文章编号：1672-3031(2017)06-0439-05

基于摆度大数据的水轮发电机故障预测方法研究

段炼达，刘晓波

（中国水利水电科学研究院 北京中水科水电科技开发有限公司，北京 100038）

摘要：水电厂积累的海量生产数据对发电机故障预测具有重要价值。本文采用蒙特卡洛方法对摆度大数据分析处理以进行故障预测。选取了某水电厂水轮发电机下导轴承摆度X峰峰值作为研究对象，在处理453 601组摆度数据时，选取3种典型工况进行研究。引入归一化加权平均值作为统计量，使用蒙特卡洛方法进行分析，发现该统计量服从正态分布。通过监测该统计量的实时值，依据正态分布3σ准则对水轮发电机工作状态做出定性评价，依据计算出的概率P值对水轮发电机工作状态定量评价，利用这些评价可以进行故障预测。

关键词：大数据；水电厂；机器学习；蒙特卡洛方法；故障预测；摆度

中图分类号：TP206.+3 **文献标识码**：A **doi**：10.13244/j.cnki.jiwhr.2017.06.005

1 研究背景

“大数据”一词由未来学家托夫勒(Alvin Toffler)在《第三次浪潮》中提出。尽管至今仍然没有“大数据”的准确定义，但维克托·迈尔-舍恩伯格(Viktor Mayer-Schönberger)和肯尼斯·克耶(Kenneth Cukier)在其合著的《大数据时代》中提出“大数据”具有4V特点：Volume(数据量大)、Velocity(输入和处理速度快)、Variety(数据多样性)、Value(价值密度低)，已得到普遍认可。

近年来我国水电机组单机容量不断增长，对水电厂安全生产提出了更高的要求。水电厂运行过程中产生了海量的监控数据。这些数据无论按照何种技术储存，都会以每年上百TB的速度增加，而随着时间的推移，数据的价值将发生变化[1]。水电厂历史数据库中包含着丰富的机组信息，可以为故障预测提供决策依据。尽管“大数据”的定义没有统一，但笔者认为电厂的海量数据就是“大数据”。采用什么方法对水电厂进行大数据分析、怎样进行分析，都是值得深入研究的问题。流行观点认为“大数据”是无法通过目前主流软件工具在合理时间内采集、存储、处理的数据集，但仍可以采用大数据与小数据结合的方法来处理问题[2]。AlphaGo在解决围棋问题中采用了蒙特卡洛方法与树搜索相结合的方法[3]，因此本文也采用蒙特卡洛方法分析研究水电厂发电机摆度大数据，用于发电机故障预测。

2 水电厂的大数据故障预测方法的探索

故障预测是在故障诊断的基础上发展起来的，目前故障预测的研究尚在发展之中。在故障诊断中，有很多处理监控数据的方法。比较传统的方法有机器学习方法、信号处理方法、多元统计分析方法等。[4]

机器学习类故障诊断方法基本思路是根据正常和故障状态下的数据作为训练集，利用神经网络或支持向量机[5]等算法来判别生产状态是否正常。由于在实际生产中缺乏故障状态的样本数据，造成该方法无法应用于水电厂监控数据的处理。信号处理类故障诊断方法是对监控数据进行谱分析或小波变换，提取与故障相关的时域或频域特征，从而为故障诊断提供依据[6]。该方法可用于早期故障的

收稿日期：2016-12-27

作者简介：段炼达(1991-)，男，河南南阳人，硕士生，主要从事水电厂自动化及计算机监控系统研究。
E-mail：duanlianda@qq.com

诊断、突变故障的检测，但是无法进行故障预测。多元统计分析类故障诊断方法根据历史数据利用多个变量之间的相关性进行故障诊断。基于该方法进行故障诊断算法简单，易于实现。但是，诊断出来的故障物理意义不明确，难以解释故障原因。[4]

蒙特卡洛方法是近年来逐渐引起关注的数理统计方法，加之以上故障诊断方法的启发，本文尝试使用蒙特卡洛方法对水电厂大数据进行统计分析，得到某项随机变化的指标对应的概率分布。从而依据该分布来定量评估生产状态，可以为故障预测提供依据。该方法的优点是：无需故障数据，而且可以根据某项指标定量评判设备的健康状态。

3　蒙特卡洛方法

蒙特卡洛方法的基本原理是使用随机抽样的方法从随机变量的概率分布中抽样，产生随机变量的样本值，确定统计量与随机变量的函数关系，重复进行模拟实验，由随机变量的样本值得到统计量的取值[7]；通过大量的统计量的取值便可以得出统计量的概率分布和数学特征[8]。

下面先将归一化加权平均值统计量[9]引入水电厂大数据处理中，然后进行蒙特卡洛模拟，最后根据模拟结果进行分析。

3.1　随机变量　同一工况下，以水轮发电机旋转周期为采样间隔采集摆度数据(摆度数据为下导摆度X峰峰值，以下简称摆度)，则摆度可视作是相互独立且具有相同分布规律的随机变量。记摆度为X，第i次采样得到的摆度值记作X_i，则根据中心极限定理大量摆度值的均值$\bar{x}$近似服从正态分布。可以根据大量的采样值估计出该正态分布的参数。本文以大量摆度值的均值作为随机变量，记作Y。

3.2　统计量　一种统计量只会从特定的角度揭示数据的部分特征[10]，因此需要更多性能优良的统计量。归一化加权平均值的加权值计算方法灵活、适用性强，而且可以在不提高硬件成本的前提下实现测量精度的提高[9]，因此采用归一化加权平均值作为本文使用的统计量。归一化加权平均值计算方法如下：

测量得到一组摆度数据，数据数量为$m\times n$，即X_i，$i\in[1, mn]$，归一化加权平均值统计量Z的计算公式如下：

$$z=\sum_{i=1}^{m}\bar{a}_i\cdot y_i \tag{1}$$

$$y_i=\frac{1}{n}\sum_{j=1}^{n}x_{j+(i-1)n} \tag{2}$$

其中：$\bar{a}_i$为y_i对应的归一化加权值；y_i为n个摆度数据的均值；m为计算z所需要的均值y_i的数量。具体算法如下：

(1)求得y_i的平均值$\bar{y}$；

(2)计算y_i相对于平均值$\bar{y}$的偏差量Δy_i，即：

$$\Delta y_i=y_i-\bar{y} \tag{3}$$

(3)将偏差量Δy_i代入权值函数$f(x)$得到$\Delta\bar{a}_i$，即：

$$\Delta\bar{a}_i=f(\Delta y_i) \tag{4}$$

(4)再对$\Delta\bar{a}_i$进行归一化处理得到加权值$\bar{a}_i$，即：

$$\bar{a}_i=\frac{\Delta\bar{a}_i}{\sum_{i=1}^{m}\Delta\bar{a}_i} \tag{5}$$

(5)由加权值得到加权平均值z，即：

$$z=\sum_{i=1}^{m}\bar{a}_i\cdot y_i \tag{6}$$

（6）权值函数 $f(x)$ 取为如下形式：

$$f(x)=\left|\frac{2}{1+\mathrm{e}^{-\frac{1}{x}}}-1\right| \tag{7}$$

按照以上步骤可以得到归一化加权平均值统计量，然后使用计算机进行蒙特卡洛模拟。

3.3　蒙特卡洛模拟　通过计算机按照分布规律进行随机抽样得到大量的摆度均值 y_i ，再通过3.2节计算得出对应统计量的值z，进而得出统计量Z的直方图和基于核密度估计的概率密度函数曲线，为评估机组状态提供依据。传统方法通过设置阈值简单判断水轮发电机的摆度数据是否越限，本文方法结合统计规律判断数据是否正常，是否更加合理。具体步骤详见实例分析。

4　实例分析

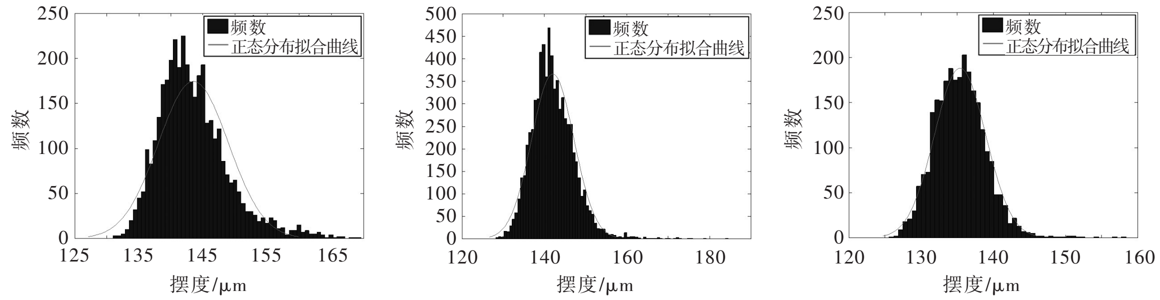
以并网条件下某水电厂的一台水轮发电机为实例，按照时间间隔为12 s取得453　601组摆度数据。为分析数据方便起见，对数据进行如下处理，选取3种工况数据进行分析研究：

工况1：水头在215 m ~ 216 m、有功功率在91%额定功率附近的3 840组数据；

工况2：水头在213 m ~ 215 m、有功功率在90.9%额定功率附近的7 101组数据；

工况3：水头在216 m ~ 217 m、有功功率在88.8%额定功率附近的2 661组数据。

使用MATLAB得到摆度频数直方图，如图1所示。



（a）工况1　（b）工况2　（c）工况3

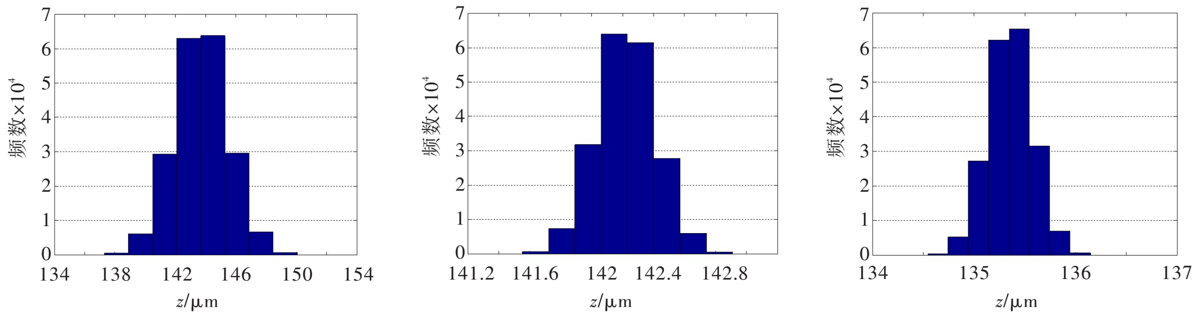
图1　摆度频数直方图

从图1可以看出3种工况摆度近似服从正态分布。下面将以工况1为例进行分析，其余工况分析步骤相同因此不再赘述。

记数据为 x_1、x_2、…、x_i、…、x_{3840} ，取n=100，有$y_i=\frac{1}{100}\sum_{j=1}^{100}x_{j+100(i-1)}$，因为时间相邻的数据时间间隔是旋转周期的整数倍，在正常运行状态下各个旋转周期的摆度相互独立，所以摆度数据相互独立，根据中心极限定理易得$Y\sim N(\mu,\sigma^2)$，使用MATLAB对Y进行jbtest测试发现服从正态分布。由样本值得到 y_1、y_2、…、y_{38}，将此组值分为两部分， y_1、y_2、…、y_{30} 作为训练组， y_{31}、 y_{32}、…、 y_{38} 作为测试组。

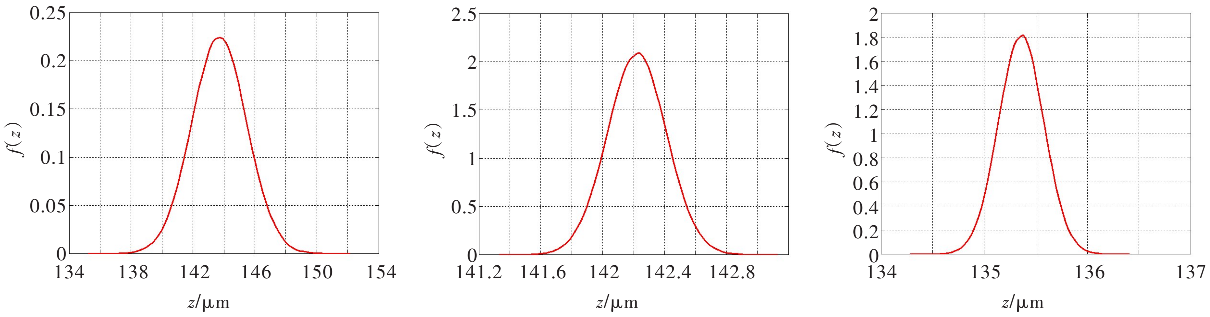
根据训练组的数据计算得到样本均值μ=143.7，样本方差为σ^2=11.9，进而得到$Y\sim N(143.7,11.9)$。根据Y的概率分布进行蒙特卡洛模拟，得到 y_1、y_2、…、$y_{1000000}$ ，按照3.2节中的算法计算归一化加权平均值统计量的值$z=\sum_{i=1}^{m}\bar{a}_i\cdot y_i$。m如果取值过大会导致数据需求量加大，计算时间变长，无法及时发现故障；如果过小则无法提高加权平均值的精度[7]，在兼顾实时性和精度的前提下取m=5。由此得到对应的 z_1、z_2、…、z_{200000} ，进而得到Z的直方图，如图2（a），其余工况如图2（b）（c）。

根据核密度估计得出其概率密度函数曲线，如图3。



(a)工况 1　(b)工况 2　(c)工况 3

图 2　统计量 Z 直方图



(a)工况 1　(b)工况 2　(c)工况 3

图 3　统计量 Z 概率密度函数图像

观察图 3 猜想该统计量近似服从正态分布，通过 jbtest 测试发现统计量 Z 服从正态分布。因此，可以参照正态分布的 3σ 准则制定评价标准。

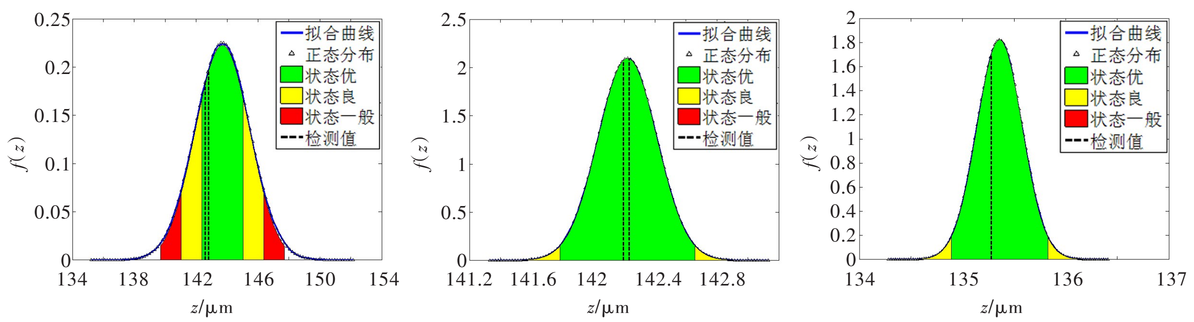
(1)当检测值处于$[\mu-\sigma,\ \mu+\sigma]$区间内时，判定水轮发电机状态为 A(状态优)；

(2)当检测值处于$[\mu-2\sigma,\ \mu-\sigma]\cup[\mu+\sigma,\ \mu+2\sigma]$区间内时，判定水轮发电机状态为 B(状态良)；

(3)当检测值处于$[\mu-3\sigma,\ \mu-2\sigma]\cup[\mu+2\sigma,\ \mu+3\sigma]$区间内时，判定水轮发电机状态为 C(状态一般)；

(4)其他区域时，判定水轮发电机状态为 D(状态异常)。

在此引入概率 P 值，P 值是一种在原假设为真的前提下出现观察样本以及更极端情况的概率，原假设为水轮发电机组状态正常。按照正态分布近似计算检测值对应的 P 值，以此作为定量评判的标准，P 值越小说明水轮发电机状态越不正常。计算得到工况 1 时进行两次测试得到的检测值对应的 P 值分别为 0.5710、0.6480，工况 2 时进行两次测试 P 值分别为 0.9527、0.8881，工况 3 时进行一次测试 P 值为 0.7026。根据测试组数据得到的 3 种工况统计量的检测值均在$[\mu-\sigma,\ \mu+\sigma]$区间内，如图 4 所示，即水轮发电机状态为 A。根据上述评价标准以及图 4，通过检测值可以判断发电机状态。



(a)工况 1　(b)工况 2　(c)工况 3

图 4　综合评价图

5　结论

本文采用蒙特卡洛方法，根据3种典型工况水轮发电机摆度大数据，得出归一化加权统计量的概率分布。根据归一化加权平均值统计量检测值处于的区间，可以定性判断发电机的工作状态。另外，根据该检测值的P值，可以定量评价发电机的工作状态。无需使用故障数据，通过以上的定性分析和定量计算即可对水轮发电机进行故障预测。另外，这种摆度大数据的处理方法可以应用到水电厂其他大数据的处理中，具有较大的参考价值。

参　考　文　献：

[1] 王德文，杨力平．智能电网大数据流式处理方法与状态监测异常检测[J]．电力系统自动化，2016，40(14)：122-128．

[2] 李国杰．对大数据的再认识[J]．大数据，2015(1)：1-9．

[3] SILVER D，HUANG A，MADDISON C J，et al．Mastering the game of Go with deep neural networks and tree search.[J]．Nature，2016，529：484-489．

[4] 周东华，胡艳艳．动态系统的故障诊断技术[J]．自动化学报，2009，35(6)：748-758．

[5] 周叶，唐澍，潘罗平，等．基于支持向量机的水电机组轴系运行故障诊断及预测研究[J]．水利学报，2013，44(S1)：111-115．

[6] 唐拥军，潘罗平，水电机组故障诊断系统信号预处理[J]．中国水利水电科学研究院学报，2005，3(3)：173-178．

[7] 康崇禄．蒙特卡罗方法理论和应用[M]．北京：科学出版社，2015．

[8] 杜林，袁蕾，熊浩，等．应用蒙特卡罗模拟法的电力变压器热特性绝缘寿命评估[J]．高电压技术，2010，36(4)：856-861．

[9] 龙宏波，叶晓慧，谭思炜．归一化加权平均值算法在测量中的应用[J]．电光与控制，2010，17(12)：68-70．

[10] 胡俭勇，苏锦海．一种随机性实时检测方案[J]．计算机工程，2009，35(9)：136-138．

Study on failure prediction for hydro-generator based on big data of run-out value

DUAN Lianda，LIU Xiaobo

(*China Institute of Water Resources and Hydropower Research*，*Beijing*　100038，*China*)

Abstract： The hydropower plants have accumulated mass production data，which is of great value to failure prediction. In this paper，the Monte Carlo method is applied to analyze the big data of run-out value to predict the failure. The X peak-peak values of the run-out of the lower guide bearing of a hydropower plant are taken as the investigation object. In the treatment of 453，601 sets of the run-out value，three typical operating conditions are selected to study. The normalized weighted average is used as the statistic，and the Monte Carlo method is adopted to analyze it. It is found that the statistic obeys normal distribution. By monitoring the real-time value of the statistic，the working status of the hydro-generator is evaluated qualitatively according to the normal distribution 3σ criterion，and evaluated quantitatively according to the calculated probability P-value. Failure can be predicted by using these evaluations.

Keywords： big data；hydropower plant；machine learning；Monte Carlo method；failure prediction；run-out value

（责任编辑：杨　虹）