智能用电数据的采集与预处理

邓东林[1],徐 尤[1],陈 剑[1],杨仁增[2]
(1. 贵州电网有限责任公司遵义供电局,贵州 遵义 563000;
2. 贵州理工学院贵州省电力大数据重点实验室,贵州 贵阳 550003)

摘 要:解决好智能用电网络数据采集和传输过程中的数据缺失和噪声问题,提高其用电数据的数据质量,才能在智能用电云平台中有效的运用各种用电大数据分析与预测算法。本文在总结智能用电网络的数据采集与数据传输特点,及分析智能用电云平台对用电数据的数据质量要求的基础上,提出了智能用电网络的用电数据预处理方法。对智能用电终端采集的用电数据归一化处理后,利用聚类算法从噪声、模糊、随机数据中提取出正常数据,本文对比验证了 K-均值聚类和基于密度的空间聚类两种算法的聚类效果。相比 K-均值聚类算法,基于密度的空间聚类算法在检测数据噪声点的同时,可自动获取复杂形状数据集的聚类数量,更适合智能用电网络的用电数据预处理。

关键词:智能用电网络;数据质量;数据预处理;聚类算法
文章编号:2096-4633(2019)03-0081-06 **中图分类号**:TM93 **文献标志码**:B

终端电力用户的负荷数据是配电网重要的基础数据之一。对负荷数据进行分析和预测,电力企业可加强对配电网的系统规划和日常运营,提高配电网的安全可靠性和经济性。对负荷数据的分析和挖掘,电网可以获得电力用户类型,了解用户的个性化、差异化服务需求,进一步拓展电力服务的深度和广度。电力用户可了解自身的用电行为和日常能耗情况,并能根据实时电价进行需求侧响应,采取及时的节能降耗措施。负荷数据的有效采集和分析对智能配电网的构建必不可少[1-7]。

智能用电网络是一种随着智能电网的发展而形成的用户侧物理信息系统,它以能量信息网关为数据传输中心节点,通过智能终端连接各种用电设备而形成,具备本地计算、数据通信、远程控制和自治等基本功能。基于需求侧响应而构建的智能用电网络,能够有效提升配电网的安全、经济运行水平,实现配电网的智能化[8-12]。

智能用电网络运行过程中,由于测控终端故障、冲击性负荷投切、通信故障等随机因素,会导致系统采集到的负荷数据包含大量异常值、缺失值、噪声数据等。由这些数据进行分析与挖掘得到的用电负荷曲线,不能准确的反应负荷变化的正常规律,质量较差的用电数据还会导致错误的决策。因此在分析和挖掘用电数据之前,辨识与修正异常负荷数据的数据预处理过程及其重要[13-15]。

本文在介绍智能用电网络系统架构的基础上,总结智能用电网络的数据采集与数据传输特点,分析智能用电网络用电数据的标准化处理方法,然后基于聚类算法设计智能用电网络适用的数据预处理方法,以提高用电数据的数据质量,满足后续各种用电数据分析与挖掘算法提高算法运用效果的需要。

1 智能用电网络的系统架构

智能用电网络具备对用电终端电器的监测与控制、并对终端负荷进行能效管理、参与电网优化运行的三个主要功能。智能用电网络主要由智能插座、信息能量网关、数据处理服务器和云端大数据服务器等组成。智能用电网络的系统架构如图 1 所示。

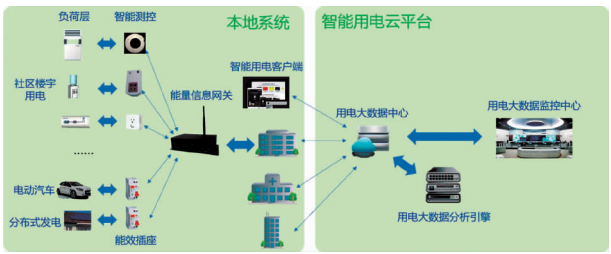

图 1 智能用电网络示意图

Fig. 1 Schematic diagram of smart electric appliance network

数量众多的智能测控终端是整个智能用电网络的底层数据采集部件,可以实现电压、电流、功率等

电气量及设备开关状态的高精度采集,具有低待机功耗及智能无线传输自组网的特点。智能终端采集的遥测、遥信数据由 ZigBee 无线通信网络传输到能量信息网关,并通过能量信息网关获得智能用电云平台的遥控、遥调指令。目前微型智能终端的待机功耗在控制在 300 毫瓦以内,采用基于事件驱动的数据采集和通讯机制,及改进的数据实时压缩技术,可使智能用电网络的终端数据采样时间缩短到 100 ms。

能量信息网关基于内嵌 Linux 系统的芯片搭建,是连接智能测控终端与智能用电云平台的动态数据交换单元,具有上下行数据传输和本地计算功能。通过串口与内置 ZigBee 通信模块实现信息交互,通过以太网或无线公网与智能用电云平台进行通信,分别实现智能终端遥测、遥信数据的上行传输,及智能用电云平台遥控、遥调数据的下行传输。

汇集到智能用电云平台的用电负荷数据,通过进行数据分析和挖掘,可对用电终端设备进行能效评估、状态评价及故障检测,从而提高用户终端的供电可靠性,并实现用电设备的节能降耗,还可进行电力用户用电行为特征分析、用电负荷预测与用电模式优化,从而全面提升用户用电体验,提高用电能效,实现需求侧管理与运行优化。用电数据的有效分析和挖掘,能够有效提升配电网的用电安全水平、能效管理水平和智能化水平,提高用户用电满意度,实现电网与用户之间能量和信息的双向互动,实现电力资源的优化配置,实现分布式新能源的有效消纳,及满足电动汽车随机接入的用电需求。

2 采集数据的标准化

智能用电云平台中进行数据分析与挖掘的各种机器学习算法,如神经网络、支持向量机等,对数据的异常值非常敏感。数据预处理主要是消除异常数据对算法模型的影响,使经过数据清洗、数据特征调节等方法预处理后的数据表示更适合于负荷分析及预测算法。

智能用电网络中的异常数据产生的原因较多。首先,智能测控终端的损坏和异常,可能会导致采集数据缺失。其次,线路维护或者安检等配电网的正常活动可能会导致用电数据缺失,而用电数据本身从测控终端到用电云平台的数据传输也有可能导致数据异常。

异常数据包括离群值、噪声、偏差等,智能用电网络中的异常数据分为两类即数据突变和数据缺失。电网上的事故或事件会导致用电数据突变,以及毛刺数据即多个负荷数据点的突然增加或减少。丢失的数据通常是因为在数据收集过程中,测控终端产生故障,或者数据文件的传输过程中的数据丢失,会使得记录的数据为 NULL 或者为 0,使得这些数据偏离真实负荷数据。

采集和传输过程中的数据缺失和数据突变,会使得智能用电数据产生空缺值、出现噪声数据、出现不一致数据,不利于后续的数据分析与处理。因而在智能用电云平台,需要对采集的用电数据做相应的数据预处理。

聚类算法能从大量含噪声、模糊、随机数据中提取出正常数据,用电数据的聚类还可直接挖掘出负荷模式,得到用户用电模式,有利于提高后续负荷分析算法的精度,因而聚类算法较适合于智能用电网络中用电数据的预处理。

聚类算法自身对异常数据也较为敏感,用电数据中的异常数据可能会影响负荷聚类的效果,产生错误的分类,所以有必要在前端对用电数据进行标准化处理,同时数据标准化也利于提高智能用电网络中的数据传输速率。

数据标准化处理也就是数据的归一化,将数据归一化以后,很多数据分析算法都能够发挥最佳的效果。归一化的数值对于聚类算法而言,可以均衡化不同用电指标尺度,避免采集数据由于数量级差别过大对聚类结果的影响,可以消除用电数据量的大小对聚类分析中距离的影响,更加注重用户用电模式的特征信息。

用电数据的归一化是将智能终端的采集数据和输出数据变换到[0,1]区间内,常使用下列的变换式:

$$\bar{x}_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

式中,$x_{\max}$ x_{max} 和 x_{min} $x_{\min}$ 分别为采集数据集合中负荷数据的最大值和最小值,x_i x_i 为实际的负荷数据,$\bar{x}_i$ 为归一化后的用电数据。

图 2 给出了用电数据进行归一化处理的算例示意图。

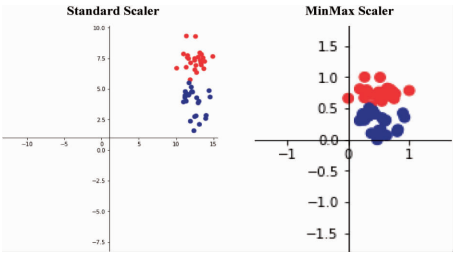

图 2　用电数据归一化算例图

Fig. 2　Example diagram of normalization calculation for electric appliance network data

3　用电数据的 k－均值聚类算法

聚类是对数据集按照某种距离测度将它们聚成多个“簇”的过程。聚类分析将数据集划分成多个簇,使得同一个簇中的对象彼此相似,但与其他簇中的对象不相似,最终使得同一簇内的点之间距离较小,而不同簇内点之间距离较大。聚类分析要得到较好的聚类结果,就应该使各个聚类中心的距离尽可能大。

聚类算法能从大量含噪声、模糊、随机数据中提取出正常数据,还可直接从众多平行用电数据中获得聚类负荷特征曲线。

k－均值聚类是最简单、最常用的聚类算法。k－均值聚类以欧式距离作为距离测度,其计算式为:

$$d([x_1,x_2,\cdots,x_n],[y_1,y_2,\cdots,y_n]) = \sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \tag{2}$$

k－均值聚类算法一般采用均方差作为标准测度函数,其表达式为:

$$J_c = \sum_{j=1}^{k}\sum_{i=1}^{n_j} \| x_i - m_j \|^2 \tag{3}$$

式中,k 为要形成聚类的个数,n_j 为第 j 类中样本的个数,m_j 为第 j 类样本的均值,代表此分类数据集的中心,即:

$$m_j = \frac{1}{n_j}\sum_{j=1}^{n_j} x_j, j = 1,2\cdots,k \tag{4}$$

K-均值算法的工作流程:首先从 n 个数据对象任意选择 k 个对象作为初始聚类中心;对于数据库中剩下的其它对象,则根据它们与这些聚类中心的欧式距离,分别将它们分配给与其最相似的(聚类中心所代表的)聚类;然后再计算每个所获新聚类的聚类中心(该聚类中所有对象的均值);迭代这一过程直到标准测度函数开始收敛为止。

由 K-均值算法的工作流程可知,聚类参数 k 值是认为设定的,对于复杂数据集的 K-均值聚类处理,k 值的选定是非常难以估计的,事先并不知道采集的海量用电数据集应该分成多少个类别才最合适。K-均值聚类算法随机选取初始聚类中心,会导致过多次的迭代,影响算法的实时性,并得到质量不高的聚类结果。K-均值算法通常需要根据初始聚类中心来确定一个初始划分,然后对初始划分进行优化。而这个初始聚类中心的选择对聚类结果有较大的影响,一旦初始值选择不当,可能根本得到有效的聚类结果。

4　基于空间密度聚类的用电数据预处理

DBSCAN(density based spatial clustering of applications with noise)算法是一种基于密度的空间聚类算法,它将簇定义为密度相连的点的最大集合,能够把具有足够高密度的区域划分为簇,并可在有“噪声”的空间数据集中发现任意形状的聚类[16－18]。

DBSCAN 算法的主要思想是:只要临近区域的密度(数据点数目)超过某个阈值,就把它加到与之相近的类中,一般而言,高密度的数据点区域被低密度的数据点区域(通常认为是噪声数据点)所分割。因此,DBSCAN 算法可过滤噪声点数据,自动识别任意形状的类簇,较适合智能用电云平台数据的聚类分析。

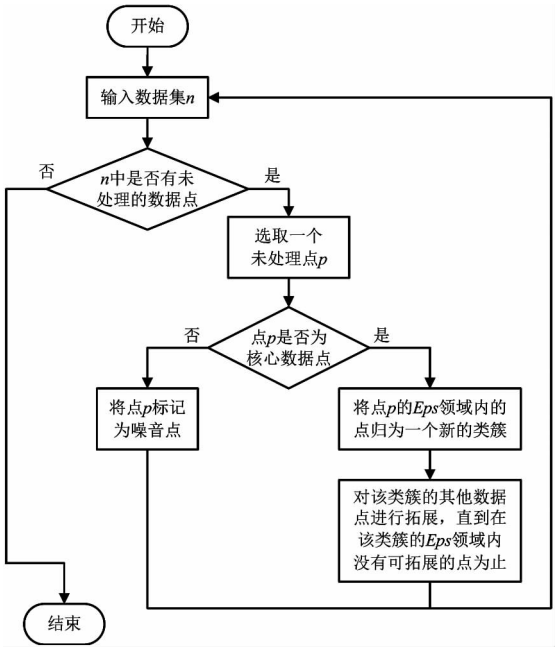

图 3　DBSCAN 算法流程图

Fig. 3　Flow chart of DBSCAN algorithm

图 3 给出了 DBSCAN 算法的流程,说明如下。

(1)输入:包含 n 个对象的数据库,半径 Eps,最少簇数目 $MinPts$。

(2)输出:所有生成的簇,达到簇密度要求。

(a)从数据集中任意选取一个点 p,并对其进行区域查询;

(b)如果 p 是核心点,则寻找所有从 p 密度可达的点,最终形成一个包含 p 的簇;

(c)否则,p 被暂时标注为噪声点;

(d)访问数据集中的下一个点,重复上述(a)-(c)的过程,直到数据集中所有的点都被处理。

聚类算法通常选择调整兰德指数(adjusted rand index,ARI)作为定量评价指标,ARI 的基础是兰德指数(rand index,RI),RI 的表达式为:

$$RI = \frac{a + b}{c_2^n} \tag{5}$$

式中,a 表示在实际类别 c 与聚类结果 k 中都是同类别的元素对数,b 表示在 C 与聚类结果 k 中都 是不同类别的元素对数,c_2^n 表示数据集中可以组成的对数,RI 取值范围为[0,1],值越大表示聚类效果准确性越高。

由 AR 得到的 ARI 的表达式为:

$$ARI = \frac{RI - E(RI)}{\max(RI) - E(RI)} \tag{6}$$

ARI 取值范围为[-1,1],值越大意味着聚类结果与真实情况越吻合。作为聚类算法的评价指标,ARI 比 RI 具有更高的区分度。

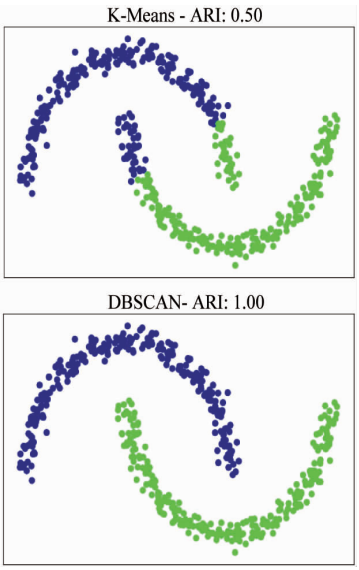

图 4 两种聚类算法的测试结果

Fig. 4 Test results of two clustering algorithms

以 500 个采集点、噪声均值为 0、方差为 0.05 的数据集,分别测试 K-均值算法与 DBSCAN 算法的聚类效果。图 4 给出两种算法的 ARI 评价指标结果,由图可见 DBSCAN 完美地自动实现了受噪声污染、复杂形状数据集的期望聚类。

对比分析可知,K-均值算法需要根据初始聚类中心来确定一个初始划分,然后对初始划分进行优化,初始聚类中心的选择对聚类结果有较大的影响,一旦初始值选择的不好,可能无法得到有效的聚类结果。而 DBSCAN 算法可检测到没有分配任何簇的“噪声点”,还可自动判断簇的数量,它允许簇具有复杂的形状,更适合智能用电云平台的数据预处理。

5 结束语

在总结智能用电网络的数据采集与数据传输特点,分析智能用电云平台对用电数据的数据质量要求的基础上,本文提出了将智能用电网络采集数据进行归一化数值处理,然后进行 DBSCAN 聚类运算的用电数据预处理方法。采集数据在智能用电云平台前端的归一化处理可均衡化不同用电指标尺度,提高聚类算法的聚类效果,聚类算法能从大量含噪声、模糊、随机数据中提取出正常数据,并可直接挖掘出负荷模式,得到用户用电模式,有利于提高后续负荷分析与处理算法的运用效果。相比常用的 k-均值算法,DBSCAN 算法可自动判断聚类簇的数量,更好的检测数据“噪声点”,识别形状复杂的数据集,提升智能用电云平台用电数据的数据质量,为后续用电大数据的分析与预测工作,奠定了良好的基础。

参考文献:

[1] 张素香,赵丙镇,王风雨,等. 海量数据下的电力负荷短期预测[J]. 中国电机工程学报,2015,35(01):37-42.
ZHANG Suxiang, ZHAO Bingzhen, WANG Fengyu, et al. Short-term power load forecasting based on big data[J]. Proceedings of the CSEE,2015,35(01):37-42.

[2] 吴润泽,包正睿,王文韬,等. Hadoop 架构下基于模式匹配的短期电力负荷预测方法[J]. 电工技术学报,2018,33(07):1542-1551.
WU Runze,BAO Zhengrui,WANG Wentao,et al. Short-term power load forecasting method based on pattern matching in hadoop framework[J]. Transactions of China Electrotechnical Society,2018,33(07):1542-1551.

[3] 胡丽娟,刁赢龙,刘科研,等. 基于大数据技术的配电网运行

可靠性分析[J]. 电网技术,2017,41(01):265 - 271.
HU Lijuan, DIAO Yinglong, LIU Keyan, et al. Operational reliability analysis of distribution network based on big data technology[J]. Power System Technology, 2017, 41 (01): 265 - 271.

[4] 凌德祥,黄 拓,关晓林,等. 基于大数据的电力客户行为分析体系研究及实践[J]. 电力大数据,2018,21(10):13 - 17.
LING Dexiang, HUANG Tuo, GUAN Xiaolin, et al. Research and practice of power client behavior analysis system based on large data[J]. Power Systems and Big Data,2018,21(10):13 - 17.

[5] 苗新,张东霞,孙德栋. 在配电网中应用大数据的机遇与挑战[J]. 电网技术,2015,39(11):3122 - 3127.
MIAO Xin, ZHANG Dongxia, SUN Dedong. The opportunity and challenge of big data's application in power distribution networks [J]. Power System Technology,2015,39(11):3122 - 3127.

[6] 刘科研,盛万兴,张东霞,等. 智能配电网大数据应用需求和场景分析研究[J]. 中国电机工程学报,2015,12(02):287 - 293.
LIU Keyan, SHENG Wanxing, ZHANG Dongxia, et al. Big data application requirements and scenario analysis in smart distribution network[J]. Proceedings of the CSEE,2015,12(02):287 - 293.

[7] 王鹏,林佳颖,郭岫,等. 配用电数据分析及应用[J]. 电网技术,2017,41(10):3333 - 3340.
WANG Peng, LIN Jiaying, GUO She, et al. Distribution system data analytics and applications[J]. Power System Technology,2017,41 (10):3333 - 3340.

[8] 赵雪霖,何光宇,杨文轩,等. 智能用电网络的设计与初步实现[J]. 电工电能新技术,2014,33(10):52 - 57.
ZHAO Xuelin, HE Guangyu, YANG Wenxuan, et al. Design and initial implementation of smart electric appliance network [J]. Advanced Technology of Electrical Engineering and Energy,2014, 33(10):52 - 57.

[9] XIN S, GUO Q, SUN H, et al. Cyber-physical modeling and cyber-contingency assessment of hierarchical control systems[J]. IEEE Transactions on Smart Grid,2015,6(05):2375 - 2385.

[10] KHAITAN S K, MCCALLEY J D. Design techniques and applications of cyberphysical systems: a survey[J]. IEEE Systems Journal,2014,9 (02):350 - 365.

[11] 刘博,栾文鹏. 基于负荷分解的用电数据云架构方案及应用场景[J]. 电网技术,2016,40(03):791 - 796.
LIU Bo, LUAN Wenpeng. Conceptual cloud solution architecture and application scenarios of power consumption data based on load disaggregation [J]. Power System Technology, 2016, 40 (03):791 - 796.

[12] 郏琨琪,何光宇. 智能用电网络数据采集与通信机制的研究[J]. 中国电机工程学报,2016,36(06):1544 - 1551.
JIA Kunqi, HE Guangyu. Research of smart electric appliance network data collection and communication mechanism [J]. Proceedings of the CSEE,2016,36(06):1544 - 1551.

[13] 张铁峰,顾明迪. 电力用户负荷模式提取技术及应用综述[J]. 电网技术,2016,40(03):804 - 811.
ZHANG Tiefeng, GU Ming. Overview of electricity customer load pattern extraction technology and its application [J]. Power System Technology,2016,40(03):804 - 811.

[14] 苏适,李康平,严玉廷,等. 基于密度空间聚类和引力搜索算法的居民负荷用电模式分类模型[J]. 电力自动化设备,2018 (01):129 - 136.
SU Shi, LI Kangping, YAN Yuting, et al. Classification model of residential power consumption mode based on DBSCAN and gravitational search algorithm [J]. Electric Power Automation Equipment,2018(1):129 - 136.

[15] 林顺富,谢潮,李东东,等. 基于灰色关联与模糊聚类分析的负荷预处理方法[J]. 电测与仪表,2017,54(11):36 - 42.
LIN Shunfu, XIE Chao, LI Dongdong, et al. Load preprocessing method based on grey relational analysis and fuzzy clustering [J]. Electrical Measurement & Instrumentation,2017,54(11): 36 - 42.

[16] WU Y, ZHANG Z. Research on a new density clustering algorithm based on MapReduce[C]// International Conference on Geo-Spatial Knowledge and Intelligence. Springer, Singapore, 2017: 552 - 562.

[17] 邱宁佳,李宾,王鹏,等. 基于 MapReduce 的密度聚类改进算法[J]. 计算机应用,2017,37(S1):63 - 67.
QIU Ningjia, LI Bin, WANG Peng, et al. New density clustering algorithm based on MapReduce [J]. Journal of Computer Applications,2017,37(S1):63 - 67.

[18] 杨晨光. 基于高斯混合模型 DBSCAN 算法的换乘站乘客群体行为特性研究[D]. 北京:北京交通大学,2017.

收稿日期:2018 - 12 - 14

作者简介:

邓东林(1978),男,硕士研究生,高级工程师,主要从事配网生产运行管理,配网自动化、信息化应用管理工作。

(本文责任编辑:范　斌)

Acquisition and preprocessing of smart electric appliance network power data

DENG Donglin, XU Yin, CHEN Jian, YANG Renzeng

(1. Zunyi Power Supply Bureau of Guizhou Power Grid Co., Ltd., Zunyi 563000 Guizhou China;
2. Guizhou Key Laboratory of Electrical Power Big Data of Guizhou Institute of Technology, Guiyang 550003 Guizhou China)

Abstract: Solved the problem of data loss and noise in the process of data collection and transmission of smart electric appliance network, and improved the power data quality, which can be effectively use all kinds of big data analysis and prediction algorithms for the electric appliance cloud platform. Based on the summary of the characteristics of data collection and data transmission of the smart electric appliance network, and the analysis of the data quality requirements of the cloud platform, this paper proposes a method of power data preprocessing of the smart electric appliance network. After normalizing the electric appliance data collected by smart electric appliance terminals, normalization data are extracted from noise, fuzzy and random data by means of clustering algorithm, and the clustering effect of k-means clustering and density based spatial clustering of applications with noise (DBSCAN) algorithms is compared and verified. Compared with k-means clustering algorithm, DBSCAN algorithm can automatically obtain the clustering number of complex shape data sets while detecting data noise points, which is more suitable for electric appliance data preprocessing of the smart electric appliance network.

Key words: smart electric appliance network; data quality; data preprocessing; clustering algorithm