Journal of Oil and Gas Technology 石油天然气学报, 2017, 39(4), 193-198
Published Online August 2017 in Hans. http://www.hanspub.org/journal/jogt
https://doi.org/10.12677/jogt.2017.394055

Hans汉斯

Study on Drilling Engineering Prewarning Based on Random Forests

Guang Li[1,2], Jie Wang[1*], Jing Liang[1], Caitong Yue[1], Yehuo Fan[2], Dianguang Song[2], Zepeng Lv[3]

[1]Electric Engineering Institute of Zhengzhou University, Zhengzhou Henan
[2]The Seventh Section of 22nd Research Institute of China Electronics Technology Group Corporation, Xinxiang Henan
[3]China Petroleum Chemical Co. Ltd., Zhengzhou Henan
Email: *wj@zzu.edu.cn

Received: Apr. 24th, 2017; accepted: Jul. 22th, 2017; published: Aug. 15th, 2017

Abstract

In consideration of the problems of instability in operation, limitation in installation position that caused low accuracy of drilling abnormal prewarning for data distortion and loss, low accuracy of drilling abnormal prewarning from unstable operation of sensors and newly developed cuttings flow monitoring instrument, the mud logging sensors, down hole sensors and cuttings flow monitoring instrument were used as object; the low warning precision of logging sensors, down hole sensors and cutting flow monitor was analyzed from the aspect of data distortion, data loss, data transmitting difficulty and single system. Abnormal condition discriminating parameters were designed with random forest algorithm in four dimensions, including Euclidean distance, Mahaton distance, GMBR distance and Marsh distance, and with conditions of no abnormality, abnormality increase and abnormality decrease were used as its discriminating target space. A drilling engineering accident warning model was built by using random forest algorithm with Euclidean distance of each parameter as its dimensions and various drilling accident conditions as its target space. Finally, emulation is made with actual field data, and the results show that the warning accuracy for both abnormal engineering parameters and accidents is improved significantly.

Keywords

Drilling Engineering, Data Flow, Random Forest

*通信作者。

文章引用: 李广, 王杰, 梁静, 岳彩通, 范业活, 宋殿光, 吕泽鹏. 基于随机森林的钻井工程预警研究[J]. 石油天然气学报, 2017, 39(4): 193-198. DOI: 10.12677/jogt.2017.394055

基于随机森林的钻井工程预警研究

李　广[1,2]，王　杰[1*]，梁　静[1]，岳彩通[1]，范业活[2]，宋殿光[2]，吕泽鹏[3]

[1]郑州大学电气工程学院，河南 郑州
[2]中国电子科技集团公司第二十二研究所第七研究部，河南 新乡
[3]中国石油化工股份有限公司华北油气分公司，河南 郑州
作者简介：李广(1980-)，男，博士研究生，现主要从事智能计算、机器学习、故障检测等方面的学习与研究。
Email: *wj@zzu.edu.cn

收稿日期：2017年4月24日；录用日期：2017年7月22日；发布日期：2017年8月15日

摘　要

针对因录井传感器工作性能不稳定、安装位置受限等原因造成数据失真和丢失导致的钻井工程异常预报准确率不高，因传感器传输问题无法获知钻井状态导致工程事故预报准确率不高，因新研制的岩屑流量监测仪系统单一无法准确预警等问题，以录井传感器、井下传感器和岩屑流量监测仪为对象，结合其数据失真、数据不完整、数据传输困难、动态数据流、单一系统等特点，从欧氏距离、曼哈顿距离、GMBR距离、马氏距离4个维度，以无异常、异常上升、异常下降为目标空间，采用随机森林算法设计参数异常判断。以各个参数的欧氏距离为维度，以各种钻井事故复杂程度为目标空间，采用随机森林算法设计钻井工程事故复杂预警模型。利用该模型在现场真实数据集上进行仿真，结果表明工程参数异常和事故复杂预报准确率均明显提升。

关键词

钻井工程，数据流，随机森林

1. 引言

石油钻井事故严重威胁着钻井安全。录井传感器受测量原理或安装位置等原因，某些参数存在失真现象，如总池体积传感器受搅拌机影响，测量数据波动较大；把式出口流量传感器受泥浆结饼粘结影响，存在传感器失效，获取不到真正出口流量值。录井传感器均在地面，而预测目标如井漏、溢流发生位置在井下，间接测量加上失真数据导致钻井工程事故预报不准确。针对上述问题，中国电子科技集团公司第二十二研究所研发了井下测量系统，能够对近钻头钻具内外压力、扭矩、温度和密度等参数直接测量，获取井下数据，并研发了岩屑流量监测仪，通过测量岩屑上返质量，结合进尺情况判断地层变化及可能发生的工程事故。随着传感器井下化系统和岩屑流量监测仪技术越来越成熟，在钻井工程事故复杂预警中的作用越来越大，如何综合利用各个系统信息为安全钻井保驾护航成为越来越紧迫的任务。

钻井工程事故异常预报[1] [2]的首要工作是检测参数异常，组合各种参数异常类型及程度，对钻井事故是否发生及严重程度进行报警。目前，主要的机器分类方法有神经网络[3]、SVM (支持向量机) [4]、随机森林(RF, random forests) [5] [6]。神经网络容易陷入局部最小点，SVM 运算量过大，不利于在线运行。RF 是由 Leo Breinan 在 2001 年提出的一种统计学习方法，在生物学、医学、工程、互联网应用等领域取得了很好的效果。为此，笔者初步探讨了 RF 在钻井工程预警中的应用，只针对 RF 在钻井工程参数趋势异常判断和钻井事故模型检测 2 个方面的内容进行研究。

2. 随机森林算法

RF 是一种由多个决策树[7]进行分类的方法。RF 是决策树的一种，随机决策树是组成 RF 的最小决策单元。RF 有 2 个特征：①有放回地随机选择样本数据；②无放回地随机选择训练样本特征来进行随机决策树的节点分裂。采用 CART (classification and regression trees)的方式生成决策树，通常不需要进行剪枝，最大程度生长[8]。通过 Bootstrap Aggregating Method [9]采集样本数据作为每棵树的训练集。结合 Bootstrap Aggregating Method [9]和 Random Subspace Method [10]两种思想，构建多个决策树，组合分类预测结果。

2.1. 生成决策树

在机器学习技术中，数据采样一般采用 Bagging 和 Boosting 两种方式[6]。Bagging 是采用有放回的方式，即从样本库中采其总量的约 2/3 数据作为训练样本，采集后再放回到样本库中，重复 K 次采样，得到 K 个训练样本库。Boosting 采用一次抽样、迭代训练的方式。

RF 采用 Bagging 方式，从 N 个样本库中，随机抽取 n 个数据，抽取后再放回样本库中，再次抽取 n 个数据，重复 K 次，得到 K 个训练数据集，经过训练分类得到 K 个决策树$\{h(X,\theta_k)$，$k=1, 2, \cdots, K\}$，$n < N$。训练数据中，n 以外的数据称为袋外数据 OOB (out of bag) [11]，OOB 数据可以用来预测分类器的精度，综合 K 次评估结果，得到错误率的 OOB 估计，用于评估集成分类器的正确率，但大多采用的是训练数据以外的数据作为测试数据，即测试数据和训练数据无交叉，但两者测试的精度是一样的。由于 Bagging 的训练样本抽取方式，每次生成的决策树训练样本都不一样，可以一定程度地避免过拟合现象。RF 的第 2 个特征是无放回的抽取分裂属性，从 M 个属性中随机抽取 m 个作为单个分类器的属性集，$m < M$。当原始数据集中 M 较少时，可以通过线性组合的方式确定新的属性，提高分类器之间的差异。RF 的 2 个随机性特征，确保不会出现过拟合，同时也能提高精确度和抗噪声能力。

设数据有 M 个输入特征，则在生成随机树时，无放回地随机固定选取 M 个特征中的 m 个，以不纯度最小原则选取 1 个特征进行分支生成；再通过同样的方式，生成下面的所有分支和叶子。RF 相比于决策树有 2 个特征：有放回地随机抽取固定数量训练数据；无放回地随机抽取固定数量 m 个属性。2 个随机性特征保证了 RF 具有较好的抗噪声干扰能力，避免了固定数据训练导致的过拟合现象。

2.2. 投票决策

在分类阶段，最终结果由所有决策树的结果综合而成。常用的方法是概率平均和投票法。由于该次研究的决策基于不同测量系统数据，因此采用投票法进行决策。

$$C_P = \arg\max\left[\frac{1}{N}\sum_{i=1}^{N}\left(I\frac{n_{h_{ic}}}{n_{h_i}}\right)\right] \tag{1}$$

式中：C 是分类结果；P 是类别；N 是森林中决策树的数目；I 是森林中第 i 个决策树的权重；$n_{h_{ic}}$ 是树 h_i

对类 C_P 的分类结果；n_{h_i} 是树 h_i 的叶子节点数。

如果把数据看成一个表，行表示数据，列表示属性，则 RF 算法具有随机选择行数据的特征。采用 Bagging 方式对样本数据集采取有放回的抽取策略，构建与测试数据集相对应的决策树。无放回的随机选择列属性，采用 CART (classification and regression tree) 方式选择一个最优分裂属性。由多个决策树组成一片森林。根据每个决策树的结果，通过投票方式确定最终分类结果(图 1)。

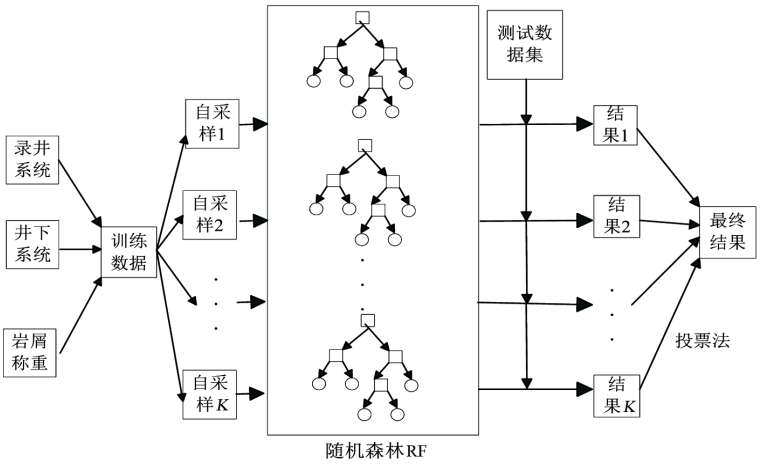

Figure 1. The pre-warning system of drilling engineering based on forest decision
图 1. 钻井工程预警 RF 决策系统

3. 基于 RF 的钻井工程预警研究

基于 RF 的钻井工程事故预警主要应用在 2 个方面：一个是数据异常判断，另外一个是预警模型判断。数据异常判断采用不同特征量的距离作为属性，目标空间为异常上升、异常下降和无异常。预警模型判断采用不同参数的欧氏距离，目标空间为井漏、溢流和刺漏等。由于现场工况复杂，很多正常操作也会导致参数类似于故障异常的特征，再加上传感器数据丢失和失常等原因，导致现场参数趋势异常、事故误报和漏报增多。RF 可以结合几个领域的决策优势，经过综合判断输出一个比单一系统效果好的结果。

地面和井下参数异常判断均采用 RF 方法进行。岩屑流量监测仪数据并入录井数据辅助决策。由于受传输带宽影响，井下传感器参数不能够实时传输到地面，因此井下传感器采用 RF 技术独立运行，只传送事故复杂预报结果到地面，再联合录井和岩屑流量监测仪进行投票决策。压缩感知 CS (compressed sensing) [12]具有前端直接采集压缩后数据、后端计算量大的特点，完全满足了钻井井下数据传输特点和要求。随着 CS 技术的发展，井下传感器数据实时传输到地面后，将提高钻井工程事故检测准确率，降低误报率和漏报率。

图 2 是对某油田某井井漏数据的 RF 仿真试验结果，数据是各个系统单独采样，标签化是通过专家经验，对故障数据开始、结束时刻进行标志化工作。从结果可以看出，井下参数(井底外压、井底内压)对事故的重要性更大，能够更直接地反映井漏事故发生。

图 3 为 RF 训练与实测效果对比图，可以看出，经过专家标定的数据库监督学习，训练 AUC (area under curve)的值(A_{UC})最高能达到 93%；实测 A_{UC} 平均能达到 92%，选择最合适的决策树颗数，最佳 A_{UC} 能达到 92.6%。图 3 与图 4 对比可知，RF 能够获得比传统单一模型高的准确率，及比单一模型低的误报率和漏报率。

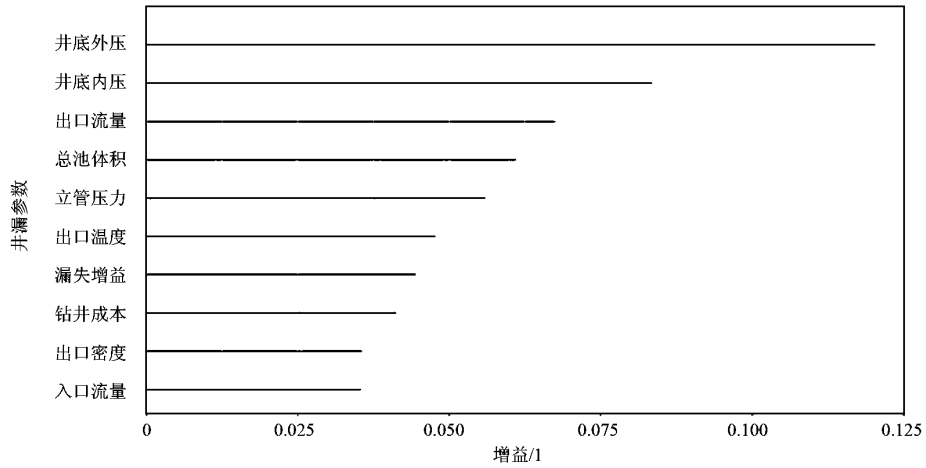

Figure 2. RF diagram of leakage parameter importance analysis
图 2. RF 井漏参数重要性分析图

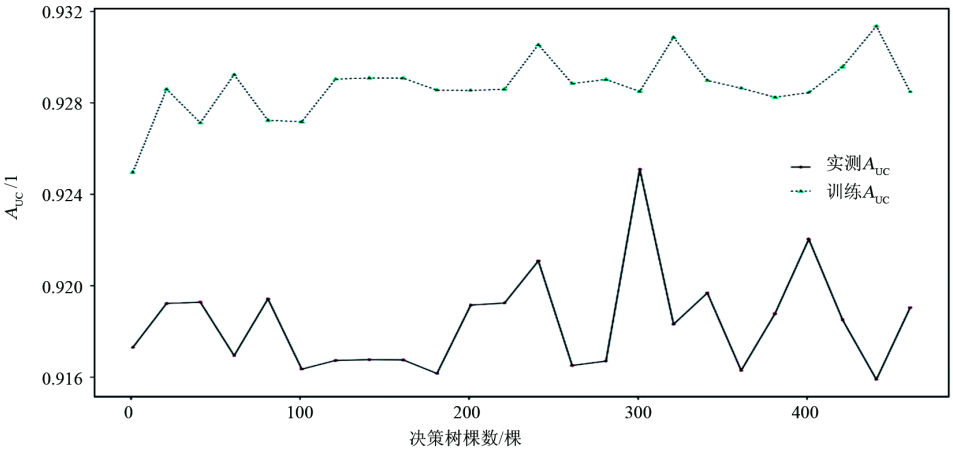

Figure 3. The contrast diagram of RF training and test effect
图 3. RF 训练与测试效果对比图

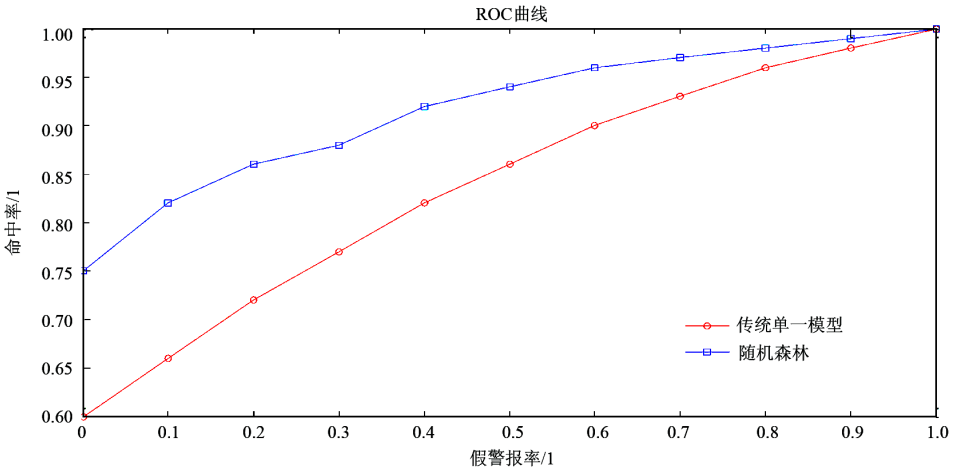

Figure 4. The contrast diagram of traditional single model and RF
图 4. 传统单一模型与 RF 对比图

4. 结语

该次研究的主要成果在于探索了钻井工程事故预警在数据失真、数据丢失、信息片面和测量范围局限性等问题。引入了 RF 算法，利用 RF 算法的随机抽取数据、随机抽取属性，集成多个模型综合决策的特点，解决由数据失真、数据丢失和系统单一造成的钻井工程事故误报率高、漏报率高等问题。通过现场真实数据仿真实验得知，RF 算法不但能够提升事故复杂预报准确率，降低误报率和漏报率，还能够分析出事故复杂的关联参数，为海量钻井数据知识挖掘提供技术支持和探讨，为钻井工程事故预警技术的提升进行有效探索。

基金项目

国家自然科学基金项目(61473266)。

参考文献 (References)

[1] 王杰, 李广, 朱晓东. 基于分层模糊推理的石油钻井事故预警系统[J]. 微计算机信息, 2008, 7(25): 177-178.

[2] 朱晓东, 王杰. 基于分层模糊系统的石油钻井参数预测模型[J]. 石油学报, 2010, 31(5): 838-848.

[3] Zhai, M., Roshtkhari, M.J. and Mori, G. (2016) Deep Learning of Appearance Models for Online Object Tracking.

[4] Weston, J. (2014) Support Vector Machine.

[5] Wager, S. and Athey, S. (2015) Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests.

[6] Zhang, F., Du, B. and Zhang, L. (2016) Scene Classification via a Gradient Boosting Random Convolutional Network Framework. *IEEE Transactions on Geoscience and Remote Sensing*, **54**, 1793-1802. https://doi.org/10.1109/TGRS.2015.2488681

[7] Wang, S., Fan, C., Hsu, C., Sun, Q. and Yang, F. (2014) A vertical Handoff Method via Self-Selection Decision Tree for Internet of Vehicles. *IEEE Systems Journal*, **99**, 1-10.

[8] Petersen, M., Tolver, A., Husted, L., *et al.* (2016) Repeated Measurements of Blood Lactate Concentration as a Prognostic Marker in Horses with Acute Colitis Evaluated with Classification and Regression Trees (CART) and Random Forest Analysis. *The Veterinary Journal*, **213**, 18-23. https://doi.org/10.1016/j.tvjl.2016.03.012

[9] Hassan, A. and Bhuiyan, M. (2016) Computer-Aided Sleep Staging Using Complete Ensemble Empirical Mode Decomposition with Adaptive Noise and Bootstrap Aggregating. *Biomed Signal Process Control*, **24**, 1-10. https://doi.org/10.1016/j.bspc.2015.09.002

[10] Hosseini, M., Hajisami, A. and Pompili, D. (2016) Real-Time Epileptic Seizure Detection from Eeg Signals via Random Subspace Ensemble Learning. *IEEE International Conference on Autonomic Computing* (*ICAC*), 17-22 July 2016, 209-218. https://doi.org/10.1109/ICAC.2016.57

[11] Janitza, S. (2017) On the Overestimation of Random Forest's Out-of-Bag Error.

[12] Bigot, J., Boyer, C. and Weiss, P. (2013) An Analysis of Block Sampling Strategies in Compressed Sensing.