

基于数据挖掘方法的风力涡轮机 状态监测技术研究

刘青凤¹, 李红兰²

(1. 安阳工学院 计算机科学与信息工程学院, 河南 安阳 455000;

2. 河南牧业经济学院 人文与社会科学系, 郑州 450011)

摘要: 目前风力涡轮机的故障模式预测成为了风力发电站发展的重要任务; 提出了一种基于数据挖掘算法的涡轮机故障状态预测方法; 这种方法包括 3 个主要的步骤: 涡轮机状态抽象, 算法训练, 状态预测; 首先利用先验知识将涡轮机的初始状态进行分类, 选择建立预测模型的参数; 为了降低计算难度, 采用数据挖掘算法进行模型参数的选择; 最终采用发电机转速、变速箱速度、温度枢纽、叶片螺距角这些参数进行预测模型的建立; 建立预测模型的过程分为 3 个阶段: 预测任意故障; 预测系统的特殊故障; 确定未知故障; 通过对各种数据挖掘算法基于大量风力涡轮机数据的性能分析, 选择了性能最优的随机森林算法模型; 这种模型的预测准确率能够达到 98%; 同时还能够预测训练数据没有包含的故障类型; 通过在实际风力涡轮机数据的验证, 表明了这种模型的稳健性。

关键词: 风力涡轮机; 数据挖掘; 随机预测

A Data Mining Approach for Wind Turbine State Monitoring

Liu Qingfeng¹, Li Honglan²

(1. Computer Science & Engineering Department, Anyang Institute of Technology, Anyang 455000, China; 2. Department of Humanities & Social Sciences, Henan University of Animal Husbandry and Economy, Zhengzhou 450011, China)

Abstract: As the rapid development of wind farms, it becomes important for wind turbine monitoring and maintenance. As the operating of wind turbine, the state may change from normal to fault. The prediction of fault modes is important for the maintenance of wind turbine. In this paper, we proposed a wind turbine fault modes prediction based on a data mining method. The prediction model contains three steps: prediction of random fault; prediction of special fault; prediction of unseen fault. We chose an optimal random forest algorithm as the data mining approach based on the comparative analysis on the data collected at a large wind farm. The prediction accuracy of the model can achieve 98%, and at the same the model can predict fault modes which are not contained in the training data. Based on the practical wind turbine data, the robustness of the model is verified.

Key words: wind turbine; data mining; random forest algorithm

0 引言

风能作为一种可再生能源, 是未来十年重点发展的方向。随着风力发电站的快速发展, 风力涡轮机的维护也变得非常重要^[1]。传统的周期性的检测和维护组件已经被条件式的检测和维护取代^[2]。条件式检测方法由安装在风力涡轮机上的传感器设备对风力涡轮机进行连续的检测。振动分析、光学应变分析、油颗粒分析是几种常用的分析方法^[3]。另一种常用的方法是性能监测。这种方法利用风力涡轮机的历史数据预测其性能参数。性能监测是一种费用很高的方法, 因为 SCADA 系统记录了风力涡轮机的各种参数, 而这些数据可能会包含错误的信息。

数据挖掘方法是另一种用于监测风力涡轮机性能参数的方法。与数据挖掘方法相关的应用包括故障诊断、异常行为建模、能量曲线监测、状态模式识别等^[4]。本文提出一种基于数据挖掘方法进行风力涡轮机状态预测的方法。通过对各种数据挖掘算法基于大量风力涡轮机数据的性能分析, 选择了性能最

优的随机森林算法模型。这种模型的预测准确率能够达到 98%; 同时还能够预测训练数据没有包含的故障类型。通过在实际风力涡轮机数据的验证, 表明了这种模型的稳健性。

1 风力涡轮机状态监测模型

建立预测模型如图 1 所示。利用专家的经验, 用抽象的涡轮机状态将输出数据归结为不同的状态。模型的建立涉及到不同的数据挖掘算法, 然后选择最优的数据挖掘算法。由得到的数据库进行第一阶段和第二阶段的预测模型的建立^[5]。第一阶段预测的主要目的在于预测任意种类的故障; 第二阶段预测的目的在于预测特定的故障; 在第三阶段, 不同涡轮机的未知故障将得到确认。接下来我们首先描述风力涡轮机的各种故障状态。

1.1 涡轮机状态抽象

典型的涡轮机的状态可能有很多种, 包括正常操作、空转、维护模式、故障模式、天气原因导致的停工等。涡轮机故障模式的预测显得非常重要, 因为这样能够预测到系统存在的潜在风险。涡轮机故障模式的原因可能会有上百种^[6]。基于故障模式的频率, 在本文中我们选择第 12 号涡轮机作为研究对象。为了降低数据挖掘算法的计算量, 我们将记录的风力涡轮机的状态进一步分类。初始的 44 种状态被分为 4 类: 正常 (OK)、故障 (Fault)、天气原因导致的停工 (weather down-

收稿日期: 2014-01-16; 修回日期: 2014-02-26。

基金项目: 国家科技重大专项课题 (2012zx04011-012)。

作者简介: 刘青凤 (1969-), 女, 河南安阳人, 硕士, 讲师, 主要从事计算机软件与理论方向的研究。

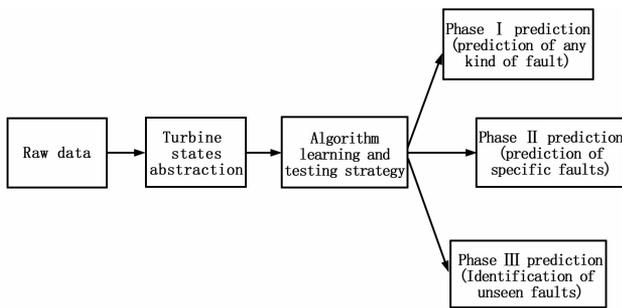


图 1 本文预测模型的建立过程

time)、维护导致的停工 (maintenance downtime)。OK 包含了涡轮机的正常操作状态; 故障包含了系统中实际发生的或者潜在的故障。weather downtime 表示由于恶劣天气导致的停工; 而其它原因导致的停工都包含在 maintenance downtime。

1.2 算法设计

SCADA 系统记录了风力涡轮机超过 100 种的参数。这些参数大概可以分为 3 类: 风力涡轮机性能参数, 风力涡轮机控制参数, 风力涡轮机非控制参数。像功率、发电机转速这一类参数称为性能参数; 叶片螺距角、发电机转矩这一类参数称为控制参数; 风速是唯一的非控制参数^[7]。在本文中我们联合性能参数、控制参数和非控制参数建立风力涡轮机状态的预测模型。为了使最小化数据维度和去掉不相关的参数, 我们采用了参数选择算法。我们采用了记录的涡轮机一个月的数据进行参数选择和算法训练。原数据的一个分层子集被用作参数选择以使得计算更高效。分别采用了三种数据挖掘算法来确定相关参数以便建立预测模型: WGS、WBFS 和 BTA^[8]。最终采用发电机转速、变速箱速度、温度枢纽、叶片螺距角这些参数进行预测模型的建立。

数据挖掘算法的性能评估是基于每一类输出的预测准确度来进行的。考虑到输出类别的不平衡性, 可以采用输出种类的几何平均作为参数选择的标准。式 (1) 所示为输出种类的几何平均:

$$gmean = \sqrt[n]{\prod_{i=1}^n acc_i} \quad (1)$$

式中, acc_i 表示种类 i 的准确度, n 为输出种类的总数^[9]。

最初我们选择了 5 种数据挖掘算法来建立时间点 t 的预测模型。这 5 种数据挖掘算法分别为神经网络 (NN)、支持向量机 (SVN)、随机森林法 (RFA)、增强树算法 (BTA)、通用卡方自动交互检测算法 (CHAID)^[10]。NN 采用后向传播的方法进行示例分类。文中建立了 20 种具有不同核函数的 NN 模型和结构, 然后选择其中准确度最高和最稳定的模型。SVM 在高维空间建立一个或者一组超平面, 并将其用于分类。在 SVM 中, 与任意类的最近训练点距离最远的超平面精度最高。RFA 是一种集中训练方法, 在分类中会产生集中随机树, 在每一个节点选择 k 非随机输入参数。BTA 产生多个节点, 然后将不同模型的预测进行加权以得到一个单独的预测模型。CHAID 是一种基于树的数据挖掘算法^[11]。

表 1 所示为第一阶段预测中每一类的预测精度。基本上所有的算法在预测 ‘OK’ 和 ‘Fault class’ 状态的时候都具有较高的精度; 而在预测 ‘weather downtime’ 和 ‘maintenance downtime’ 状态时精度相对较低。几何平均准则 $gmean$ 表明

表 1 第一阶段的预测结果

Algorithm	Output class				Overall Accuracy (%)
	Turbine OK (%)	Fault (%)	Maintenance downtime (%)	Weather downtime (%)	
SVM	99.1	91.0	30.4	54.5	95.8
CHAID	99.0	89.6	35.5	67.1	96.0
NN	99.5	93.5	62.3	84.9	97.6
BTA	99.6	93.2	83.3	96.9	98.8
RFA	99.8	99.6	78.4	97.9	99.4

当所有种类的预测精度都很高时, 其值为 1。具有最高 $gmean$ 值的算法被选作建立不同时间点的预测模型。从图 2 和表 3 所示的结果可以看出, 增强树算法和随机森林法性能优于其它 3 种数据挖掘算法。然而, 由于 RFA 对数据规模不敏感, 因此我们选择 RFA 作为建立模型的数据挖掘算法。图 3 所示为随机森林算法的树的复杂度与错误分类的关系。从中可以看出最优的树的数量为 91。

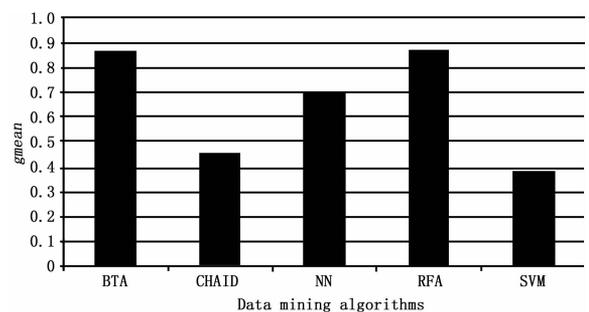


图 2 采用 $gmean$ 作为评价准则的不同数据挖掘算法性能

在第二阶段也采用这 5 种相同的数据挖掘算法建立预测模型。在第一阶段预测中的 ‘fault’ 被实际的故障类别所取代。图 3 所示为时刻 t 的数据分布结果。当有动叶片的限位开关的角度为非正时会触发状态 ‘pitch overrun 0’。当晶闸管没有准备好就会触发状态 ‘Pitch thyristor 2 fault’, 即使网格线已经打开。‘Axle 1 fault pitch controller’ 表示轴扰动。‘Pulse sensor rotor monitor defect’ 是由于发电机超速以后没有冲激加速监测器。表 2 所示为时刻 t , 不同数据挖掘算法的性能。从中可以看出大多数算法在预测数量较少的输出状态时会出现错误, 从而会导致 $gmean$ 值为 0, 如图 4 所示。如期望的一样, RFA 比其它算法的性能更优, 预测耳的准确度也更高。

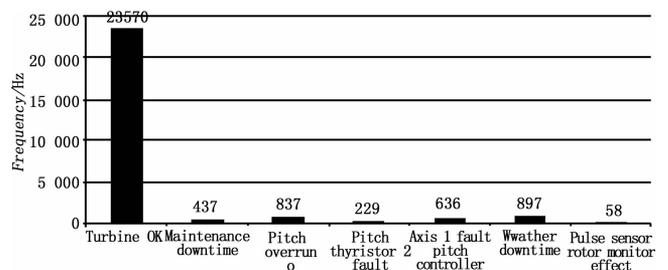


图 3 时刻 t 的输出种类分布

表 2 时刻 t 不同数据挖掘算法的性能

Algorithm	Output Class							Overall Accuracy (%)
	Turbine OK (%)	Maintenance downtime (%)	Weather downtime (%)	Axle 1 fault pitch controller (%)	Pitch overrun 0 (%)	Pitch thyristor 2 fault (%)	Pulse sensor rotor monitor defect (%)	
SVM	99.16	26.82	56.73	51.14	67.12	0.00	0.00	93.08
CHAID	98.98	0.00	97.17	100.0	08.30	30.55	0.00	93.42
NN	99.73	87.19	93.10	99.54	97.23	54.16	37.50	98.65
BTA	99.03	84.96	24.46	35.43	67.37	0.00	0.00	92.88
RFA	99.64	82.70	99.29	100.0	98.93	87.83	61.90	98.83

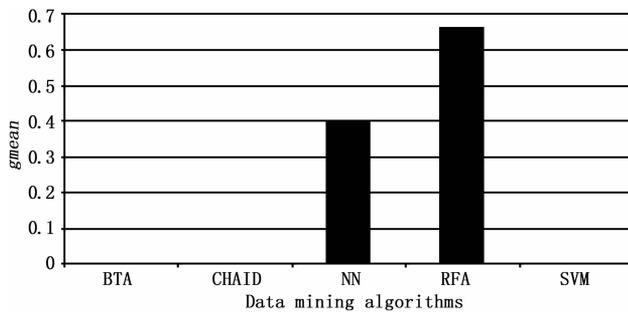


图 4 第二阶段预测中不同算法的性能对比

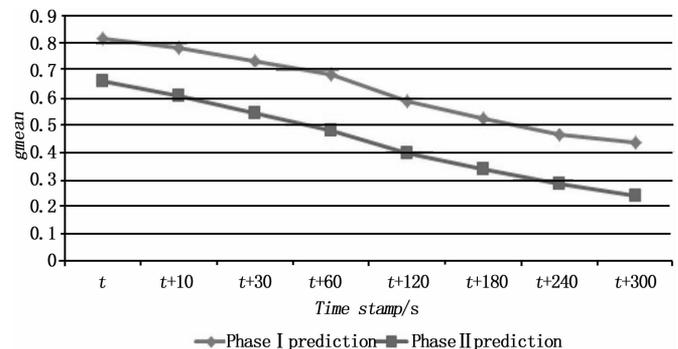


图 5 第一阶段和第二阶段预测中不同时刻的 $gmean$ 值

2 试验结果与分析

在本节我们以 RFA 作为数据挖掘算法建立不同时刻的预测模型，最大的预测时间长度为 5 min。RFA 中树的最大规模为 300。如表 3 所示所有种类的预测精度的范围为 81%~99%。

表 3 第一阶段中不同时刻 RFA 的预测精度

Time stamp(s)	Output Class				Overall Accuracy (%)
	Turbine OK (%)	Fault (%)	Maintenance downtime (%)	Weather downtime (%)	
t	99.88	99.67	78.41	97.91	99.45
$t+10$	99.56	99.00	77.22	95.04	98.39
$t+30$	97.64	96.41	74.59	94.62	96.54
$t+60$	95.70	95.64	71.67	92.55	94.43
$t+120$	91.87	90.00	67.49	88.47	90.89
$t+180$	88.58	87.34	64.94	84.43	86.82
$t+240$	85.62	84.64	60.31	82.44	83.93
$t+300$	83.05	82.76	59.67	80.39	81.76

2.1 故障检测

在这一阶段中，‘fault’被实际中的故障种类所代替，比如 ‘pitch overrun 0’, ‘Pitch thyristor 2 fault’, ‘Axle 1 fault pitch controller’, ‘Pulse sensor rotor monitor defect’。

表 4 所示为 RFA 在不同时刻的预测结果。每一种类的预测精度的范围为 68%~100%，除了对 ‘Pulse sensor rotor monitor defect’ 预测结果以外。图 5 所示为两个预测阶段的 $gmean$ 值。第一阶段的 $gmean$ 值比第二阶段的 $gmean$ 值明显更高。这是因为对 ‘Pulse sensor rotor monitor defect’ 预测精度较低造成的。

2.2 结果分析

基于测试数据的结果表明了 RFA 算法的有效性，为了验证这种算法的稳健性，我们还需要验证这种模型对未知状态类型的预测结果。由于风力涡轮机内在的变化性，风力涡轮机的故障状态也千变万化。因此有必要验证这种新模型对未知故障状态的预测性能。在本节，我们采用其它风力涡轮机的数据，第 10 号，第 14 号，第 17 号来验证算法的稳健性。利用第一阶段建立的预测模型对这些数据进行分析。研究的故障状态包括偏航失控、碳刷磨损报警等。图 6 所示为这几台涡轮机的实际输出状态分布图。这几台涡轮机的故障数量不同，但是所有涡轮机在大多数时间段都是正常工作的。从中我们可以看出，这种基于 RFA 的预测模型能够准确预测未知的故障状态。对未知故障状态的预测准确率在 60%~100%，除了对与变速箱故障相关的预测结果以外，比如变速箱温度过高，变速箱油压过低等。这些状态通常会被预测为正常操作。这其中的主要因为模型中缺少相关的输入参数。表 5 结果表明这种模型能够预测大多数涡轮机的故障状态。

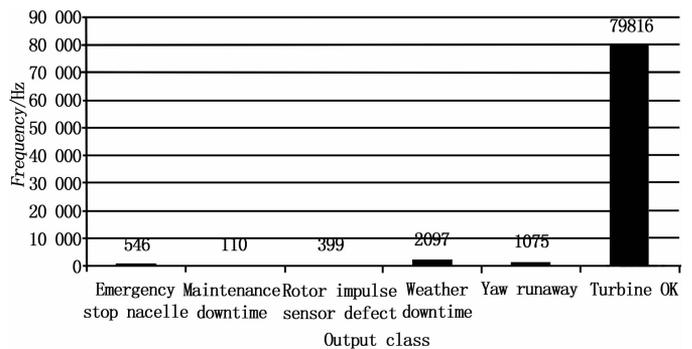


图 6 第 10 号涡轮机的输出种类分布

表 4 第二阶段中 RFA 在不同时刻的预测结果

Time stamp[s]	Output Class							Overall Accuracy(%)
	Turbine OK(%)	Maintenance downtime(%)	Weather downtime(%)	Axle l fault pitch controller(%)	Pitch overrun 0°(%)	Pitch thyristor 2 fault(%)	Pulse sensor rotor monitor defect(%)	
t	99.64	82.70	99.29	100.0	98.93	87.83	61.90	98.83
$t+10$	99.34	81.08	97.84	98.67	97.77	85.43	58.94	96.44
$t+30$	97.15	79.26	95.23	97.13	95.52	83.29	55.03	94.09
$t+60$	95.28	76.87	92.86	95.45	93.26	80.09	51.98	91.68
$t+120$	90.10	71.58	88.41	91.90	90.23	77.67	48.83	87.82
$t+180$	87.98	68.71	85.73	86.39	86.88	74.87	46.29	84.69
$t+240$	84.45	65.32	83.66	82.55	83.34	71.45	43.91	81.53
$t+300$	82.76	62.43	81.45	80.76	79.55	68.32	40.67	78.35

表 5 第 10 号涡轮机的状态预测

Actual output	Anticipated output	Correctly identified cases(%)
Emergency stop nacelle	Fault	85.66
Maintenance downtime	Maintenance downtime	100
Rotor impulse sensor defect	Fault	60.90
Weather downtime	Weather downtime	69.84
Yaw runaway	Fault	99.62
Turbine OK	Turbine OK	99.96

3 结论

本文提出了一种基于数据挖掘算法的涡轮机故障状态预测方法。为了降低计算难度,我们采用数据挖掘算法进行模型参数的选择。最终采用发电机转速、变速箱速度、温度枢纽、叶片螺距角这些参数进行预测模型的建立。建立预测模型的过程分为 3 个阶段:预测任意故障;预测系统的特殊故障;确定未知故障。通过对各种数据挖掘算法基于大量风力涡轮机数据的性能分析,选择了性能最优的随机森林算法模型。这种模型的预测准确率能够达到 98%;同时还能够预测训练数据没有包含的故障类型。通过在实际风力涡轮机数据的验证,表明了这种模型的稳健性。

参考文献:

[1] Kusiak A, Li W. The prediction and diagnosis of wind turbine faults [J]. *Renew. Energy*, 2011, 36 (1): 16-23.
 [2] Hyers R, McGowan J, Sullivan K, et al. Condition monitoring and

prognosis of utility scale wind turbines [J]. *Energy Mater*, 2006, (13): 187-203.

[3] Caselitz P, Giebhardt J. Rotor condition monitoring for improved operational safety of offshore wind energy converters [J]. *J. Solar Energy Eng*, 2005, 127 (2): 253-261.
 [4] Becker E, Posta P. Keeping the blades turning: Condition monitoring of wind turbine gears [J]. *Refocus*, 2006, 7 (2): 26-32.
 [5] Kusiak A, Verma A. A data-driven approach for monitoring blade pitch faults in wind turbines [J]. *IEEE Trans. Sustain. Energy*, 2011, 2 (1): 87-96.
 [6] Kusiak A, Verma A. Prediction of status patterns of wind turbines: A data-mining approach [J]. *Trans ASME J. Solar Energy Eng*, 2011, 133 (1): 1-10.
 [7] Tavner P J, Bussell G J W, Spinato F. Machine and converter reliabilities in wind turbines [A]. in *Proc. Inst. Elect. Eng. 2nd Int. Conf. Power Electronics, Machine & Drives [C]*, Dublin, 2006.
 [8] Kohavi R, John G H. Wrappers for feature subset selection [J]. *Artif. Intell*, 1997, 97 (1-2): 273-324.
 [9] Sbihi A. A best first search exact algorithm for the multiple-choice multidimensional knapsack Problem [J]. *J. Combinatorial Optimization*, 2007, 13: 337-351.
 [10] Kudo T, Matsumoto Y. A boosting algorithm for classification of semi-structured text [J]. in *Proc. EMNLP*, 2004, 301-308.
 [11] McCormick A C, Nandi A K. Classification of rotating machine condition using artificial neural networks [J]. *Proc. IMechE: Part C*, 1997, 11 (6): 439-450.

(上接第 1326 页)

端不需要开发客户程序,直接利用 IE 浏览器作为客户端。因此开放式网络化自动测试系统软件采用这种基于 COM 技术和 Web 技术的计算体系结构可以保证系统软件的可移植性、互操作性和可重用性,以及高度的可扩展性^[5]。

3 结论

本文采用 C/S 和 B/S 混合结构模式,建立了开放式的网络化自动测试系统硬件模型。以国际标准为基础,建立了开放式网络化自动测试系统的软件体系结构,从框架模型、功能模型、计算模型多个视角对软件体系结构进行了建模,达到了描述软件体系结构的目的,为开放式网络化自动测试系统的软件开发奠定了坚实的理论基础。

参考文献:

[1] 陈国顺, 宋新民. 网络化测控技术 [M]. 北京: 电子工业出版社, 2006.
 [2] 丁媛媛, 宋 杨. 基于 C/S 模式多层结构的通用计算机在线考试系统的实现 [J]. *微型机与应用*, 2011, 30 (14): 23-27.
 [3] 丁伟华. UML 在软件体系结构描述中的应用 [D]. 大连: 大连海事大学, 2007.
 [4] 郭 瑞, 姜玉海. 基于 COM 的分布式自动测试系统的设计与实现 [J]. *现代电子技术*, 2006, (22): 146-152.
 [5] 李文星. 船舶机电设备故障诊断方法研究 [J]. *计算机测量与控制*, 2013, 21 (3): 2094-2096.