**基于SCADA数据挖掘的风电场能效评估**

**汤蔚**

**tangwei03@mywind.com.cn1**

(1. 深圳量云能源网络科技有限公司，广东省 深圳市518000)

**摘要：**风电机组是利用自然风能将气流的动能转化为机械能，并连接和带动发电机运转用来发电的一种发电设备，其过程是一种多种能量并行转换运行的复杂状态。本文提出以风机的正常数据采集与监视控制数据（SCADA）作为数据输入建立深度自编码算法模型（DAE）进行预训练，再通过训练后模型对无标签数据进行能效计算并提出一种基于极值理论的自适应阈值算法作为风机底层部件能效劣化诊断基准。通过对风电机组底层部件能效劣化的有效定位并结合故障诊断树及时下发工单给予最优维修措施，再利用底层部件能效值结合机组实际工况建立随机森林模型（Random Forest)进行动态指标权重计算，并借由工程经验建立从部件级、风机级、风场级层次化能效评估体系。同时通过对风电机组每个部件的能效机理研究，借由风电机组部件级能效评估确定风能利用效率、机械能效率以及电能效率等三大能量系统对风电机组出力的影响。本文以明阳智能某风场双馈机型MY1.5MW机组为案例分析对象，运用理论研究与风电机组实际工况进行对比验证了其方案的可行性。

**关键字：**能效评估；风电；自适应阈值；随机森林（Random Forest)；数据采集与监视控制数据（SCADA）；深度自编码算法（DAE）

**文献标志码：** **中国分类号：**

**Effciency Evaluation of Wind Farm Based on SCADA Data Mining**

**Wei Tang**

**tangwei03@mywind.com.cn1**

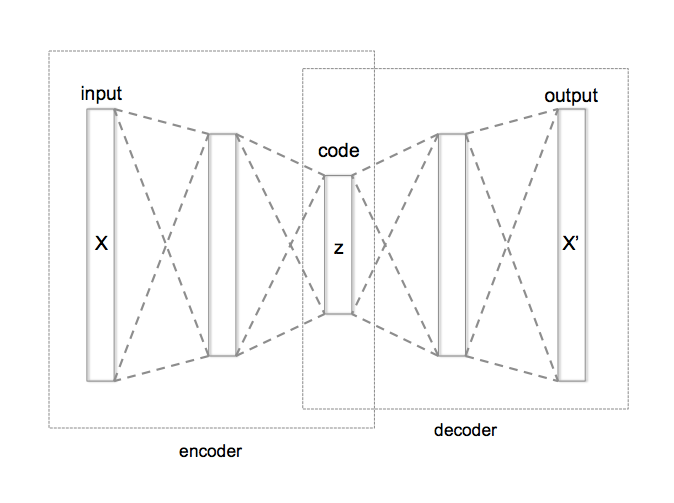
(1. Shenzhen Quant-Cloud Energy Network Technology Co., Ltd., Shenzhen, 518000,China)

**Abstract:** Wind turbine is a device that converts the kinetic energy of wind into electrical energy, connect to and drive the generator producing electricity, which of the process is a complicate state of parallel conversion and operation of multiple kinds of energy. This paper proposed to use normal supervisory control and data acquisition data(SCADA) as data input, pre-training deep autoencoder algorithm model, then, the trained model could be used to eastablish new unlabeled SCADA data and an adaptive threshold based on extreme value theory was presented as the criterion of energy effciency deterioration diagnosis of wind turbine bottom components. By effectively locating the energy efficiency deterioration of the bottom components, and combining with the fault diagnosis tree to issue orders for giving optimal maintence measures. Besides, combining the erengy efficiency condition of bottom components with the actual working conditions of wind turbine, the daynamic feature weights will be established by random forest, and with the help of engineering experiemce, a hierarchical energy efficiency evalustion system was built which composed of component level, wind turbine level and wind farm level. At hte same time, based on the research on the erengy efficiency mechanism of component of wind turbine, the impact of wind energy efficiency, mechanical erengy efficiency and electrical energy efficiency on the output of wind turbines would be judged by wind turbine component level energy efficiency condition. In case anlysis, one wind farm containing Mingyang MY1.5MW wind turbines is used to verify the feasibility by comparing the theoretical research with hte actual working conditions of wind turbine.

**Keywords** efficiency evaluation ; wind power; wind farm; adaptive threshold; Random Forest;supervisory control and data acquisition data(SCADA); deep autoencoder(DAE)

## 0 引言

近年来随着全球气候环境恶劣趋势日益增加，碳排放对生态环境所产生的不良影响已越来越走进全球的视线， 尤其是国家于2021年提出2060 零碳行动倡议， 力争在2030年之前实现碳达峰，在2060年前实现碳中和。众所周知，电力行业的排碳量是非常高的，为了实现电力减碳、能源减碳、实现双碳目标，扩大已风电为代表的非化石能源的消纳比例成为了一项重要措施。而随着风电产业高速发展的同时[1]，行业也把焦点从风电装机容量逐渐转移到风电场运行质量以减少风电机组运维成本。而目前风电场由于受到其运行环境的影响具有不稳定性特点导致了风电行业的高成本运营[2]。为了降低成本本文主要集中于提高风电场的能效水平，通过对风电场的能效性能进行深入研究从而实现对风机部件级、整机级、风场级能效状态精准评估和定位，及时给出能效劣化的优化措施以及维修工单，从而达到风电场从宏观到微观故障提前预防以及维护效果。

针对风电场能效评估，本文主要包含风机底层部件指标能效评估以及诊断，部件级-整机级-风场级能效评估两大部分。其中风机底层部件指标能效评估以及诊断方法主要步骤包括：1）根据专家经验从SCADA数据提取风机各部件有效特征指标；2）根据SCADA故障信息对数据进行人工标注标签爱从而筛选出正常SCADA数据；3）使用正常SCADA数据建立深度学习神经网络模型进行预训练，在输入无标签SCADA数据进行能效计算；4）使用基于极值理论的动态阈值诊断底层部件能效水平。然后根据能效劣化指标根据故障树分析法找到有效优化或维修措施并将预警下发给现场，从而实现能效优化的措施落地目的。而部件级-整机级-风场级能效评估过程可概括为：1）通过对底层部件指标的能效评估，再借由机器学习随机森林（Random Forest)等方法分析得到各项能量载体各项能量权重分布，从各分项能量损失机理计算得到机组各部件的能量转化效率值；2）基于部件级能效评估分析研究风电机组各系统能效关联结构从而构建数学机理计算得到风机整机能效；3）利用能效框图分析法建立风机-风场关联模型从而进行风电场能效评估。

## 1 风机底层部件指标能效评估&诊断

### 1.1 深度自编码器模型构建

深度自编码器算法（DAE）[3]是一种基于神经网络的无监督机器学习降维算法，其隐藏层主要包含encoder（编码器）、bottleneck(瓶颈层）和decoder(解码器）两部分，编码器主要将高维度输入层数据集压缩，而解码器则用于重建输入数据。通过神经网络学习最有信息量的特征从而转化为低维的隐变量，编码器会选择输入层最有用的信息特征量压缩至瓶颈层，而后解码器再将隐变量近似还原为初始输入数据集 从而达到降噪效果。所以隐藏层的维度必须比输入层或输出层的维度少，其神经网络结构示意图如图一所示。

通过上述部分可了解到DAE神经网络共包含三部分，而合理的选择神经网络层数以及隐藏层神经元个数，会影响训练模型的最终损失函数。对于输入层的数据集，如果数据集能够通过线性可分函数或决策分开即只无需隐藏层，而当输入数据集非线性分离时则需更多层隐藏层，理论而言，层数越深模型拟合函数能力越强效果会越好，但模型也可能会导致过拟合，同时增加模型训练的时长。同时神经网络包含过多的神经元节点时，输入数据集不足以训练隐藏层中所有的节点由此导致过拟合，而当训练数据足够，隐藏层过神经元过多也会增加训练时间。因此一般而言，隐藏神经元个数应保持在输入层与输出层大小之间，后通过不断试验找到最佳数量。

图1：DAE算法结构

Figure 1：Structure of DAE Alogrithm

从输入数据集**X**映射到隐藏层**z**满足：

(1)

其中 表示激活函数，**W** 表示权重矩阵，**b** 表示偏差向量；

从隐藏层映射到输出层满足：

(2)

DAE会基于反向传播算法和最优化收敛方法计算最小重构误差，即最小损失值即为MSE误差：

(3)

其中DAE算法得到的损失值即可定义为用于训练的SCADA数据特征指标的损失值 [4] ，对于风机SCADA数据，异常数据与正常数据的比例本质上分布极其不均衡，且需人工定义正常与异常标签。所以在无监督的情况下，可以选取 无故障风电场机组数据训练DAE神经网络 ，然后训练好的模型再可以很好还原正常样本，一旦样本出现异常，模型会无法很好还原异常样本，从而导致损失值较大。所以本质上基于DAE算法的能效异常诊断是一种使用半监督学习的评估方法。

### 1.2 基于自适应动态阈值诊断方法

在能效异常诊断种，利用正常数据训练好的DAE模型预测输入SCADA数据的原始值，并利用预测值于实际值的误差作为能效损失值，即能效值Re可定义为:

由于输入SCADA数据为无标签数据，通过计算自适应阈值即可判断底层部件指标能效是否异常，指标正常运行时，SCADA数据就在阈值范围内；一旦出现异常，其指标能效就会超过阈值[5] 。其定义规则如下所示：

(4) (4)

其中为阈值； 从分析来看，风机故障分为可预测故障以及突发性故障由于突发性故障具备不可预测性，所以本文不做详细分析。而可预测性故障可定义为在故障发生之前风机部件就表现出一定的异常状态，此部分通过阈值计算即可找到部件的异常状态。由于风机各部件的工作效果可随着时间、环境时刻变化，如果阈值是人为设定的，需要用户有足够的使用经验，而且这种设定方式随机性很大。所以基于极值理论的自适应阈值被提出作为判断子部件是否处于异常状态的依据，此方法需要手动设定概率值q和漂移窗口数据量p [6] 。

提取某一子部件一段时间的正常数据（约半年时间的数据量）作为先验数据集，通过应用极值理论来分析极值Re的分布用来设计异常监测和风机零部件的条件评估，假设样本向量分布为F分布，通常表示每十分钟某一参数的数值。对于这些样本向量，其分布尾部符合统计公式：

(5)

其中是每十分钟Re的最大值，对于一组，其概率分布公式如下：

(6)

根据分布可知，随着, ,为了避免极限到一个值，假设存在,使

(7)

其中为参数，为广义极值分布函数;

当使用SCADA数据计算Re时，Re的平均值并没有很大波动，所以Re的平均值可以近似为常数，但是方差却波动很大，因此此部分需要根据公式计算出来，其阈值Re公式定义为：

(8)

其中由于F分布未知，通过最大似然估计计算，p为置信限制，即为最初需手动设置参数概率p; 以上述计算为基础，输入要检测的数据集手动输入参数p提取数据集一段时间内的初始数据个数p作为初始平均值,继而得到初始阈值Re，循环往后流动一个数据计算滑动平均值得到流动阈值，即可得到自适应阈值，而超出阈值即可定义为能效异常状态。

## 2 层次化能效评估系统

### 2.1 部件级能效评估模型

根据式（3)损失函数可知风机底层部件指标损失能效，基于随机森林权重分配可定义风机部件的主要特征，考虑各项底层部件指标之间的强耦合性，从SCADA多维指标中解析出主要影响因素，从而将交错复杂的问题简化为数学模型。本文采用基于选取的底层部件指标使用随机森林提取特征权重，随机森林一般使用基尼值作为切分节点的标准，赋予小类较大的权重，大类较小的权重。通过基尼值来进行评价，可将指标的重要性评分用score表示，假设风机部件有m个特征,要计算每个特征的基尼指数评分,即第i个特征在随机森林所有决策树中节点分裂不纯度的平均该变量，则基尼指数(GI)的计算公式如式所示：

(9)

其中,k表示有k个类别，表示节点m中类别k所占的比例；

假设随机森林有n棵树,则特征权重可定义为：

(10)

(10)

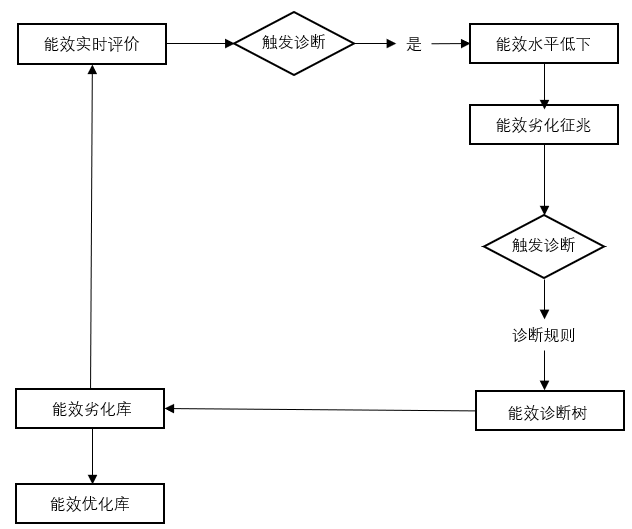
基于风机部件的底层部件指标权重，再通过加权求和方式计算得到部件级能效值。在实际情况中，由于风机部件级的故障记录一直处于动态变化状态，通过随机森林计算出的实际权重是以故障记录为数据基础进行动态计算，所以各子部件特征权重也会随之变化。 能效诊断的目的是分析设备运行情况，找到能效劣化的底层原因，定位劣化部件，确定各影响因素的关联性，找到主要因素，指导进行故障处理和消缺,其结构图如二所示。

图2:能效诊断结构图

Figure2:Structure of Energy Efficiency Diagnosis

能效诊断模块将针对能效评估模块的评估结果做进一步数据分析诊断，对机组的主要大部件建立多个能效诊断模型，对相关的关联参数的大量运行数据进行分析，从而得到细分到子部件具体运行参数的能效值，找到影响整机能效水平的关键原因。

### 2.2 风机-风场级能效评估模型

风电机组主要利用自然风将风能转化为电能，风能被叶轮系统转换为机械能 ，并经由主轴-齿轮箱-发电机系统传递至发电机-变流器发电系统，由此可知，风电机组同时存在风能、机械能、电能的能量转换过程。通过将风机组成系统拆分成多个串联系统即为风能捕获系统、机械能传递系统和电能转换系统。 风能捕获系统能效分析是对叶轮吸收风能水平的评估。风能捕获效率用风能转换系数 为主表征指标，并综合考虑该系统内设备（叶片、变桨系统、偏航系统、风速仪、风向标等）关联参数，对风能捕获系统的能效进行评价。的计算可采用经验函数或基于历史数据拟合函数，函数的形式应为：

(11)

其中，为桨距角， 为叶尖速比，V是自由来流风速，R为叶片半径；

而风力机械传动链则由主轴传动轴、齿轮箱、发电机等构建。通过将每个部件看作一个系统模块，低速轴轴通过主轴传动柔性转至高速轴，从而建立串联关联系统[7] 。根据系统结构可知三个独立系统可看作为串联关系，由数学公式可知机械能传递效率为

同理可知，建立发电机-变流器两个模块串联系统可基于其监测参数水平计算得到电能转换系统能效效率，其关联参数应至少包括发电机、变流器等部件的监测参数，从而可知电能转换效率为：

(13)

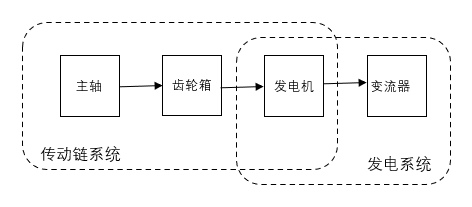
通过上述三个模块的计算，通过能效框图可知，如果系统的一个模块发生故障，则整个系统故障，此为串联关系。应用到风机领域，上述三大子系统模块可近似看作串联成一个风机系统完成从风能到电能的传递过程，一旦任意模块发生故障，则整个能量传递过程也会失败。其结构示意图如图三所示

图3:子系统能效关联结构

Figure3: Subsystem Structure of Energy Efficiency Correlation

而风机整机系统的一系列的能效效率即可表达为：

(14)

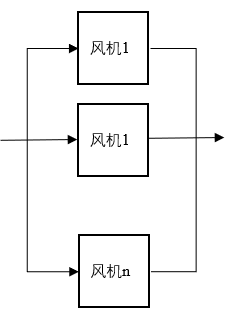
并行框图则可应用于冗余系统，并行系统一般指所有的子系统模块并行运行，如果并行系统失效则需要所有子系统失效才会发生。而在风电场应用中，会发现单个风机失效并不会影响风电场的全部能效为零，即风电场的能效效率并行框图如图四所示：

图4:风场能效关联结构

Figure4:Wind Farm Structure of Energy Efficiency Correlation

在风电场运行过程中，只要有至少一台风机还在运行，其子系统就会是风电场具有冗余特性，所以风电场的能效效率可表达为：

(15)

其中n为风电场包含的风机数量；

借由上述数学计算模型即可建立底层部件指标与部件之间的能效关联，部件与整机之间的能效关联，和整机与风场之间的能效关联 [8], 从而建立其整个层次化能效评估系统。

## 3 实例验证

### 3.1 风机底层部件指标选定

为了测试理论的可行性，实验数据主要提取3个风电场明阳双馈机械MY1.5MW共250台风机约1年的SCADA数据。其中风机部件选取了共5大部件（主轴、发电机、齿轮箱、变流器、偏航系统）进行能效评估，其中主轴、发电机、齿轮箱、偏航系统根据专家经验主要选取了温度、速度、角度等底层部件指标进行分析，变流器则主要选取电流、电压等指标进行分析。

### 3.2 风机底层部件指标能效评估&诊断结果

以明阳双馈机型为例，现统计出明阳SCADA数据一共包含约270个特征标签，然而实际情况每个部件的故障特征不会包含所有SCADA标签，遂需进行特征工程对每个部件进行特征降维，直至找到只与每个部件相关的故障特征。以发电机为例，可先通过专家经验筛选出与发电机相关联的约6个特征标签具体如表一所示。

表1：发电机底层部件能效指标

|  |  |
| --- | --- |
| 特征参数 | 编码 |
| 发电机驱动端轴承温度（℃） |  |
| 发电机非驱动端轴承温度（℃） |  |
| 发电机定子U绕组温度（℃） |  |
| 发电机定子V绕组温度（℃） |  |
| 发电机定子W绕组温度（℃） |  |
| 发电机冷风温度（℃） |  |

提取发电机一种故障数据（例如发电机定子U绕组温度异常），加入正常数据数据总量约2万条，其正常数据：异常数据比例约为90：1，后通过XGBOOST分类器对筛选出的SCADA数据进行分类，得到分类混淆矩阵其中由于正负样本不均衡造成数据准确率约100%，召回率约97%，精准率约98%。

### 3.3 部件级能效分析

根据上述10007003风机的2天SCADA数据进行验证，其能效诊断全部结果包含5大部件，19个子部件的能效值以及能效状态，而以其中一个子部件的能效动态阈值诊断为例其效果如图所示；

图5：能效诊断

Figure5:Diagnosis of Energy Efficiency

其中图五横坐标表示时间序列，纵坐标表示子部件可靠度；黄色线条表示流动阈值，青色线条表示子部件实际值，红色散点表示子部件在这一时刻处于异常或是故障状态；

如图所示，可观察到发电机驱动端轴承温度在测试数据集约第800个时间序列状态发生异常，对比实际故障记录结果，发现此时间段发电机驱动端轴承温度偏高，由此可得能效阈值诊断结果是准确的[9] 。 通过上述图二子部件阈值诊断经过一定计算可转化为表三形式，表三观察可得到发电机部件每日能效值以及能效等级,其10007003机位号风机在2021-09-02号能效等级由良好降为一般，原因正是由于发电机驱动端轴承温度当日发生异常。

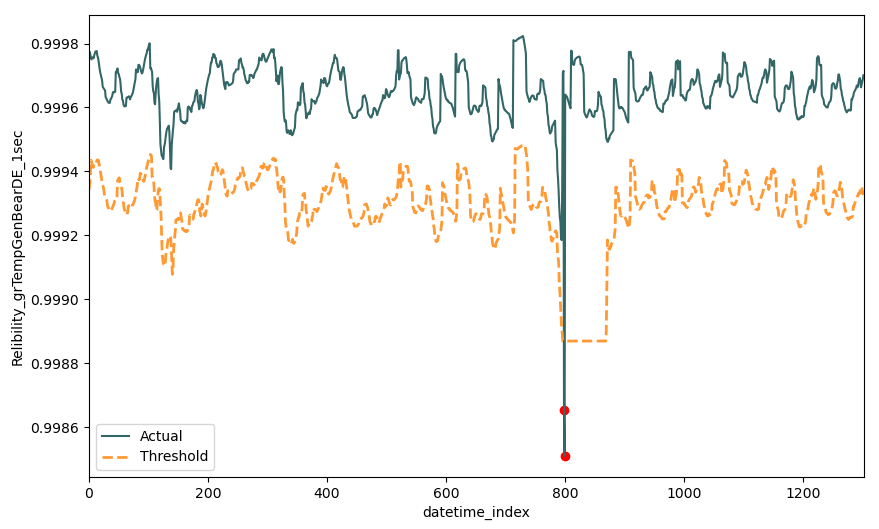
### 3.4 风机级-风场级能效分析

通过上述图五子部件阈值诊断经过一定计算可转化为表二形式，表二观察可得到发电机部件每日能效值以及能效等级，其10007003机位号风机在2021-09-02号能效等级由良好降为一般，原因正是由于发电机驱动端轴承温度当日发生异常。

表2：发电机部件能效评估

|  |  |  |  |
| --- | --- | --- | --- |
| 风机 | 时间 | 发电机能效值 | 能效 等级 |
| 10007002 | 2021-09-01 | 0.99989 | 良好 |
| 10007002 | 2021-09-02 | 0.99982 | 良好 |
| 10007002 | 2021-09-03 | 0.99983 | 良好 |
| 10007003 | 2021-09-01 | 0.99986 | 良好 |
| 10007003 | 2021-09-02 | 0.99967 | 一般 |
| 10007003 | 2021-09-03 | 0.99984 | 良好 |

## 4 结论

本文提出了基于半监督学习的DAE模型计算风电机组底层部件能效评估，再基于动态阈值进行能效诊断。再根据底层部件能效建立层次化能效评估系统。基于实际案例的分析可得出如下结论： 1）结合DAE模型和自适应阈值，能够预测底层部件异常状态，根据异常种类关联到相关维修手段，可用于风电机组能效异常监测和诊断； 2）基于动态误差分析方法，自适应阈值可一定程度减少因风速波动和外界干扰对风电机组的监测影响而造成的失误预警； 3）建立层次化能效评估系统可监测大至一整个风电场的能效水平，小至风电场某一风机的部件能效水平。从而达到对风电场一整个生命周期的全面监测效果。

**参考文献**

[1] YE X, QIAO Y, LU Z. Cascading tripping out of numerous wind turbines in China: Fault Evolution analysis and simulation study[C][J]. Power and Energy Society General Meeting,

2012 : 1-11.

[2] 王勇, 刘金宁, 曹曼. 基于概率论与证据理论的风电场电能评估方法研究[J][J]. 华北电力大学学报, 2012, 39(3) : 65 –70.

[3] VINCENT P, LAROCHELLE H, LAJOIE H. Stacked denoising autoencoders：learning useful representations in a deep network

with a local denoising criterion[J]．[J]. Journal of Machine Learning Research, 2010,11:3371-3408.

[4] YANG W, COURT R, JIANG J. Wind

turbine condition monitoring by the approch of SCADA data analysis[J][J]. Renewable Energy, 2013, 53 : 365 –376.

[5] LI M, WANG S. Dynamic fault monitoring of pitch system in wind turbines using selective ensemble small-world neural networks[[J]. Energies, 2019, 12 : 1 – 20.

[6] BEIRLANT J, GOEGEBEUR Y, SEGERS J, et al. Statistics of extremes: theory and applications[J]. John Wiley and Sons,2006.

[7] N S, S C, R J. Operations Management[J]. Prentice Hall, 2010, Sixth edition.

[8] NGAI E, HU Y, WONG Y, et al. The application of data mining techniques in

nancial fraud detection: A classication framework and anacademic review of literature[J]. Decision Support Systems, 2011, 50(3) : 559 – 569.

[9] SADIK S, GRUENWALD L. Research issues in outlier detection for data streams[J]. ACM SIGKDD Explorations Newsletter, 2014, 15(1) : 33 –40.