

「电网」

凝聚层次聚类算法在电厂设备故障预测上的研究

文/北京工商大学 刘翠玲 李天瑞 于重重

针对电厂设备产生的数据浪费问题，并为了保证电网的健康运行，提出了利用设备的时间序列数据并运用凝聚层次聚类算法对电厂设备故障进行预测的方法。首先提出了预测方法的框架，后运用凝聚层次聚类算法对时间序列数据聚类，然后采用多项式拟合方法进行了Matlab曲线拟合仿真，根据动态时间弯曲方法度量运行曲线和拟合曲线的相似性，最后通过相似性和预设的阈值判定设备运行状态。结果表明该方法能充分利用时间序列数据，有效地进行复杂电力设备故障预测，有利于及时调度和处理有关事故或缺陷，提高企业经济效益和管理水平。

电网的运行需要许多大型设备维持，包括变压器、发电机和励磁系统等。这些设备会随着监测时间的增长，成为大规模海量时间序列数据集。从海量数据中进行分析 and 筛选可以挖掘出数据中有价值的隐藏信息，并且作为电厂中的核心设备，一旦发生故障，将会影响电厂的正常运行，并可能会造成极大损失。为避免故障的产生，采用合适的数据挖掘方法和故障诊断方法对大型设备的正常运行及电网的经济效益都具有重大意义。

目前故障预测方法存在多种分类方式，大致将预测方法分为3类，即基于模型的故障预测、基于信号处理的故障预测与人工智能的故障预测。

第一类是基于模型的预测方法，需要较为精确的待诊断对象的解析模型，能够解决知识获取瓶颈问题，具有较高的诊断精度。包括灰色模型方法、卡尔曼滤波器等。不过该方法通常适用于线性系统，而复杂装备往往是非线性系统，以至于难以建立精确的数学模型，所以该方法会受到很大限制。

第二类是基于信号处理的预测方法。此方

法在故障领域应用较早，不需要建立对象的精确模型。通常利用基于小波变换、经验模态分解法等故障诊断方法对信号模型进行分析和故障监测。

第三类为基于人工智能的故障预测方法。包括专家系统、人工神经网络等，此外近几年也出现了数据挖掘、粗糙集理论和贝叶斯网络等技术以及上述方法的综合应用。此类别是通过各种数据找出的隐含信息并与所获的数据分析基础上预测，可以避免基于模型故障预测方法的不足，从而更实用。

通过利用电力设备产生的时间序列数据找出其中隐含的信息，并对这些数据进行分析。于是提出了一种故障预测方法，是用层次聚类算法对数据进行聚类，Ward方法计算簇之间的邻近度，进行了多项式拟合并用Matlab仿真得出拟合曲线，再用动态时间弯曲处理数据之间的相似性，最后根据相似性测定的距离预设的阈值判定设备运行状态。通过这样的数据挖掘与分析可以避免基于模型故障预测方法的不足，能够更有效地使用实时历史数据，使得数据创造出其价值。

故障预测总体框架

针对时间序列数据，采用凝聚的层次聚类算法设计的设备故障预测功能的整体技术框架如图1所示。

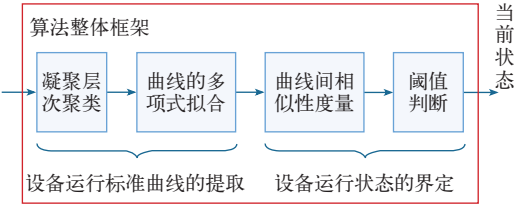


图1 设备故障预测方法整体技术框架图

图1中该算法主要由2部分组成，分别为标准曲线的聚类提取和设备运行状态的界定。运用层次聚类算法对时间序列数据聚类，后对聚类的数据进行多项式的拟合，然后再根据曲线间相似性度量的结果与预设故障的阈值之间的关系，来判断设备运行是否会发生故障。

故障预测方法

1. 凝聚层次聚类

一般来说，时间序列聚类技术其主要目的是将待聚类的时间序列数据对象根据其相似性分类为多个类别，且保证不同分类类别中的对象序列之间的相似性尽可能小，而同种分类类别中的对象序列数据之间应该存在有尽可能高的相似关系。

凝聚层次聚类在许多领域都被证明是有效的问题解决方法。凝聚层次聚类，其主要是从每个数据对象（或样本）作为个体簇开始，然后在每个动作中合并最为邻近的两个簇，直到当所有的簇个体均合并成一个簇为止，具体的聚类结果常

常使用树状图进行展示，关于凝聚层次聚类的算法流程如表1所示。

表1 基本的时间序列凝聚层次聚类算法

选择合适的簇合并邻近度更新函数	
过程	1) 选择合适的相似性距离度量公式
	2) 计算初始邻近性距离矩阵
	3) 重复
	4) 使用选择的簇合并函数，合并最为相近的2个簇
	5) 更新邻近性距离矩阵，用于反映新的簇与合并前的簇之间的邻近性
	6) 直至只剩下1个簇
	7) 结束

2. 凝聚层次曲线多项拟合函数选取

由上述聚类算法特点得知，在进行时间序列凝聚层次聚类挖掘之前，必须要结合对象序列数据的特点完成邻近性距离度量函数与簇合并标准函数的选择工作。

动态时间弯曲（Dynamic Time Wrapping，DTW）与其他算法相比对于时间序列类型数据间的相似性描述具有较高的准确性，其定义如下。

设存在两条时间序列数据设存在两条时间序列数据 $Q = \{q_1, q_2, q_3, \dots, q_n\}$ 与 $S = \{s_1, s_2, \dots, s_m\}$ ，其长度分别为 n 与 m ，则存在两序列构成的大小为 nm 的矩阵 M ，其中矩阵元素 (i, j) 对应的是序列 Q 与 S 中的元素 Q_i 和 S_j 之间的基准距离，即 $(i, j) = D_{base}(q_i, s_j)$ ，其中基准距离函数 D_{base} 通常选用均方欧式距离。在此基础上，序列 Q 与 S 的元素之间距离之和的最小值的可以使用弯曲路径 $W = \{w_1, w_2, w_3, \dots, w_k, \dots, w_L\}$ 进行表达，其中 $\text{Max}(m, n) \leq L \leq m + n - 1$ ； $w_k = (i, j) k$ 。该路径的选择一般是采用动态规划方法从起点 $(1, 1)$ 出发到终点 (m, n) 的最小化累积距离路径，该路径一般被认为是最优的，则用公式

$$DTW(i, j) = \begin{cases} D_{base}(q_i, s_j) + \text{Min} \begin{cases} DTW(i, j-1) \\ DTW(i-1, j) \\ DTW(i-1, j-1) \end{cases} & i, j \geq 1 \\ 0 & i = j = 0 \\ \infty & i = 0, j > 0 \text{ or } i > 0, j = 0 \end{cases} \quad (1)$$

表述如下。

同时考虑为了时刻保证生成的新簇簇内平均误差增量最小，选择了Ward方法作为簇合并所使用的标准函数。

Ward离差平方和法是基本凝聚层次聚类方法中的一种，其基本策略为：从每个对象作为一个簇开始，逐步合并距离最近的两个簇，循环直到所有的对象都在一个簇中，或者满足预先设定的

终结条件。Ward方法将簇间的距离定义为两个簇合并时导致的平方误差和，其公式如下。

$$D(C_i, C_j) = \sum_{k=1}^g \sum_{i=1}^{n_k} |z_{ik} - \bar{z}_k|^2 \quad (2)$$

式中， g 为簇的个数； $\bar{z}_k$ 为第 k 个簇的平均值； n_k 为第 k 个簇中的对象个数。

由于，这种方法中邻近性距离矩阵的更新是随着聚类而进行的，同时在更新过程中只涉及更新前的邻近性距离矩阵中的值，而不再使用原始对象数据实值。这一特点，在实际应用中将会节省大量的内存空间。

3. 曲线拟合与设备状态的界定

对于所采集到的原始时间序列数据来说，如何通过已聚类序列数据完成相应的标准曲线提取工作已经成为目前辨别当前运行设备是否处于故障状态的主要手段。在一般的科学研究课题中，

对于上述问题一般归结于曲线拟合问题。为解决上述问题，选择多项式拟合方法来达到上述目的。

多项式拟合是指在函数关系 $y=f(x)$ 不知道的情况下，可以根据 n 对原始数据找出一个含有 $(k+1)$ 个参数的 k 阶多项式来逼近 $(k+1 < n)$ ，即 $y=f(x)=a_0+a_1x+a_2x^2+\cdots+a_kx^k$ 。一般对于同一组数据，用高一阶或低一阶的多项式拟合曲线，往往较难辨别优劣。一个拟合的多项式并不是阶数越高越好。一般总是先从低阶开始拟合，然后根据判别决定是否需要提高阶数进行拟合。

故障预测方法

1. 凝聚层次聚类结果

运用上述所定义的凝聚层次聚类算法，通过对设备采集所获得到的原始监测数据集中，进行的聚类分析结果如图2所示。

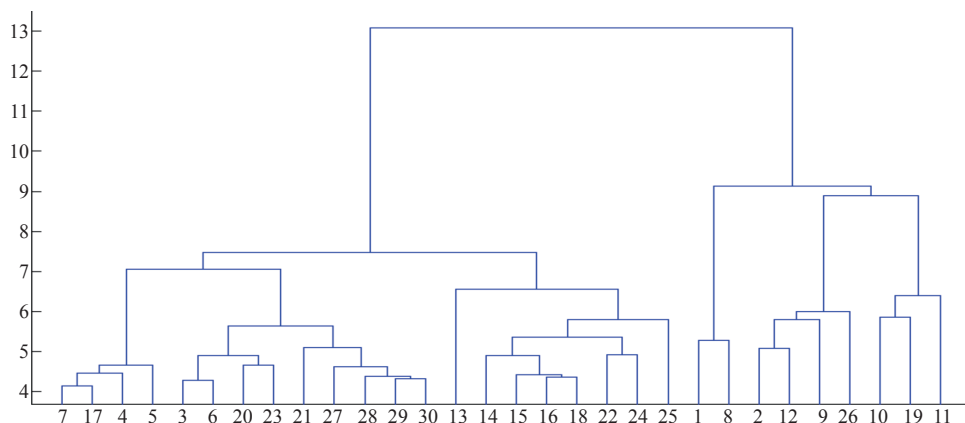


图2 凝聚层次聚类谱系图

在发电过程中，由于负载的不同，为了更好地监测相关设备的运行情况，结合上述的凝聚层次聚类谱系图将整体分为8类，其结果如表2所示。

表2 聚为8类的凝聚层次聚类结果

类别	组成
类1	4、5、7、17
类2	3、6、20、23
类3	21、27、28、29、30
类4	13
类5	14、15、16、18、22、24、25
类6	1、8
类7	2、9、12、26
类8	10、11、19

表2的聚类结果表示，设备运行曲线大致可分为8类，经详细的曲线对比发现每一种类别可以归结为一种相应的运行模式，同时发现若将每一类聚类结果进行整理则可以完成该类别标准曲线的提取，从而达到故障检测的目的。

2. 曲线拟合结果

根据前述聚类的结果，以类1为例，即（4、5、7、17）所涉及到的原始序列数据通过Matlab仿真软件进行曲线拟合所得到的结果如图3所示。

为了达到自动判别当前运行的设备是否处于正常的工作状态，DTW技术完成选定标准曲线（预测值曲线）与当前运行曲线（实际值曲线）间的相似性关系的度量工作。若当前设备运行曲

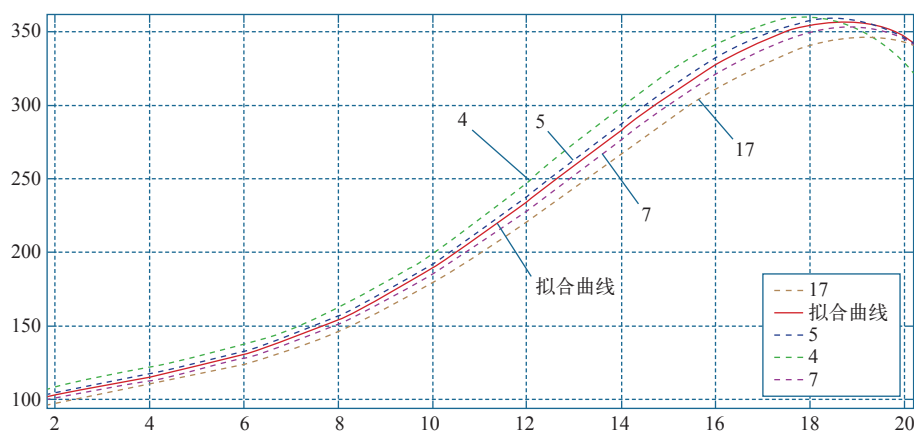


图3 曲线拟合图

线与由聚类结果获得的当前日期标准设备工作曲线间的相似性距离大于预设阈值,则认定为该设备处于异常工作状态;反之认定该设备处于正常工作状态。同时,为了保障由聚类结果所生成的标准设备工作曲线具有有效性及准确性,需要在设定时间间隔后,对其重新完成标准曲线生成工作。

3. 故障预测结果

该方法在发电集团生产信息系统应用的故障

预测结果如图4所示。

图4中红色为预测值上限曲线,绿色为预测值下限曲线,蓝色为实际值曲线。通过查看设备的实时曲线,进行趋势分析,能找到设备偏离正常运行状态的距离。从各个参数的偏离情况,提供运行人员的操作方向,使设备回到正常运行的轨道上。因而这能很好地找出设备潜在问题或者出现征兆的时刻,避免设备出现事故。

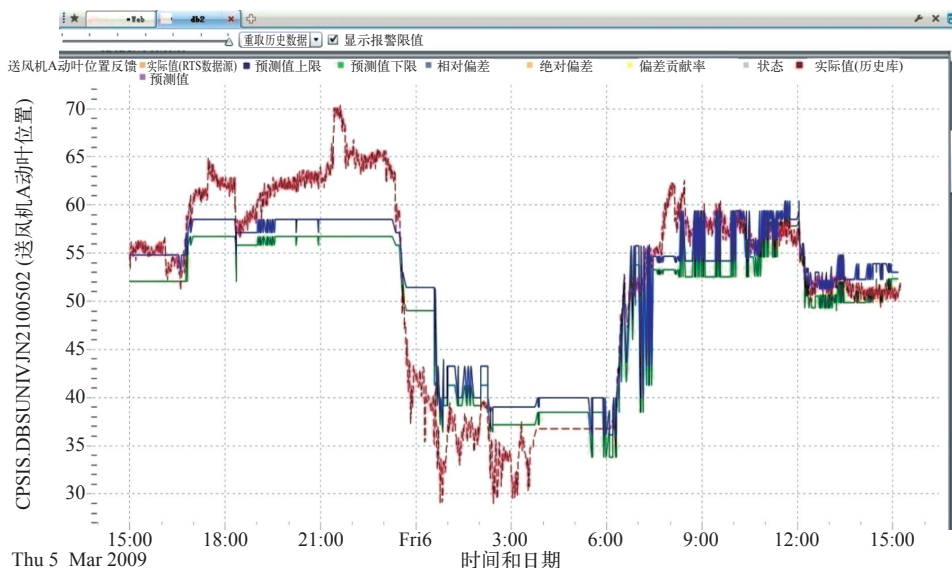


图4 预测结果示意图

结束语

1) 凝聚的层次聚类算法能较好地提供设备运行时的标准曲线以便运行人员及时发现问题,同时有效地利用设备产生的时间序列数据。

2) 层次聚类算法存在自身的局限性,不具备

很好的可伸缩性,如果合并不当会导致低质量的预测结果。EA

(收稿日期: 2017.02.09)

(编辑: 刘群焱)