

基于 AP – LOF 离群组检测的配电网连接验证

司方远¹, 韩英华², 赵强³, 汪晋宽¹
(1. 东北大学 信息科学与工程学院, 辽宁 沈阳 110819; 2. 东北大学秦皇岛分校 计算机与通信工程学院, 河北 秦皇岛 066004;
3. 东北大学秦皇岛分校 控制工程学院, 河北 秦皇岛 066004)

摘 要: 现有配电网连接验证工作将可疑异常值视为具有二元属性的独立个体, 因此难以有效识别和验证具有高度内在相关性的局部离群组. 针对这一问题, 提出了基于 AP – LOF 离群组检测的配电网连接验证方法. 通过引入近邻传播(affinity propagation, AP)聚类方法, 将待校验台区用户聚类为多簇, 并基于局部离群因子(local outlier factor, LOF)算法对所有簇心进行离群点检测, 从而准确识别出台区内的离群组用户. 以某电力公司实际用户电压数据进行算例分析, 结果证明了 AP – LOF 算法在配电网连接验证中的适用性和有效性.

关 键 词: 电压数据分析; 配电网连接验证; 局部离群组检测; 近邻传播聚类; LOF 算法

中图分类号: TP 277 **文献标志码:** A **文章编号:** 1005-3026(2020)08-1070-05

Verification of Distribution Network Connectivity Based on AP-LOF Outlier Group Detection

SI Fang-yuan¹, HAN Ying-hua², ZHAO Qiang³, WANG Jin-kuan¹
(1. School of Information Science & Engineering, Northeastern University, Shenyang 110819, China; 2. School of Computer & Communication Engineering, Northeastern University at Qinhuangdao, Qinhuangdao 066004, China; 3. School of Control Engineering, Northeastern University at Qinhuangdao, Qinhuangdao 066004, China.
Corresponding author: HAN Ying-hua, E-mail: yhhan723@126.com)

Abstract: In the existing methods for the verification of distribution network connectivity, the suspicious outliers are usually regarded as independent individuals with binary attributes, which is difficult to effectively identify and validate local outlier groups which are correlated with each other. Therefore, a verification method for distribution network connectivity is proposed based on AP-LOF outlier group detection. Users are clustered into multiple clusters by introducing affinity propagation (AP) clustering, and all of the cluster centers are then detected by the local outlier factor (LOF) algorithm. In this way, the outlier groups can be accurately identified. The actual user voltage data of a power company are used in the case study, and the results demonstrate the applicability and effectiveness of the AP-LOF algorithm in the verification of distribution network connectivity.

Key words: voltage data analysis; distribution network connectivity verification; local outlier group detection; affinity propagation clustering; LOF algorithm

电力公司通常利用电网地理信息系统 (geographical information system, GIS) 获得准确的电气设备信息及拓扑结构连接信息, 进行低压配电网的运行、维护及故障响应等工作^[1]. 然而, 电网 GIS 难以实时更新, 其记录的拓扑信息与实际的连接结构往往存在一定差异^[2], 而传统的校验工作依赖于较大的人力、物力资源和成本投入. 为此, 基于高级量测体系 (advanced metering infrastructure, AMI) 的配用电大数据特征为提高 GIS 精度提供了替代解决方案^[3-4]. 其中, 基于用

户电压时间序列数据相关性分析的离群点检测是一项重要的配电网数据挖掘技术。

在低压配电网中,由于用户负荷的不确定性,各节点电压呈现不规则波动。电气距离比较近的用户负荷,其电压曲线波动性相关度较高,反之相关度较低^[5]。文献[6]通过计算不同用户智能电表电压曲线之间的相关系数来校验用户所属台区的正确性。文献[7]基于用户电压数据的相关特性及其稀疏马尔可夫随机场景描述,提出一种面向特定区域配电网的拓扑重构算法。然而,针对大规模、复杂的低压配电网,仅通过电压数据相关性分析难以实现快速、大批量、自动化的拓扑结构校验。文献[8]提出采用离散 Fréchet 距离和剪辑近邻算法进行低压配电网拓扑结构校验;通过定义待校验用户与所在台区其他用户、相邻台区所有用户之间的智能电表电压曲线的离散 Fréchet 距离,运用剪辑近邻算法检验连接关系是否正确。但是,当同一台区用户电气距离分布不均匀时,基于距离度量的校验方法难以确定用户归属正确性的阈值。

针对上述问题,文献[9]提出了对离群程度定量分析的局部离群因子(local outlier factor, LOF)方法。但是,离群点概念本身存在一定的局部特性,与特定电气距离内的用户分布密切相关。为此,文献[10]提出采用高斯核密度函数改进 LOF 算法,使其具有较稳定的判定阈值,且能够处理非均匀分布的用户用电数据集。尽管如此,用户的 LOF 值严重依赖于 k -距离值的选取。一方面,当配电网台区中存在局部电压曲线相似性较高的离群组用户时,若 k -距离选择过小,则离群组用户的 LOF 值趋近于 1,从而被识别为正常用户;另一方面,若 k -距离选择过大,则相距较远的正常用户易被误判为离群用户,从而降低了检测的准确率。

为了在有效识别配电网离群组用户的同时保障检测准确率,文中通过聚类分析与 LOF 算法进行低压配电网拓扑结构校验。引入近邻传播^[11](affinity propagation, AP)聚类算法,将待校验台区用户按电压曲线相似性程度划分为多个簇,通过对各簇簇心进行基于 LOF 算法的离群点检测,有效地识别出所属台区错误的用户组,并将该组用户电压数据置于附近台区下进行验证,从而实现快速低压配电网的拓扑结构校验。

1 基于 AP-LOF 的离群组检测方法

1.1 LOF 算法的相关定义

LOF 离群点检测方法主要利用 k 近邻算法,

通过计算每个样本数据点所处的局部邻域内的异常程度来确定是否离群。通常采用皮尔逊相关系数 $r(x, y) \in [-1, 1]$ 作为不同用户电压曲线的相似性度量,表示如下:

$$r(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 (y_i - \bar{y})^2}}. \quad (1)$$

式中: x_i 和 y_i 分别表示两个不同用户在 i 时刻的采样数据; \bar{x} 和 \bar{y} 分别表示两个用户电压时间序列数据的均值。由此可知, $r(x, y)$ 的绝对值(即距离)越趋近于 1,表明两个用户电压曲线的相似性越强。由此,给出以下定义,对于任意配电网用户对象 p 有:

1) k 距离: 用户集 D 中与用户对象 p 相关系数绝对值最小(距离最近)的第 k 个用户 q_k (由小到大排列)与 p 的距离:

$$k_{\text{dis}}(p) = \max \{ |r(p, q)|, q \in D_k(p) \}. \quad (2)$$

式中, $D_k(p) \subseteq D$ 表示与用户对象 p 距离最近的 k 个用户的集合。

2) k 距离邻域: 用户集 D 中与用户对象 p 的距离不小于 k 距离的用户构成的集合:

$$N_k(p) = \{ |r(p, o)| \geq k_{\text{dis}}(p), o \in D \setminus \{p\} \}. \quad (3)$$

3) 可达距离: 对任意用户对象 o' , 若 $o' \in D_k(p)$, 则 o' 与 p 之间的可达距离为 p 的 k 距离 $k_{\text{dis}}(p)$, 如图 1 所示; 否则, 可达距离为二者之间的相关系数绝对值:

$$\text{reachdis}(p, o') = \max \{ |r(p, o')|, k_{\text{dis}}(p) \}. \quad (4)$$

4) 局部可达密度: p 到其邻域内所有用户的平均可达距离的倒数:

$$\text{lrd}(p) = 1 / \left(\frac{\sum_{o' \in N_k(p)} \text{reachdis}(p, o')}{|N_k(p)|} \right). \quad (5)$$

5) 局部离群因子: p 的局部可达密度相对邻域内所有用户局部可达密度的平均值的倒数。局部离群因子表征用户对象 p 的离群程度:

$$\text{LOF}(p) = \frac{\sum_{o' \in N_k(p)} \frac{\text{lrd}(o')}{\text{lrd}(p)}}{|N_k(p)|}. \quad (6)$$

显然, p 的局部可达密度越低, 且 o' 的局部可达密度越高, 则 $\text{LOF}(p)$ 的值越大。据此, 可以有效识别用户集中的离群用户个体。

由图 1 可知, p 为远离用户集的一个用户, 设 $k=3$, 则 $D_k(p) = \{q_1, q_2, q_3\}$, 用户对象 p 的 k 距

离为 $k_{\text{dis}}(p) = |r(p, q_3)|$. 在 p 的 k 距离邻域内计算可达距离, 若 $o' \in D_k(p)$, 则 $\text{reachdis}(p, o') = k_{\text{dis}}(p)$; 否则, $\text{reachdis}(p, o') = |r(p, o')|$. 由此可以得到 p 的局部可达密度, 并进一步根据式(6) 计算得到 $\text{LOF}(p)$.

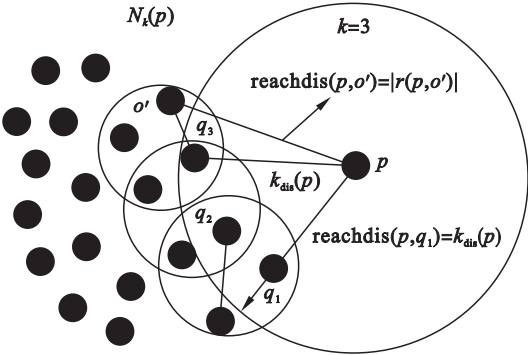


图 1 LOF 离群点检测方法原理示意图
Fig. 1 Schematic diagram for LOF outlier detection method

然而实际的配电网用户数据集通常不均匀分布, 正常用户和离群用户各自呈现相关度较高的局部簇拥, 如图 2 所示. 在这种情况下, 采用传统的 LOF 离群点检测方法将出现两种检测结果: ①将离群组用户识别为正常用户; ②将距离较远的正常用户识别为离群用户.

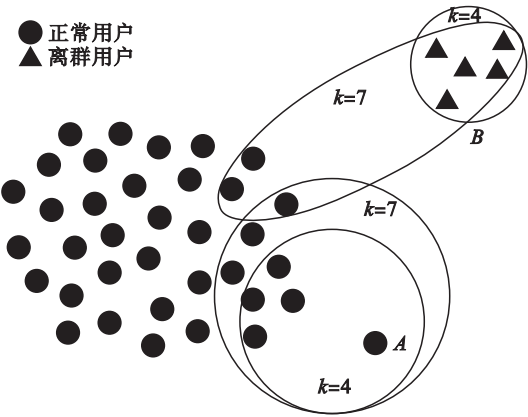


图 2 存在离群组用户情况的 LOF 离群点检测
Fig. 2 LOF outlier detection in the presence of an outlier group

由图 2 可知, 数据集呈现正常用户组和离群用户组两簇局部相关度较高的子集, 其中 A 为距离正常用户组相对较远的正常用户个体, B 为一组包含 5 个用户的离群组. 当 $k \leq 4$ 时, 由于 B 组局部簇拥, 导致组内用户间局部可达密度均较高, 根据式(6)可知, B 组内用户的 LOF 值均趋近于 1, 被识别为正常用户, 因此无法有效识别该离群组. 进一步增加 k 值, 取 $k=7$, 此时 B 组内用户与 3 个正常用户构成一组检测对象, 并形成局部可

达密度差异, B 组内用户的 LOF 值大于 1, 因此可将 B 组识别为 5 个独立的离群用户. 此外, 随着 k 值的增加, A 相对于其他正常用户的局部可达密度进一步减小, 则 A 的 LOF 值可能大于 1, 并被识别为离群用户. 另外, 对 5 个独立的离群用户需要进行 5 次重复性验证, 这个过程严重限制了验证效率.

1.2 AP-LOF 离群组检测方法

基于以上分析, 配电网连接验证工作亟需一套既可以准确识别离群组用户又能够保留离群组内相关性特征的检测算法, 为此, 引入 AP 聚类算法. 作为一种高效的基准聚类算法, 其核心思想是以数据点偏离簇中心误差最小化为目标条件来寻找一组聚类中心. 相较于其他聚类算法, AP 聚类算法无需人工设定聚类簇数, 而是依靠“信息传递”机制迭代循环寻找最优聚类簇数. 这种信息传递机制中主要包含两类信息: 吸引度和归属感. 通过数据点之间的相似度迭代计算更新数据点的吸引度矩阵 $R = (r_{ij})$ 和归属感矩阵 $A = (a_{ij})$, 依据 $r_{ii} + a_{ii} > 0$ 是否成立来判断是否为聚类中心. 当迭代次数超过最大值或是连续多次迭代计算质心不发生变化时, 终止迭代, 同时将其余的数据点分配到相应的簇中.

对于数据集 $X = \{x_1, x_2, \dots, x_i\}$, 其中 x_i 代表一个数据点, $x_i = [x_{i1}, x_{i2}, \dots, x_{id}]$, d 表示数据点的维度. AP 聚类算法计算步骤如下:

1) 计算相似度矩阵 $S = (s_{ij})$. 利用式(1)各数据点之间的皮尔逊相关系数表示相似度, 即

$$s_{ij} = \begin{cases} r(x_i, x_j), & i \neq j; \\ P(i), & i = j. \end{cases} \quad (7)$$

式中, $P(i)$ 为初始偏向度, 表示数据点 i 作为聚类中心的偏向程度, 通常取数据点之间相似度最小值. $P(i)$ 越大, 表示该数据点越有可能作为聚类中心.

2) 信息的相互传递. 吸引度矩阵 R 的元素 r_{ij} 表示从用户 i 到用户 j 的信息, 具体表示数据点 x_j 作为数据点 x_i 聚类中心的适合程度; 归属感矩阵 A 的元素 a_{ij} 表示从用户 j 到用户 i 的信息, 具体表示数据点 x_i 选择数据点 x_j 作为聚类中心的适合程度.

$$r_{ij} = s_{ij} - \min_{j' \neq j} \{a_{ij'} + s_{ij'}\}. \quad (8)$$

$$a_{ij} = \begin{cases} \min \{0, r_{ji} + \sum_{i' \in \{i, j\}} \max \{0, r_{ji'}\}\}, & i \neq j; \\ \sum_{i' \neq j} \max \{0, r_{ji'}\}, & i = j. \end{cases} \quad (9)$$

然而,式(8)和式(9)存在一定的振荡,导致收敛速度较慢,因此引入阻尼因子 $\lambda \in (0,1)$,则信息传递过程可以表示为

$$r_{ij}^{t+1} = (1 - \lambda) r_{ij}^t + \lambda r_{ij}', \quad (10)$$

$$a_{ij}^{t+1} = (1 - \lambda) a_{ij}^t + \lambda a_{ij}'. \quad (11)$$

3) 确定聚类中心. 如果 \mathbf{x}_j 要作为 \mathbf{x}_i 的聚类中心,那么 j 需满足

$$j = \operatorname{argmax} \{ a_{ij} + r_{ij} \}. \quad (12)$$

即当 i 一定时,使 $a_{ij} + r_{ij}$ 最大的 j 值.

4) 终止迭代. 当达到规定最大迭代次数或经多次迭代聚类中心未发生变化时,算法结束.

利用 AP 聚类算法,迭代计算最佳的聚类中心,使数据点偏离簇中心误差最小化,确保将相关系数较高的用户归为一簇. 簇心计算如下:

$$v_{ij}' = \frac{\sum_{k=1}^{|D_i|} v_{kj}}{|D_i|}. \quad (13)$$

式中: v_{kj} 表示第 i 簇内用户 k 在 j 时刻的电压数据; D_i 表示第 i 簇内的用户集合. 计算得到簇心集 $v_i' = \{v_{i1}', v_{i2}', \dots, v_{in_j}'\}$. 对簇心集进行 LOF 离群程度校验,当簇心 i 被识别为离群点时,表示第 i 簇用户集为原数据集的离群组. 基于 AP-LOF 离群组检测的配电网连接验证方法具体步骤如下:

步骤 1 读取台区用户电压时间序列数据集,并根据式(7)计算用户间相似度矩阵 \mathbf{S} ,同时初始化吸引度矩阵 \mathbf{R} 和归属度矩阵 \mathbf{A} ;

步骤 2 迭代计算聚类中心是否变化,当达到最大迭代次数或聚类中心不再变化时,终止迭代;

步骤 3 依据聚类中心将用户划分至各簇,并利用式(13)计算簇心集;

步骤 4 利用式(1)计算各簇心间的皮尔逊相关系数,并根据式(6)计算簇心局部离群因子,从而得到原数据的离群组用户.

2 结果与讨论

原数据集来自某市电力公司采集的台区 A 内 324 个用户的电压-时间序列数据,时间范围为 2017 年 4 月 1 日至 4 月 30 日,采集间隔为 1 h. LOF 算法的邻域用户数 k 分别选取用户总数的 5%~20%,用阈值 g 表示.

基于 LOF 算法的台区 A 离群点检测结果如图 3 所示. 当 $g=5\%$ 时,所有用户的 LOF 值均接近于 1,此时所有用户均被识别为正常用户;当 $g=10\%, 15\%, 20\%$ 时,可以看到,用户 309~324

的 LOF 值大于 1,同时其他部分用户 LOF 值亦偏离 1,此时用户 309~324 及其他部分用户被识别为离群点. 这是因为检测结果受阈值 g 的影响,无法确定一个合适的离群点判别阈值,导致判别结果模糊,从而无法准确识别离群用户. 此外,从检测结果无法得出离群用户间的相关性特征,影响验证效率.

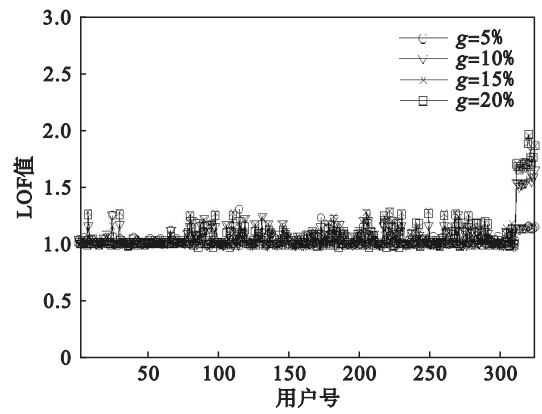


图 3 基于 LOF 算法的台区 A 离群点检测结果
Fig. 3 Outlier detection result of transformer A based on the LOF algorithm

采用 AP-LOF 离群组检测方法,台区 A 内 324 个用户数据集的 AP 聚类结果如图 4 所示. 可以看到,原数据集被划分为 29 簇,其中簇 1~28 相对集中于两个区域,簇 29 (左下角) 偏离其他簇.

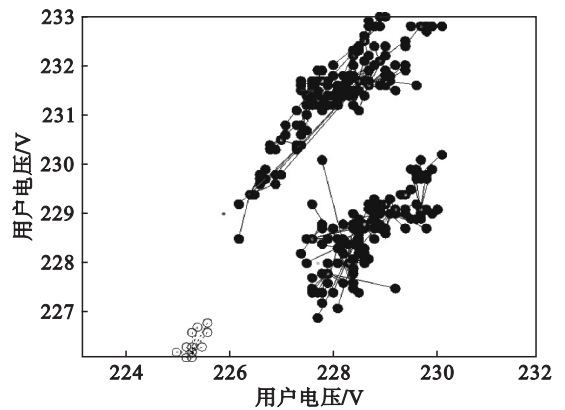


图 4 台区 A 用户数据集的 AP 聚类结果
Fig. 4 AP clustering result of the user data in transformer A

进一步对各簇簇心进行 LOF 离群程度校验,结果如图 5 所示. 其中,簇心 29 的 LOF 值远大于 1,其他簇心的 LOF 值围绕 1 小范围波动,这表明簇心 29 的局部密度低于其他簇心,即第 29 簇内的用户整体偏离于其他用户,为离群组用户.

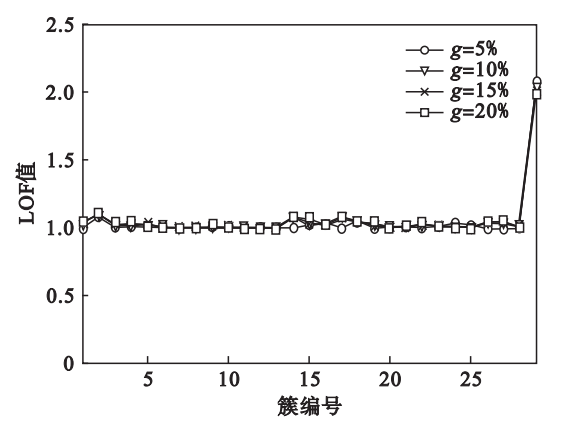


图5 基于AP-LOF算法的台区A离群组检测结果
Fig. 5 Outlier group detection result of transformer A based on the AP-LOF algorithm

针对两种算法的现场实地考察结果如表1所示,其中TP,FP,FN分别表示检索到的离群用户个数、检索到的离群用户中实际正常的用户个数,以及未检索到的离群用户个数^[12]。由表1可知,当 $g=5\%$ 时,采用LOF算法无法校验出离群用户。当 $g=10\%$ 时,可以校验出19个离群个体,其中包括3个实际正常连接的用户。因此,当选取的邻域用户数 k 小于实际离群组用户数时,LOF算法无法校验出离群用户;当 k 大于离群组用户数时,可以校验出离群用户,但发生误判的用户数随 g 的增加而增多。而采用AP-LOF算法可以准确地校验出离群组用户309~324且无误判用户,进一步验证了本文方法的有效性。

表1 两种算法离群用户检测的对比结果
Table 1 Comparison result of two algorithms for outlier detection

$g/\%$	LOF 算法			AP-LOF 算法		
	TP	FP	FN	TP	FP	FN
5	0	0	16	16	0	0
10	19	3	0	16	0	0
15	23	7	0	16	0	0
20	28	12	0	16	0	0

3 结 语

本文提出一种基于AP-LOF离群组检测的配电网连接验证方法,通过引入AP聚类算法保留了用户组内的相关性特征,并将聚类簇心用于基于LOF的离群程度校验。实验结果表明,与传统的LOF算法相比,AP-LOF算法避免了判定

阈值对检测结果的影响,能够准确有效地对台区内的离群组用户进行校验,提高了配电网连接验证效率。

参考文献:

[1] Short T A. Advanced metering for phase identification transformer identification, and secondary modeling[J]. *IEEE Transactions on Smart Grid*, 2013, 4(2): 651-658.

[2] Luan W, Sharp D, LaRoy S. Data traffic analysis of utility smart metering network[C]//IEEE Power Energy Society General Meeting. Vancouver: IEEE, 2013. DOI: 10. 1109/PESMG. 2013. 6672750.

[3] Wang Y, Qiu H, Tu Y, et al. A review of smart metering for future Chinese grids[J]. *Energy Procedia*, 2018, 152: 1194-1199.

[4] 栾文鹏, 余贻鑫, 王兵. AMI 数据分析方法[J]. *中国电机工程学报*, 2015, 35(1): 29-36.
(Luan Wen-peng, Yu Yi-xin, Wang Bing. AMI data analytics [J]. *Proceedings of CSEE*, 2015, 35(1): 29-36.)

[5] Luan W P, Peng J, Maras M, et al. Smart meter data analytics for distribution network connectivity verification[J]. *IEEE Transactions on Smart Grid*, 2015, 6(4): 1964-1971.

[6] Luan W P, Peng J, Maras M, et al. Distribution network topology error correction using smart meter data analytics [C]//IEEE Power Energy Society General Meeting. Vancouver: IEEE, 2013. DOI: 10. 1109/PESMG. 2013. 6672786.

[7] Bolognani S, Bof N, Michelotti D, et al. Identification of power distribution network topology via voltage correlation analysis[C]//IEEE Conference on Decision and Control. Florence: IEEE, 2013: 1659-1664.

[8] 耿俊成, 张小斐, 郭志民, 等. 基于离散Fréchet距离和剪辑近邻法的低压配电网拓扑结构校验方法[J]. *电测与仪表*, 2017, 54(5): 50-55.
(Geng Jun-cheng, Zhang Xiao-fei, Guo Zhi-min, et al. Topology verification of low-voltage transformer areas based on discrete Fréchet distance and editing nearest-neighbors method[J]. *Electrical Measurement & Instrumentation*, 2017, 54(5): 50-55.)

[9] Breunig M M, Kriegel H P, Ng R T, et al. LOF: identifying density-based local outliers[C]//Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2000: 93-104.

[10] 孙毅, 李世豪, 崔灿, 等. 基于高斯核函数改进的电力用户用电数据离群点检测方法[J]. *电网技术*, 2018, 42(5): 1595-1604.
(Sun Yi, Li Shi-hao, Cui Can, et al. Improved outlier detection method of power consumer data based on Gaussian kernel function[J]. *Power System Technology*, 2018, 42(5): 1595-1604.)

[11] Wang C, Lai J, Suen C Y, et al. Multi-exemplar affinity propagation[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2013, 35(9): 2223-2237.

[12] Jiang M, Faloutsos C, Han J. CatchTartan: representing and summarizing dynamic multicontextual behaviors [C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2016: 945-954.