基于 CFSFDP 算法的边缘电力数据异常检测

张哲敏[1]，李琪林[2]，严　平[2]，成贵学[1]
（1. 上海电力大学计算机科学与技术学院，上海　200090；
2. 四川省电力公司计量中心，四川 成都　610045）

摘　要：随着智能电网的不断发展，电力设备产生的数据量逐渐增多，如何利用电力数据成为电网发展的关键。为了保障电力数据的准确性，在边缘端快速检测并处理异常数据，提出了一种基于 CFSFDP 算法的电力数据异常检测的方法。该方法基于 CFSFDP 的假设，将局部密度较低且距高密度点较远的样本点定义为异常值，并创新使用了一种根据前后 k 值自动选择异常值的策略，解决了人工选择时存在主观因素影响的问题。通过与 DBSCAN 和 LOF 的比较表明，该方法能够快速、高效地找出电力数据中的异常值，适用于边缘电力数据异常检测。

关键词：异常值检测；CFSFDP 算法；边缘电力数据；自动选择策略
中图分类号：TM 73　**文献标志码**：A　**文章编号**：1003 - 6954(2021)04 - 0006 - 05
DOI：10. 16527/j. issn. 1003 - 6954. 20210402

Detection of Edge Power Data Anomaly Based on CFSFDP Algorithm

Zhang Zhemin[1], Li Qilin[2], Yan Ping[2], Cheng Guixue[1]
(1. School of Computer Science and Technology, Shanghai University of Electric Power, Shanghai 200090, China;
2. Metering Center of State Grid Sichuan Electric Power Company, Chengdu 610045, Sichuan, China)

Abstract: With the continuous development of smart grid, the amount of data generated by power equipment is gradually increasing. How to use power data becomes the key to the development of power grid. In order to ensure the accuracy of power data and detect and process abnormal data quickly at the edge, a detection method for power data anomaly based on CFSFDP algorithm is proposed. Based on the hypothesis of CFSFDP, the sample points with low local density and far away from high density points are defined as outliers, and a new strategy of automatically selecting outliers based on the k values before and after is used to solve the problem of subjective factors in manual selection. The comparison with DBSCAN and LOF shows that the proposed method can quickly and efficiently find the outliers in power data, and is suitable for outlier detection of edge power data.

Key words: outlier detection; CFSFDP algorithm; edge power data; automatic selection strategy

0　引　言

如今，随着智能电网的发展，电力系统中产生的数据量也不断增多。但安装在发电、输电、配电、用电各个环节各种类型的计量装置和系统，由于外部干扰等原因，会不可避免地出现异常数据，及时有效地检测出异常数据能够保障电力系统的稳定性和安全性。各种离群点检测算法能够检测出那些与正常数据行为或特征属性差别较大的异常数据或行为，有利于降低安全风险，减少经济损失。

目前，已经有一些文献研究了电力数据领域的异常值检测算法。它们可以大致分为基于距离的异常值检测、基于密度的异常值检测和基于聚类的异常值检测等。基于距离的异常值检测方法由 EM Knorr、RT NG[1] 等人在 20 世纪末提出，该方法认为与大多数样本的距离都大于某个固定阈值的点就是异常值点。但这种方法不能判断含有密度不同的多个类簇的数据集。基于密度的异常值检测的原理认为正常样本点所处的类簇密度要高于异常点样本所处的类簇密度。最具有代表性的是基于局部异常因子(local outlier faction, LOF)的异常值检测方法[2]。

基金项目：国家电网有限公司科技项目(基于全息感知和边缘计算的新型电能信息交互设备研究项目 52199719001M)

基于聚类的异常检测其目标是将数据点按照一定的规则划分到某一类中,而异常值检测的目标不属于任何簇的样本点 k 均值聚类算法,据此与正常样本点进行区别。目前,主要的基于聚类的异常值检测 k 均值聚类算法采用 k – means 和 DBSCAN(density – based spatial clustering of applications with noise)算法进行聚类[3-4]。文献[5]针对传统电量数据异常检测方法的不足,提出了一种基于三次指数平滑模型和 DBSCAN 聚类的电量数据异常检测方法。文献[6]采用一种基于孤立森林的异常检测算法,实现大规模电能量数据的异常检测。文献[7]将 DBSCAN 和 LOF 算法相结合,即 KDBLOF,将 k 近邻(k – nearest neighbors,KNN)思想引入到 DBSCAN 中,解决了原 DBSCAN 参数确定困难的问题。

电力数据经采集后会将所有数据上传至集中式数据中心,再使用异常值检测算法做数据清洗,其中异常数据的传输会造成大量的带宽浪费。在边缘端进行异常值检测,可以减少异常数据传输,节省带宽资源。但边缘端一般不具备较高计算能力的计算处理单元,所以需要复杂度低的算法。

基于密度峰值的快速聚类(clustering by fast search and find of density peaks, CFSFDP)算法是 Alex Rodriguez[8]在 2014 年于《Science》上提出的一种快速寻找聚类中心的聚类算法,具有简洁、高效、参数少的特点,十分适合在边缘计算平台中使用。目前,已有不少研究将该算法应用于电力数据异常检测。文献[9]利用 KNN 思想重新定义局部密度和距离,将 CFSFDP 用于电力大数据的异常值检测,但该方法需要人为设置经验参数,不具有普适性。文献[10]采用 LOF 算法和 CFSFDP 算法相结合的聚类算法进行电力数据异常值检测,弥补了 CFSFDP 算法对于局部密度变化大的数据识别能力弱的缺点;但该方法是通过人工选择决策图来实现聚类中心选取,存在主观因素的影响。

下面将 CFSFDP 算法应用于电力数据的异常检测,并提出了一种异常点的选择策略来实现异常点的自动选择。所提方法避免了原算法需要通过决策图人工输入来实现聚类,再从聚类后的数据中寻找异常点的繁琐步骤,降低了算法的冗余性并提高了寻找异常值的效率。

1　CFSFDP 算法

CFSFDP 算法在所提方法中主要基于两个重要的假设思想:一是假设聚类中心相较于其他的样本点局部密度较高,且被局部密度较低的点包围;二是假设各类簇聚类中心之间的距离较远。为了实现这 2 种假设,定义了两种度量方法。

第一个定义是每个点的局部密度,对于每个点 i,它的局部密度 $\rho(i)$ 的表示有 2 种方法,其中:式(1)为截止距离法;式(2)为核距离方法,适用于数据量较小的数据样本。

$$\rho(i) = \sum_j \chi(d(i,j) - d_c) \tag{1}$$

$$\rho(i) = \sum_{i \neq j} \exp\left(-\left(\frac{d(i,j)}{d_c}\right)^2\right) \tag{2}$$

式中:$d(i,j)$ 为点 i 和点 j 之间的欧氏距离;d_c 为截止距离;$\begin{cases}\chi(x)=1, x<0\\ \chi(x)=0, x\geqslant 0\end{cases}$,当 $\chi=1$ 时意味着点 i 和 j 之间的距离小于截止距离。

第二个定义是每个点距离高密度点的距离。对于每个点 i,它距离高密度点的距离 $\delta(i)$ 的定义公式为

$$\delta(i) = \begin{cases}\min\limits_{j,\rho_i<\rho_j} d(i,j), \text{if}\ \exists j \quad \text{s.t.}\ \rho_i < \rho_j\\ \max\limits_{j} d(i,j), \text{其他}\end{cases} \tag{3}$$

根据定义,只有局部密度较大或者全局最大的点,$\delta(i)$ 才能够足够大。

CFSFDP 算法计算局部密度 ρ 和更高密度距离 δ,将数据集映射成二维图并构造一个决策图(如图 1 所示)。在决策图中,ρ 和 δ 都很大的点(靠右靠上的点)即为聚类中心。在选择聚类中心后,再将剩余点分配给距离最近的聚类中心完成聚类。

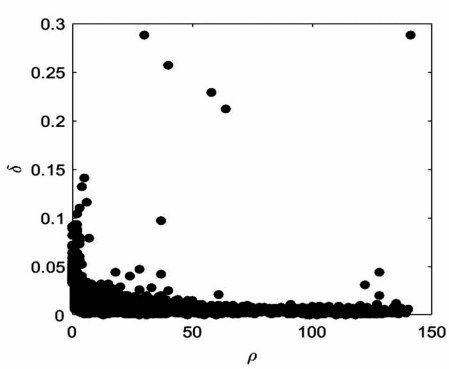

图 1　CFSFDP 算法决策

CFSFDP 算法能够在不确定聚类数目时快速地找到聚类中心,但只适用于特定结构的数据集。对于一些稀疏的数据集,如果经验参数设置不当,可能会取得较差的效果。此外,由于选取聚类中心时采用人为框图框选聚类中心的方式,存在主观因素,不同的选取会得到不同的结果,增加了算法冗余性的同时也不利于实现算法的批量自动化应用。

2 基于 CFSFDP 算法的异常值检测

2.1 CFSFDP 检测异常值思路

根据 CFSFDP 算法提出的假设,从异常值检测的角度来看,可以认为局部密度较低且距离高密度点较远的样本点为异常值点。虽然异常值点距离密度较高的点的距离较正常样本点远,但聚类中心之间的距离同样也很远。如果此时该聚类中心的局部密度不够大,很有可能在人工选择异常值时出现将聚类中心误划分为异常值的情况。对此,引入了一个离群值的概念,将样本点的异常度进行量化,方便进行异常值的选择。

对于每个点 i,它的离群值λ_i的定义公式为

$$\lambda_i = \begin{cases} \dfrac{\delta_i}{\rho_i}, & \rho_i \neq 0 \\ \infty, & \rho_i = 0 \end{cases} \tag{4}$$

当点 i 的局部密度ρ_i 等于 0 时,此时离群值 λ_i 为无穷大,可以直接定义点 i 为异常值点。其他情况下,λ_i 越高,点 i 成为异常值点的概率越大。

2.2 异常值点自动选择策略

通过离群值的定义,为了找出异常值点,可以将离群值大于一定标准的点定义为异常值点。但该标准通常为人工指定,仍然存在主观因素的影响,所以制定了以下策略来实现异常值点的自动选取。

将所有样本点按照离群值进行降序排列,取出前 $m\%$ 的点得到离群值排列图,如图 2 所示。可以看出,虽然离群值整体呈现下降趋势,但下降的程度有所不同,前面下降得快,后面下降得慢。即前半部分离群值相差大、不稳定,可以认为是异常值点;后半部分因为趋向稳定,离群值下降缓慢,可以认为是正常点。在下降程度发生最大变化的点是离群值总体下降由急变缓的拐点。拐点前的是异常值点,拐点后的是正常样本点。

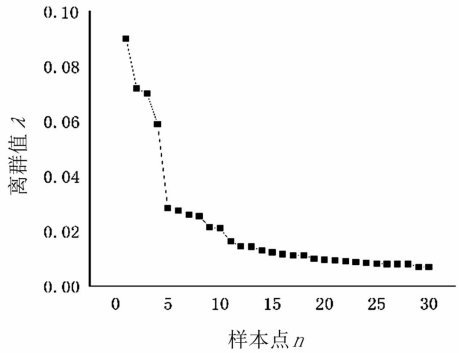

图 2 离群值降序排列

当表示下降趋势时,可以采用斜率进行表示,即

$$k_{i,m} = \frac{\lambda_{i+m} - \lambda_i}{m} \tag{5}$$

式中,$k_{i,m}$ 表示区间$[i, i+m]$ 内的离群值λ 变化率,该参数描述了这一区间 λ 的总体变化趋势。

对于某点前后下降趋势,可以用与前一点线段的斜率和后一点线段的斜率的比值来表示。

$$k_i = \begin{cases} 0, & i = 1 \\ \dfrac{\lambda_{i-1} - \lambda_i}{\lambda_i - \lambda_{i+1}}, & i > 2, \lambda_i \neq \lambda_{i+1} \\ k_{i-1}, & \text{其他} \end{cases} \tag{6}$$

第一个点的下降趋势默认为 0,且当该点的离群值与后一点相同时,该点的变化趋势与前一点相同。计算所有点的变化趋势比值,绘制出图 3 所示的变化率趋势图。拐点为使变化率 k 取得最大值时的点。

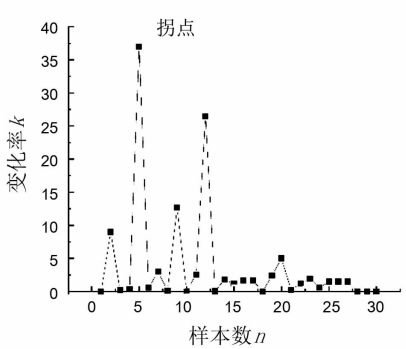

图 3 变化率趋势

得到拐点后,可将拐点前的所有点,视为异常值点,使用 CFSFDP 算法寻找异常值点的具体步骤如下:

1)根据 d_c 确定每一个点的局部密度ρ_i 和距离δ_i。

2)计算每个点的离群值 λ_i 并从高到低排序。

3)取样本点前 $m\%$ 的点计算变化趋势 k_i。m 为经验参数,一般选择 5% ~10%。

4)取使 k 取得最大值的拐点 x。

5)挑选出拐点之前的点$\{1,2\cdots,x\}$作为异常值点。

3　仿真验证

采用2017 年 1 月至 10 月某公司的日用电数据作为研究对现象,采样间隔为 15 min。用户日用电数据作为电力数据的一种,经常因为电能表故障和传输异常等原因,造成上传数据存在异常。但在电力数据的异常值检测场景中,异常值所占比例远低于正常对象。因此,只提取了数据集中的部分数据,使得最终实验数据中异常值与正常值的比值满足异常值检测的一般要求。并且,为了衡量用电数据异常检测算法的有效性,采用的数据提前进行了人工标注,即异常数据已经被标识,方便检验异常检测算法的效果。

在预处理阶段对数据进行了降维和归一化处理,是为了消除因为量纲不同和数量级差距所带来的影响,且可以加快算法的识别速度。按照式(7)对数据进行归一化处理。

$$x_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{7}$$

为了评估基于 CFSFDP 寻找异常值算法的性能,与 DBSCAN 直接检测异常值、局部异常因子 LOF 算法进行了对比试验。DBSCAN 直接检测异常值是先对数据进行聚类,获得不同的类簇;然后求取各个类簇聚类中心间的距离,如果距离过大则认为是异常用电数据。这里设置 DBSCAN 的参数 ρ 为 0.2。

将算法检测出的异常值与数据样本的真实标签作对比,计算并选取检测率(detection rate)和误检率(false positive rate)作为算法评价标准,检测率和误检率的计算公式如下:

$$\text{检测率} = \frac{\text{被检测出的异常数据个数}}{\text{异常数据总数}} \tag{8}$$

$$\text{误检率} = \frac{\text{被检测为异常的正常数据个数}}{\text{正常数据总数}} \tag{9}$$

检测率和误检率的实验结果如图 4、图 5 所示。由图可以看出:1)基于 CFSFDP 算法的异常检测在检测异常值时总体检测率较高,误检率较低,明显优于直接利用 DBSCAN 算法检测异常值和利用局部异常因子算法 LOF 检测异常值;2)对于不同月份的检测样本,直接利用 DBSCAN 算法的异常检测算法的检测率和误检率不同且波动较大,这是因为算法对不同数据样本具有独特性,DBSCAN 只适用于部分样本。相对地,基于 CFSFDP 算法的异常检测就具有较好的适应性,对于不同月份的数据都能维持一个很高的检测率和很低的误检率,变化不大。其中部分月份检测率较其他月份有所降低,原因为该月平均用电量较其他月份有差别,需要提取更多该月样本进行单独检测。

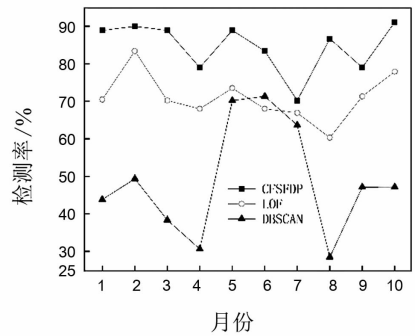

图 4　检测率实验结果

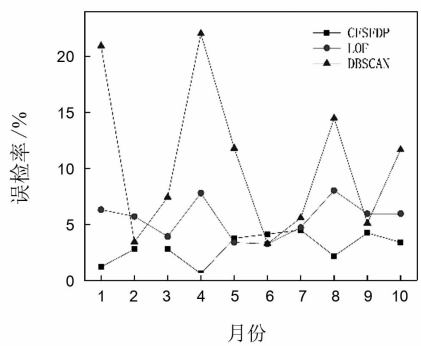

图 5　误检率实验结果

同时,基于 CFSFDP 算法的异常检测还具有快速查找异常值的特点。在实验内存为 8 GB、CPU 为 1.6 Hz 的运行条件下,3 种算法的计算耗时如表 1 所示。

表 1　3 种算法的计算时间

算法	平均运行时间/s
CFSFDP	1.06
LOF	12.31
DBSCAN	30.78

从表 1 可以看出,基于 CFSFDP 算法的异常检测运行时间是比其他两种算法都要短。这不仅证明了基于 CFSFDP 算法的异常检测可以减少计算量,具有快速找到异常值的特点,而且证明了其对大规模数据集具有更好的适应性。

综上,所提出的基于 CFSFDP 算法的异常检测同时具有检测率高、误检率低和运行时间少的特点。在电力生产、调度和决策过程中,可以起到良好的监督防范作用。在用户防窃电方面也能为电力企业提供有力的依据,能够更好地为电力生产和电力缴费服务。

4 结 语

上面对于电力数据的异常检测问题进行了研究,提出了一种基于CFSFDP聚类算法的电力数据异常值检测方法。该方法基于原本的密度峰值快速搜索算法提出的两点有关于聚类中心的假设,设立了离群值指标,在该指标的判断下寻找异常值点,实现了异常值点的快速寻找。同时根据离群值下降趋势,提出一种不需要进行人工选择的自动选择异常值点的策略,避免了进行人工选择时主观因素的影响。通过对比该方法与利用DBSCAN直接寻找异常值和利用局部异常因子LOF寻找异常值的方法,发现该方法能够有效、快速地寻找出异常值点,且该算法复杂度低,耗时短,适合作为边缘设备检测电力数据的算法。

参考文献

[1] Knorr EM, Ng R T, Tucakov V. Distance-based Outliers: Algorithms and Applications[J]. The VLDB Journal, 2000, 8(3):237-253.

[2] Breunig M M, Kriegel H P, Ng R T, et al. LOF: Identifying Density-based Local Outliers[C]// Acm Sigmod International Conference on Management of Data, ACM, 2000.

[3] 费欢,李光辉. 基于K-means聚类的WSN异常数据检测算法[J]. 计算机工程, 2015,41(7):124-128.

[4] 巴建军. 基于DBSCAN算法的异常检测方法研究[D]. 天津:中国民航大学,2019.

[5] 肖勇,郑楷洪,余忠忠,等. 基于三次指数平滑模型与DBSCAN聚类的电量数据异常检测[J]. 电网技术, 2020,44(3):1099-1104.

[6] 余翔,陈国洪,李霆,等. 基于孤立森林算法的用电数据异常检测研究[J]. 信息技术,2018,42(12):96-100.

[7] Hongyan Gang, Bo Liu, Peng Cui, et al. An Outlier Detection Algorithm for Electric Power Data Based on DBSCAN and LOF[C]// In book: Proceedings of the 9th International Conference on Computer Engineering and Networks, 2020:1097-1106.

[8] Alex Rodriguez, Alessandro Laio. Clustering by Fast Search and Find of Density Peaks[J]. Science, 2014, 344(6191):1492-1496.

[9] 刘凤魁, 邓春宇, 王晓蓉,等. 基于改进快速密度峰值聚类算法的电力大数据异常值检测[J]. 电力信息与通信技术, 2017, 15(6):36-41.

[10] 李航. 基于LOF的快速密度峰值聚类的电力数据异常值检测方法研究[D]. 兰州:兰州理工大学, 2019.

作者简介:

张哲敏(1997),男,硕士研究生,主要研究方向为电力大数据挖掘;

李琪林(1973),男,教授级高级工程师,从事营销电能计量和电力信息通信技术服务与研究工作;

严 平(1966),男,高级工程师,从事营销电能计量和高压设备技术管理与研究工作;

成贵学(1971),男,博士,副教授,硕士生导师,研究方向为过程自动化、电力系统设备控制与检测、电力信息化及新能源。

(收稿日期:2021-01-15)

(上接第5页)

[11] S Kang, S Yang, H Kim. Non-intrusive Voltage Measurement of Power Lines for Smart Grid System Based on Electric field Energy Harvesting[J]. Electronics Letters, 2017,53(3):181-183.

[12] S Kang, J Kim, S Yang, et al. Electric Field Energy Harvesting under Actual Three-phase 765 kV Power Transmission Lines for Wireless Sensor Node[J]. Electronics Letters,2017,53(16):1135-1136.

[13] 苏超,黄绍川,吴细秀,等. 高压输电线路感应取电试验研究与分析[J]. 农村电气化,2019(1):19-22.

[14] M Zhu, A Reid, S Finney, et al. Energy Scavenging Technique for Powering Wireless Sensors[C]//2008 International Conference on Condition Monitoring and Diagnosis,2008:881-884.

[15] M Zhu, M D Judd, P J Moore, et al. Energy Harvesting Technique for Powering Autonomous Sensors within Substations[C]//2009 International Conference on Sustainable Power Generation and Supply,2009:1-5.

[16] M Zhu, P C Baker, N M Roscoe, et al. Alternative Power Sources for Autonomous Sensors in High Voltage Plant[C]//2009 IEEE Electrical Insulation Conference, 2009:36-40.

[17] 王黎明,李宗,孟晓波,等. 一种交流电场无线取能电源的优化设计[J]. 高压电器,2020,56(5):121-127.

作者简介:

倪 源(1997),女,硕士研究生,研究方向为电力电子与电力传动。

(收稿日期:2021-04-05)