策树进行规则提取。

3 结束语

决策树建模方法不需要涉及钻井系统内在机理模型参数,可以根据实际钻井过程的输入输出数据样本进行自适应学习,实现钻井过程故障诊断。通过对样本数据的测试,应用 ID3 算法建树,系统可以在较短的时间内识别钻井故障。

文中在论述了决策树算法的基础上,结合钻井工程与工艺,将决策树应用于这一领域。利用决策树知识表示与获取集于一身的优点,对基于决策树的钻井过程故障诊断专家系统提出了初步的应用模型。但该算法对大量数据的实现还有待下一步的工作去改进。

参考文献:

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

[2] 王江萍, 孟祥芹, 鲍泽富. 钻井参数实时监测与故障诊断技术[J]. 钻采工艺,2008,31(1):49-52.

[3] Han Jiawei, Kambr M. Data Mining: Concepts and Techniqyes [M]. 北京:高等教育出版社,2001:286-316.

[4] 于润桥. 卡钻事故预测技术研究[J]. 石油钻探技术, 1996, 24(2):15-18.

[5] 洪家荣, 丁明峰, 李星原, 等. 一种新的决策树归纳学习算法[J]. 计算机学报,1995,18(6):470-473.

[6] 毛 国,段立娟,王 实,等. 数据挖掘原理与算法[M]. 北京:清华大学出版社,2007:115-163.

[7] 刘慧魏, 张 雷, 翟军昌. 数据挖掘中决策树算法的研究及其改进[J]. 辽宁师专学报,2005,7(4):23-26.