第 30 卷第 2 期　　四川理工学院学报(自然科学版)　　Vol. 30　No. 2
2017 年 4 月　　Journal of Sichuan University of Science & Engineering(Natural Science Edition)　　Apr. 2017
文章编号:1673-1549(2017)02-0021-05　　DOI:10.11863/j.suse.2017.02.05

电网历史数据缺失及补录研究

谢翘楚[1,2], 姚　毅[1,2]

(1.四川理工学院自动化与信息工程学院, 四川　自贡　643000;2.人工智能四川省重点实验室, 四川　自贡　643000)

摘　要:电网历史数据是智能电网信息化发展的基础,确保历史数据的完整非常必要。在分析电网数据采集与传输过程中产生数据缺失原因的基础上,对缺失数据进行了类型划分,并给出了发现和判定数据缺失的方法。根据数据缺失类型,采用缺失数据清洁法和缺失数据补录法处理规律缺失数据和不规律缺失数据,使用 SPSS 验证上述方法,结果表明补录准确度高达 90%;运用随机森林算法处理不完全规律缺失数据,以均方根误差和填补准确度为评判指标,实验结果证明了该方法的准确性和有效性。用这些方法处理电网的数据缺失问题,能有效地提高电网历史数据的质量,使现有的数据得到最大程度的利用。

关键词:电网历史数据;数据缺失;数据补录;随机森林

中图分类号:TP274　　**文献标志码:**A

引　言

近年来,随着全球智能电网的飞速发展,国家电网公司为我国的智能电网建设提出了新的要求,发展高速、高效的智能电网系统变得势在必行[1]。

智能电网技术就是实现信息化、自动化、互动化,构建以特高压为骨干网架、各级电网协调发展的统一。电网的历史数据就是智能电网信息化建设的数据基础。然而在实际中,各个变电站的数据在提取和传输时,会产生海量的杂乱无章的数据,其数量级别是呈指数级增长的,这些数据在传输和使用的过程中,有相当一部分数据因为人为因素或客观因素发生了缺失的现象,对智能电网的信息化建设带来了很大的不便。为了提高电网数据的质量,保障数据的完整性,为智能电网的发展扫清障碍,解决电网数据缺失是很有必要的。

本文阐述了智能电网变电站监控系统所产生的数据传输过程,并针对在传输过程中所产生的数据缺失问题,提出了处理数据缺失的方法。

1　数据缺失的产生原因及类型

电网的监控平台可以管理一部分区域内的所有变电站,并将其产生的海量数据进行数据挖掘分析,获取其中有用的数据,寻找到一定的规律,对智能电网建设起到积极作用[2]。

整个电网系统中,数据的传输大致可分为单向流传输与双向传输,本文主要研究单向流传输的数据缺失。传输过程为:个体变电站→数据集控站→县级调度→市级调度→省级调度。

数据在传输过程中,会产生很多的缺失,产生缺失的原因大致可分为两类,主观原因和客观原因。人为因素所导致的数据采集或传输造成的数据缺失可称为主观原因,如录入数据失误、工作失职或有意伪造数据所造成的数据缺失。设备故障、路线中断等客观原因所造成的数据缺失可称为客观原因,如数据存储失败、变电站机械故障、数据传输路线截断等。

尽管变电站的历史数据属性众多且繁杂,但是根据

收稿日期:2016-12-12
基金项目:四川理工学院研究生创新基金项目(20141210)
作者简介:谢翘楚(1991-),男,四川自贡人,硕士生,主要从事大数据处理方面的研究,(E-mail)luckyxc1991@163.com

数据产生缺失的原因,大致可以把数据缺失情况归为三类:无规律缺失、规律缺失、不完全规律缺失[3]。

无规律缺失是指该数据是完全随机的,其数据类型不能由已知的数据类型来判断。规律缺失是指该数据是有规律可循的,其数据类型可以由已知的数据来补充或推断。不完全规律缺失是指该数据中既有无规律缺失数据,也有规律缺失数据。

2 数据缺失的发现

数据缺失问题在基于传感器采集数据的发电厂普遍存在,严重阻碍了电力科学与工程数据分析及挖掘在变电站优化领域的发展。

变电站数据采集、存储系统组成复杂,测点工作环境恶劣等多方原因能够造成数据的缺失,主要分为:传感器故障、数据传输故障、数据存储故障、人的主观因素等。数据的不完整性给数据挖掘过程、数据分析和研究带来了重重困难,这些不完整的数据会导致分析结果发生偏置,建立错误的数据挖掘模型,导致不准确的挖掘结果,甚至会误导用户的决策,导致经济损失[4-7]。

依据数据类型的重要程度来划分数据的级别,例如首先将变压器(油中溶解气体、局部放电等),高压断路器(气体成分),高压母线(温度)设定为优先级较高的数据,其次对各级别的数据依照以往的正常数据量设定相应的阈值,如果数据量低于阈值,即可判断数据发生了缺失,再次根据即时数值与阈值的差距,对数据的缺失情况进行评级[8]。在对数据进行检测时,若发生数据缺失,系统会根据数据的优先级别和阈值来一一判断数据在哪个部位发生了缺失。

不同类型的数据缺失情况,应该有相应的缺失发现机制。

(1)规律缺失数据的发现

若数据缺失是呈规律性或遵循函数而发生的,系统会将其判定为规律缺失数据。

(2)不规律缺失数据的发现

若数据缺失是呈无规律性或随机发生的,系统会将其判定为不规律缺失数据。

(3)不完全规律缺失数据的发现

若数据缺失即存在规律数据缺失又存在不规律数据缺失,系统会将其判定为不完全规律缺失数据。

3 数据缺失的处理

传输中产生数据缺失会对整个电网监控平台的实际效果产生巨大的负面影响,因此,对这些缺失的数据进行处理变得势在必行。根据现在大数据处理技术对于数据缺失的处理办法,可以对电网产生的数据缺失使用缺失数据清理法和缺失数据补录法。在数据量较大时,普通的人工补录效率会十分低下,而一般的基于统计学原理的补录方法(如采样法、回归预测法、EM 算法等)会出现较大的偏差,这就需要设计更加适合的补录决策。

3.1 缺失数据清洁法

缺失数据清洁法主要分为删除法和权重法。

删除法是处理缺失数据最简单的方法,就是将缺失的个体直接删除。如果直接删除掉一部分个体数据就可以达到预期数据的目标,这个方法是最有效的。

权重法即当缺失值的类型为规律缺失时,通过对整体的数据加权来降低整体数据的偏差。把数据缺损的个体分别记录后,用线性回归法求得缺损数据各个部分的权重,然后将整体的数据个体给予有差异的权重。假如个体数据类型中存在对权重估计起决定性因素的变量,那该方法可以降低数据的缺损程度。假如个体数据类型中的变量和权重并不相关,那它并不能降低数据缺损程度。所以针对多个数据类型缺失的情况,就需要对不同类型的缺失组合给予有差异的权重,这将会加大数据处理的工作量,使预期结果发生偏移[9]。

缺失数据清洁法可运用于电网监控系统中表现较为良好的设备所产生的数据,但当数据类型比较复杂或设备产生的问题较多时,此类方法将会加大决策人员工作量,导致不能精确分析问题产生的原因,降低电网数据分析效率等。

3.2 缺失数据补录法

大数据处理技术的背景下,当海量数据出现一定的缺失情况时,如果单纯地使用数据清洁法,会造成许多有用数据的遗失,这会对之后的数据挖掘和分析产生巨大的负面影响。因此,对缺失数据进行预估和补录的对策(数据补录法)应运而生。

根据规律缺失数据和无规律缺失数据和不完全规律缺失数据,采用相应的方法解决。

3.2.1 规律缺失数据补录

针对规律缺失数据,运用系统已形成的规律数据,建立相应的线性回归方程式和决策树,对缺失的数据进行预估,形成相应的预测数据,使用相应的预测数据对缺失的数据进行替换,此方法的准确程度将会随着数据库中线性回归方程式和决策树的准确度的提升而提升[10]。

采用最小二乘法计算线性回归方程:

$$\hat{y} = bx + a \tag{1}$$

$$b = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{2}$$

$$a = \bar{y} - b\bar{x} \tag{3}$$

当式(1)中 a、b 取得最小值时，则称式(1)为该数据的线性回归方程，式(2)与式(3)为求解线性回归方程的方式。

这里采用 SPSS 的数据缺失处理进行规律缺失数据的实证。数据库为 1978 – 2005 年的电量使用率。首先使用 SPSS 的数据缺失值替换功能(图 1 与图 2)；然后发现缺失值(图 3)；再对缺失值进行补录(图 4)。

图 1　SPSS 选择替换缺失值

图 2　智能选择替换方法

如图 4 所示，根据以上的原理，系统对缺失的数据生成了一个新的补录值 5048，而 1995 年该变电站的实际电量使用量为 5429，准确度超过 90%，证明此方法在实际工作中有效，能有效提升电网历史数据质量。

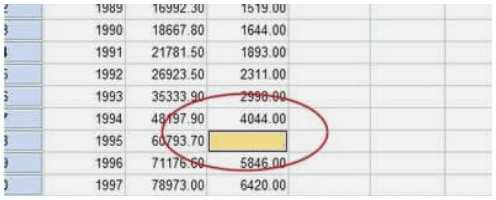

图 3　发现缺失值

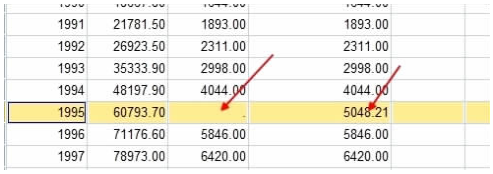

图 4　对缺失值进行补录

另外还可以采用就近补齐法和多重补录法应对不同程度数据缺失情况的补录。其中，就近补齐法是在之前未发生缺失的相近数据中找到与缺失值最为相似的一个值来补录，但相对需要的人工时间较多，适用于对于相对重要的数据缺失的补录；多重补录法是通过记录之前所有缺失的数据所形成的一个数据库来匹配相应的缺失数据，根据缺失值的规律特征从数据库里调出匹配度最高的数据来进行补录。

3.2.2　无规律缺失数据补录

针对无规律缺失数据，目前采用平均值补录最为有效，即将这些无规律的数据类型进行分类，取与该缺失数据属性相近的数据平均值与该类数据进行替换[11]。

3.2.3　不完全规律缺失数据补录

在数据量特别大且数据类型多为不完全规律缺失数据时，如何对数据缺失的类型进行分类和处理，就要运用到大数据处理中的随机森林原理。

如文献[12]所述，随机森林顾名思义，是用随机的方式建立一个森林，森林里面由很多的决策树组成，决策树相互之间是没有关联的。在得到森林之后，当有一个新的输入样本进入的时候，就让森林中的每一棵决策树分别进行判断，判断这个样本应该属于哪一类，然后判断哪一类被选择最多，就预测这个样本为哪一类。

通过总结之前发生数据缺失的数据特征，形成相应的决策树，通过这些决策树群对新的数据缺失样本进行分类。

按这种算法得到的随机森林中的每一棵都是很弱的，但是决策树的数量多了就会对决策结果准确率产生较强的正面影响。总之，在随机森林算法中，每一棵决策树就是一个精通于某一个窄领域的“专家”，这样在随

机森林中就有了很多个精通不同领域的“专家”,对一个新的问题(新的输入数据),可以用不同的角度去看待它,最终由各个“专家”,投票得到结果。这样可以较为准确的对已知数据样本的类型进行智能的分类[13]。

随机森林中的每一棵分类树为二叉树,其生成遵循自顶向下的递归分裂原则,即从根节点开始依次对训练集进行划分;在二叉树中,根节点包含全部训练数据,按照节点纯度最小原则,分裂为左节点和右节点,它们分别包含训练数据的一个子集,按照同样的规则节点继续分裂,直到满足分支停止规则而停止生长。若节点 n 上的分类数据全部来自于同一类别,则此节点的纯度 $I(n)=0$,纯度度量方法是 Gini 准则,即假设 $P(Xj)$ 是节点 n 上属于 Xj 类样本个数占训练。

具体实现过程如下:

(1) 原始训练集为 N,应用 bootstrap 法有放回地随机抽取 k 个新的自助样本集,并由此构建 k 棵分类树,每次未被抽到的样本组成了 k 个袋外数据。

(2) 设有 n 个变量,则在每一棵树的每个节点处随机抽取 m 个变量,然后在 m 中选择一个最具有分类能力的变量,变量分类的阈值由通过检查每一个分类点确定。

(3) 每棵树最大限度地生长,不做任何修剪。

(4) 将生成的多棵分类树组成随机森林,用随机森林分类器对新的数据进行判别与分类,分类结果按树分类器的投票多少而定[13]。

这里采取均方根误差(Root Mean Square Error, RMSE)和填补准确度(Accuracy)评价算法的优越性。均方根误差 E_{RMSE} 是缺失值填补研究中应用最广泛的评价标准:

$$E_{RMSE}=\sqrt{\frac{\sum_{i=1}^{n}(x_r-x_i)^2}{n}} \tag{4}$$

式中:x_r 为真实值;x_i 为算法的填补值;n 为缺失值的数目;E_{RMSE} 值越小说明算法填补质量越高[14]。

填补准确度 A 评价函数能够计算出填补值中符合容忍度要求的值所占的比例:

$$A=\frac{n_T}{n}*100\% \tag{5}$$

式中:n 为缺失值数量;n_T 为正确估计值数量。填补值在真实值的 ±10% 以内都可被视为在容忍度范围之内,即为正确估计值[15]。

将随机森林算法与当前填补效果较好的 SVR - OCSFCM 算法[16](即支持向量回归与遗传算法优化的模糊聚类填补算法)进行比较,取某变电站的油纸绝缘中局部放电量、油中火花放电量及油中电弧为数据集,以油中电弧为填补量,对这两种算法得到的均分根方差和填补准确度进行分析比较。根据分析得的结果如图 5 与图 6 所示。

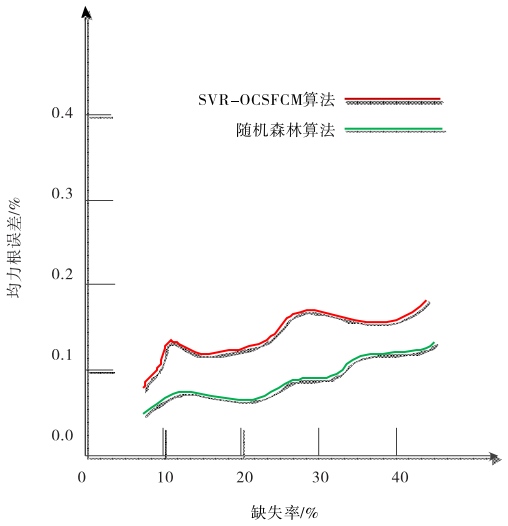

图 5 填补结果的均方根误差

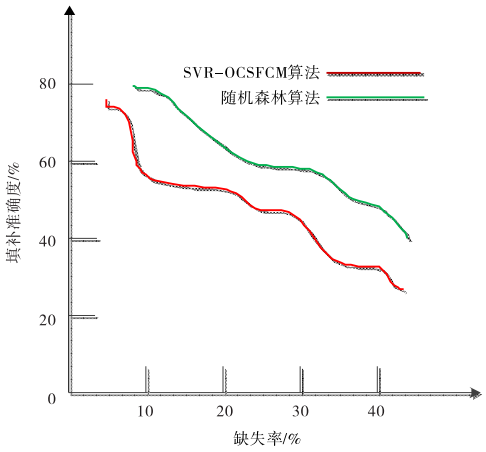

图 6 填补结果的准确率

由图 5 与图 6 对均方根误差与填补准确率分析可知,随着缺失率的提升,随机森林算法在均方根误差和填补结果准确率上都要优于 SVR - OCSFCM 算法。

4 结束语

将这些数据缺失处理方法应用于电网数据处理中,大大提升了数据的可用性,提高了电网各类数据的挖掘分析效率,将有效推动我国智能电网的建设与发展。

参 考 文 献:

[1] 李佳玮,郝悍勇,李宁辉.电网企业大数据技术应用

研究[J].电力信息与通信技术,2014,12(12):20-25.

[2] 于存水.基于智能电网调度系统的调度监控平台的设计与实现[D].长春:吉林大学,2013.

[3] 李丽.数据缺失及处理方法探析[J].湖南城市学院学报:自然科学版,2016,25(1):118-119.

[4] DRISCOLL M.Duke Energy's data modeling & analytics initiative[R].2014.

[5] 武森,冯小东,单志广.基于不完备数据聚类的缺失数据填补方法[J].计算机学报,2012,35(8):1726-1738.

[6] 韦钢,王飞,张永健,等.负荷预测中历史数据缺损处理[J].电力科学与工程,2004,20(1):16-19.

[7] DONG L J,LIU X,ZHANG Q,et al.Design and implementation of metering abnormal and online diagnosis system of new generation intelligent substation[J]. Applied Mechanics & Materials,2014,678:343-351.

[8] 侯广松.变电站故障数据处理与分析系统研究与开发[D].济南:山东大学,2014.

[9] 叶素静,唐文清,张敏强,等.追踪研究中缺失数据处理方法及应用现状分析[J].心理科学进展,2014,22(12):1985-1994.

[10] 吴刘仓,张家茂,邱贻涛.缺失偏态数据下线性回归模型的统计推断[J].统计与信息论坛,2013,28(9):22-26.

[11] 赵志文,何静花,杨慧超.Rayleigh 分布总体参数的均值填补估计和检验[J].佳木斯大学学报:自然科学版,2016,34(2):285-288.

[12] AURET L,ALDRICH C.Change point detection in time series data with random forests[J].Control Engineering Practice,2010,18(8):990-1002.

[13] 曹正凤.随机森林算法优化研究[D].北京:首都经济贸易大学,2014.

[14] 卜范玉,陈志奎,张清辰.基于聚类和自动编码机的缺失数据填充算法[J].计算机工程与应用,2015,51(18):13-17.

[15] 李建强,赵凯,潘文凯,等.电站历史数据缺失值填补策略研究[J].电力科学与工程,2017,33(1):43-48.

[16] 唐阔,胡国圣,车喜龙,等.基于遗传算法优化支持向量回归机的网格负载预测模型[J].吉林大学学报:理学版,2010,48(2):251-255.

Research on the Data Missing and Data Completion of Power Grid

XIE Qiaochu[1,2], *YAO Yi*[1,2]

(1. School of Automation & Information Engineering, Sichuan University of Science & Engineering, Zigong 643000, China; 2. Artificial Intelligence Key Laboratory of Sichuan Province, Zigong 643000, China)

Abstract: The completion of data is needed in the development of smart grid, so it is necessary to improve the data quality of smart grid. The transmitting procedure of the smart grid's big data is introduced and the reasons of the data missing and the type of missing data in the process of data transmission are analyzed. According to the analysis of the missing data cleaning and the missing data collection, the problems of irregular missing data and missing data patterns are solved. Then SPSS is used to validate the methods. The results show that the accuracy rate is as high as 90%. The random forest algorithm is introduced to deal with the incomplete data. And the accuracy and effectiveness of the above methods are proved by the experiments. The methods to the data missing problems of smart grid above will effectively improve the quality of the smart grid data and get the most use of existing data.

Key words: smart grid; data missing; data completion; random forest