基于云计算的电力大数据分析算法研究

任敬佩,邢敬创,白晓伟
(西安思安云创科技有限公司,陕西 西安 710000)

摘 要:为了解决传统聚类算法检测准确性低,复杂性高不适于电力大数据异常值检测的问题,提出了一种在云计算平台上基于距离三角不等式的类轮廓聚类算法处理电力异常数据。文中首先根据三相不平衡、功率等计算分析要求,针对源数据进行降维与清洗处理,然后,利用距离三角不等式的类轮廓聚类算法计算与识别处理后的电力运行数据,最终,利用轮廓系数、簇密度、时效性和正确率为评价指标确定算法的优劣性,快速检测出孤立点和噪声数据,减少了 I/O 以及网络传输的消耗。该算法能够有效处理任意形状的簇,一定程度上防止出现线形类或蛇形类,从而确定的最优聚类数处理企业电能质量曲线,针对不符合要求的数据,认为相应电力数据点为电力数据异常值。该算法通过某企业的三相电流、三相电压与功率数据进行聚类分析,验证了该算法的可行性和有效性。

关键词:距离三角不等式;类轮廓;轮廓系数;时效性;正确率;噪声点
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2021)0047-05

Research on Algorithm of Big Data Analysis of Power Based on Cloud Computing

REN Jing-pei, XING Jing-chuang, BAI Xiao-wei
(Xi'an Sian Yunchuang Technology Co., Ltd., Xi'an 710000, China)

Abstract: In order to solve the problem of low accuracy and high complexity of traditional algorithm which is not suitable for abnormal value detection of large power data, a clustering algorithm based on distance trigonometric inequality and class profile is proposed for large power data. Firstly, according to the calculation and analysis requirements of three-phase unbalance and power, the dimension reduction and data cleaning are carried out. Secondly, clustering algorithm of class profile and distance triangle inequality calculate and recognize the processed power operating data. Finally, the silhouette coefficient, cluster density, time-efficient and correctness rate are used to determine the pros and cons of the algorithm, which quickly detects outliers and noise data and reduces the computational redundancy and operation time. The proposed algorithm can effectively process clusters of any shape, and prevent linear or serpentine from appearing to a certain extent. The optimal number of clusters determined is used for cluster processing of enterprise power quality curve. For data that does not meet the requirements, the corresponding power is considered data points are abnormal values of power data. The algorithm uses the three-phase current, three-phase voltage and power data of a certain company to perform cluster analysis, which verifies its feasibility and effectiveness.

Key words: distance triangle inequality; class profile; silhouette coefficient; time-efficient; correctness rate; noise data

0 引 言

随着物联网的迅速发展,在电力系统的运行与维护过程中,伴随着数据量的越来越大,针对电力数据的质量分析要求越来越高[1];然而设备常受到电压、电流、温度、负荷等多种因素的影响[2-4],造成数据的异常,致使在企业在获得有效数据的过程中,常常与预测的目标不一致,造成了运维人员维护与优化控制的困难,因而对于电力大数据的异常检测往往显得至关重要。当前关于电力大数据的异常点检测都是基于传统的密度算法[5-8]、K-means 算法[9-10]、模糊 C 均值算法[11]等常用聚类算法,虽然这些算法从某种程度上提高数据质量的处理,但是其处理速度与正确性并不能满足当前大数据量的电力数据处理需求。

鉴于以上改进后的聚类算法的问题,文章利用单玉双等[12]在 K-means 算法引进的三角不等式原理与孟海东等[13]类轮廓定义的原理的基础上,设计了基于类轮廓与三角不等式的聚类数据处理方法。文章主要基于开源云计算平台,利用 Flink 分布式框架,融合了距离三角不等式定理与类轮廓的定义,结合电力大数据的特性,进一步实现了 K-means 算法在聚类分析过

收稿日期:2020-12-23
作者简介:任敬佩(1987-),男,硕士,研究方向为软件工程、大数据分析。

程中的改进;并通过针对某企业采集的电压、电流与功率进行计算分析,依据电力大数据特点,提出一种基于类轮廓与距离三角不等式的聚类的电力大数据异常值检测算法,距离三角不等式的类轮廓聚类算法(research on clustering method based on class profile and trigonometric inequality, CP-TRI-K-means)。其基本思想为优化初始聚类中心,减少了 I/O 以及网络传输的消耗,改善电力大数据异常值检测复杂度,提高了检测的准确性与检测效率。经实例分析,所提方法可有效检测电力大数据异常值。

1 距离三角不等式的类轮廓算法基本思想

基于云计算平台的 Flink 框架,利用传统的 K-means 算法与类轮廓定义方式和距离三角不等式定理相结合,实现 K-means 算法的优化,提出了基于距离三角不等式的类轮廓聚类算法。该算法首先基于电力数据的处理规则进行数据预处理,然后通过类轮廓定义的方式寻找所有的边界点,并把所有边界点作为初始的聚类数;然后根据给定的阈值和簇的密度选择出最优的聚类个数,把剩下非边界点作为独立的簇,通过距离三角不等的原理快速计算出非边界的点所属簇,同时计算该点到所属簇的最短距离是否满足给定阈值,满足则标记为该簇所属点,反之标记为异常点;依次类推,计算出所有点所属簇,如果点所属簇不满足给定密度,则标记该簇的所有点为异常点。

1.1 数据预处理

(1)数据补缺。

针对电压的数据预处理,由于峰平谷时段,整体变化不大,在进行数据补缺时,按照该时间点最近一段时间内上下变化的平均值进行补缺;针对电流与功率,由于峰平谷时段整体变化很大,根据该点所在的峰平谷时间段,求取该点在不同时段的平均值进行补缺。

(2)数据降维。

根据电流与电压的三相不平衡计算分析要求,分别计算三相电流与三相电压的最大值与最小值,进行数据降维。

1.2 类轮廓的定义

定义 1(样本类边界点):设在欧几里得空间中存在 N 维变量的样本集合,样本集合中每个样本点视为分布在 N 维空间中的独立空间点,以每个空间点作为基点,如果在其 $2n$ 个象限内存在一个样本点到基点的近似性度量(定义 3)符合给定 δ 阈值,则该点不是边界点,反之则标记该点为样本类边界点。下列以二维空间为例。

在图 1 中的 C_1 簇里,样本 S_1 在 1、2、3、4 象限内都能找到至少一个点在它的距离阈值 δ 的范围内,则 S_1 不是边界点;同理,S_2 在其 1、4 象限内找不到样本,所以 S_2 是边界点;S_3 在第 3 象限内找不到样本,所以 S_3 也是边界点;图 2 中有两个簇,C_1 簇和 C_2 簇,S_1 是簇 C_1 的边界点,S_2 是簇 C_2 的边界点。

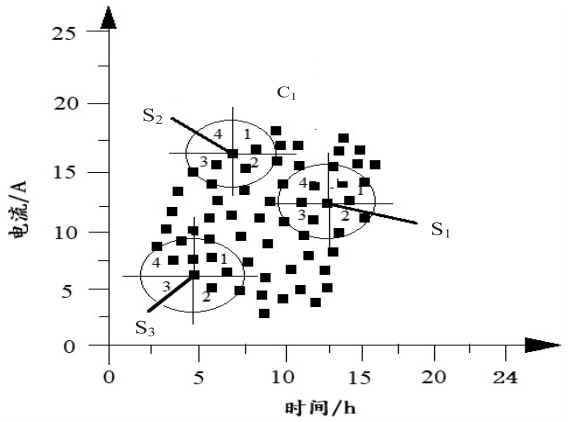

图 1 单个簇 C_1 边界点展示

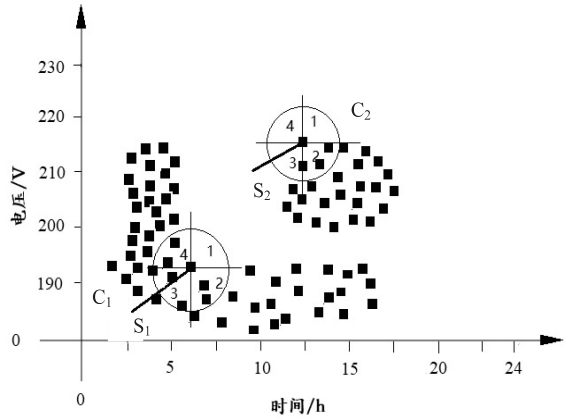

图 2 多簇 C_1 和簇 C_2 边界点展示

定义 2(类轮廓):针对上述的所有边界点形成的边界集合,定义为类轮廓。

定义 3(近似度度量):样本点到每个簇内所有点的最短距离 Min($d_1(x,y), d_2(x,y), \cdots, d_n(x,y)$)。

$$d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$ (x,y 分别表示 N 维空间的数据点) (1)

1.3 距离三角不等式原理

定义 4:设在欧几里得空间中存在 3 个样本点 X、C_1、C_2,其中 X 为欧几里得空间中的任意样本点,C_1 和 C_2 为两个簇中心点,任意两样本点之间的欧氏距离(公式 1)计算分别为 D_1、D_2、D_3,当满足以下条件时:

(1)任意两点相互不重合;

(2)不同样本点之间距离符合三角不等原理,即 $D_1+D_2 \geq D_3$ 且 $|D_1-D_2| \leq D_3$;

(3)$2 * D_3 \leq D_1$。

那么,样本点 X 标记为簇 C_1 的点。

1.4 算法实现流程

(1)针对电压、电流与功率数据按照电力数据特

性(三相不平衡计算原则,及不同时段数据在峰平谷的波动性)进行预处理,主要包括数据补缺与降维;

(2)根据类轮廓的定义计算所有的边界点,并定义近似度度量阈值为 δ;

(3)根据距离三角不等式计算出所有点到所有簇的距离;

(4)根据边界点计算出所有边界点所属簇,从而确定 K 的个数;

(5)电力数据点到某个簇的某个点的距离小于或等于规定阈值 δ,则标记为该簇的数据点;

(6)如果某个点属于多个簇 C_i,则根据到 C_i 内最近点的距离为依据,判断该数据点属于那个簇;

(7)如果某个点到所有簇的距离都不符合规定的阈值,则定义该点为一个独立的簇;

(8)根据上述过程所生成的簇,密度小于规定阈值 ρ 的所有独立簇合并并标记为一个簇,同时判定为异常簇,包含所有噪声或孤立数据。

2　CP-TRI-K-means 实现结果分析

2.1　实验平台、测试数据集和评价指标

文章所有实验环境搭建的平台的组成为:5 台 2 GHz Inter Xeon CPU、64G 内存 PC 构成,操作系统均为 Centos7.5,Hadoop 版本选用 2.7.1;Flink 版本选用 1.9,Java 开发包为 JDK1.8 版本,程序开发工具为 Eclipse-standard-kepler-SR1-linux,算法使用 Java 与 Scala 混合实现。

实验数据集采用了某企业不同类型的电力数据集,分别使用了 30 天,60 天,120 天,180 天的数据集来验证算法的时效性与有效性。

在实验中,为了测试距离三角不等式的类轮廓聚类算法的性能,文章采用了以下评价指标:轮廓系数、簇密度、时效性和正确率,来表达算法的有效性。

2.2　实验结果

为了验证该算法的时效性,文章通过针对不同类型电力数据分别进行距离三角不等式聚类计算与传统 K-means 算法进行对比;距离三角不等式类轮廓聚类算法给定功率、电流、电压距离阈值分别为 $\delta=15$, $\delta=15$, $\delta=8$,密度全为 $\rho=20$。K-means 算法初始中心随机选择,因而对初始中心点进行了 10 次随机选择,同时进行了 10 次迭代运算,最终结果利用 10 次实验结果的平均值来获得。

表 1　CP-TRI-K-means 与 K-means 算法时效性分析

数据类型	数据集/天	CP-TRI-K-means/ms	K-means/ms
电压	30	2.5	3.5
	60	3	6
	90	4	8.5
	120	7	9.5
	150	8.5	11
	180	9	13
电流	30	2	3
	60	5	6
	90	7	8
	120	8	9
	150	8.5	10
	180	9	12
功率	30	2.5	3
	60	3	4.5
	90	4	6.5
	120	6	7.5
	150	7.5	9
	180	8.4	10.5

由表 1 可知,两种算法分别从不同数据量维度与不同类型数据的角度分析了电力数据基于上述两种算法计算的时效性,基于云计算平台的距离三角不等式的类轮廓聚类算法比基于云计算平台的传统的 K-means 算法的计算速度要更快。其原因:距离三角不等式的类轮廓聚类算法使用三角不等式时,不需要每个点进行计算,很大程度上节约了时间,同时两种算法当数据量达到一定程度时计算速度相对比较稳定,基于云计算的优势可以在此充分体现。

为了证明距离三角不等式的类轮廓聚类算法与 K-means 算法在不同时间段内对不同电力数据集的划分有效性,下面分别从功率、电流和电压三种类型数据进行了分类测试,其中有效时间段的划分分别为标记 1、2、3、4 为所属簇,具体分类如表 2 所示。

表 2　CP-TRI-K-means 与 K-means 算法分类

数据类型	标记	CP-TRI-K-means(时间段)	K-means(时间段)
电压	标记 1	23:40-05:05,05:40-06:35	00:20-04:05,05:20-06:15
	标记 2	07:10-14:10,19:45-22:13	06:30-15:10,19:25-22:40
	标记 3	14:35-18:58	14:10-17:20
	标记 4	其他	其他

续表 2

数据类型	标记	CP-TRI-K-means(时间段)	K-means(时间段)
电流	标记 1	23:23-05:15,05:30-06:55	23:20-05:05,05:40-06:15
	标记 2	07:30-14:35,19:55-22:40	06:40-14:55,19:25-21:40
	标记 3	14:45-18:38	15:10-18:20
	标记 4	其他	其他
功率	标记 1	22:40-06:05,06:25-07:10	23:20-05:05,05:25-06:15
	标记 2	07:36-14:45,19:25-22:30	06:45-14:25,18:25-23:40
	标记 3	14:50-18:58	14:30-18:50
	标记 4	其他	其他

基于电力数据的特性,在数据进行有效划分时,以 3 ~5 个簇为最优,由于电力数据在不同时间段会形成比较鲜明的簇,例如文中所测试的电流与功率,电压相对稳定,但是在同等的时间段电压同样会有明显的波动。通过表 2 可知,CP-TRI-K-means 算法与传统的 K-means 算法分别将三种类型的数据分为了 4 个簇,整体符合峰时:07:00-11:00 与 19:00-23:00,平时:11:00-19:00,谷时:23:00-07:00 的时间段;其中簇 4 中所有点为孤立簇或异常数据点。CP-TRI-K-means 算法相对于传统的 K-means 算法来讲,基于电流、电压与功率在进行 CP-TRI-K-means 算法对比理论数据分析结果(给定的异常数据比),异常簇在整个时间段内占比相对比较少,对于分类整体影响较小,对数据的划分更有效。

2.3 实验分析评价指标

(1)时效性。

定义 5(时效性):通过设计严谨的运算顺序规则,并且每一个规则都是有效的,根据给定的数据集,判断其计算时间以及计算结果的正确性。

定义 6(正确率,(correct rate,CR)):根据给定的训练数据集,计算分类正确的数据点与所有数据点的一个比值:

$$CR = \frac{\text{正确的数据点}}{\text{所有的数据点}}(0 < CR \leqslant 1) \quad (2)$$

针对同样的训练数据,CR 越高,同时计算速度越快,表示算法的时效性越高。

(2)轮廓系数。

$$s[i] = \frac{b[i] - a[i]}{\max(a[i], b[i])}(-1 \leqslant s[i] \leqslant 1) \quad (3)$$

$a[i]$:样本点 i 在当前簇内到其他点的平均距离,$b[i]$:样本点到其距离最近簇内所有点的距离的平均值;其中 $s[i]$ 的系数越大,聚类的效果越好。

(3)簇密度。

定义 7(簇密度):定义一个簇边界点数量与对应簇样本点数量的一个比值。

$$r = \frac{\text{簇所有边界点数量}}{\text{簇包含的样本点数据量}}(0 < r \leqslant 100\%) \quad (4)$$

r 值越小表示簇越密集,r 越大则簇越稀疏。该指标主要是防止在聚类的过程中只考虑每个簇所包含的数据量而忽略簇聚集程度,从而导致出现线形类或蛇形类。

(4)异常比(abnormal ratio)。

定义 8(异常比):定义一个簇内的异常数据点与簇内所有样本的数据量的一个比值。

$$AR = \frac{\text{簇内异常数据点}}{\text{簇包所有样本的数据量}}(0 < AR \leqslant 100\%) \quad (5)$$

AR 值越小表示异常点处理分析正确率越高。

2.4 算法分析对比

由表 3 与表 4 对比可知,在 CP-TRI-K-means 算法与 K-means 算法在同时获得 K 最优时,CP-TRI-K-means 算法比传统的 K-means 算法的计算的正确率高。其原因:传统的 K 均值算法在进行选择最优 K 时,迭代次数与异常点对其影响较大,而 CP-TRI-K-means 算法在基于类轮廓方法进行定义和判定时获得最优的 K 值。

表 3 CP-TRI-K-means 算法正确率指标分析

数据类型	δ	ρ	CR
电压	8	20	82%
电流	15	20	87.75%
功率	15	20	89%

表 4 K-means 算法正确率指标分析

数据类型	初始中心选择	迭代次数	CR
电压	10	10	72.5%
电流	10	10	84.5%
功率	10	10	87.5%

由表 5 对比可知,CP-TRI-K-means 算法比 K-means 的类轮廓系数和 AR 高,证明了距离三角不等式的类轮廓聚类算法聚类效果及数据处理分析具有一

定程度上的优越性,其优化了初始聚类中心 K 以及 K 的值,提高了算法处理电力大数据异常的正确率。

表 5　CP-TRI-K-means 算法的轮廓系数与 AR 指标分析

算法	数据类型	K 值	轮廓系数（平均值）	AR（平均值）
CP-TRI-K-means	电压	4	0.79	79%
	电流	4	0.83	87%
	功率	4	0.87	92.75%
K-means	电压	4	0.65	76%
	电流	4	0.73	82%
	功率	4	0.84	86.5%

由表 6 可知,簇 1、2、3 的簇密度比簇 4 的簇密度低,从一定程度上证明了簇 1、2、3 的聚类比较密集,簇 4 的密集度较小,同时根据电力数据要求计算其电压与电流三相不平衡为规范化的电力数据,可以得出其电流与电压的三相不平衡不符合要求的数据簇 4 占整体比例较大,可以从一定程度上判断出该簇为异常簇,说明该簇包含了所有的异常点数据。

表 6　CP-TRI-K-means 算法的簇密度

数据类型	簇 1(r)	簇 2(r)	簇 3(r)	簇 4(r)
电压	8.5%	7.5%	8%	10.5%
电流	5%	6%	5.5%	17%
功率	8%	7%	8%	12%

3　结束语

基于距离三角不等式的类轮廓聚类算法通过实验证明了针对在空间上进行电力大数据异常值检测的可行性和有效性,且能够快速检测出给定的异常数据点;由于在数据处理过程中,采用了先连通性原则再近似性原则和距离三角不等式原理,有效处理了任意形状的簇,减少了 I/O 以及网络传输的消耗,从而能大大减少数据量的计算,提高了异常数据处理效率和正确率。

文中提出的聚类算法,具有实用的价值,将传统的聚类方法融入其中,采用通用的轮廓系数、时效性、异常比与簇密度作为评价指标有效评估了该算法的正确性与有效性,准确地帮助用户发现数据集中的异常数据,为优化控制和发现设备缺陷提供有效的帮助。

参考文献:

[1] 颜　清. 基于云计算的电力大数据分析技术应用研究[J]. 中国管理信息化,2020,18:103-104.

[2] 郝晓弘,李亚岚. 基于 DTW 直方图的电力负荷数据聚类算法[J]. 传感器与微系统,2020,12:140-142.

[3] 周　巍,崔艳林. 基于 k-Means 算法的电网调度辅助决策平台[J]. 自动化与仪器仪表,2020,11:137-140.

[4] 严英杰,盛戈皞,陈玉峰. 基于大数据分析的输变电设备状态数据异常检测方法[J]. 中国电机工程学报,2015,35(1):52-59.

[5] 王　飞,王国胤,李智星. 一种基于网格的密度峰值聚类算法[J]. 小型微型计算机系统,2017,38(5):1034-1038.

[6] 夏庆亚. 基于密度峰值和网格的自动选定聚类中心算法[J]. 计算机科学,2017,44(B11):403-406.

[7] 陆春光,叶方彬. 基于密度峰值聚类的电力大数据异常值检测算法[J]. 科学技术与工程,2020,20(2):654-658.

[8] 荣秋生,颜君彪,郭国强. 基于 DBSCAN 聚类算法的研究与实现[J]. 计算机应用,2004,24(4):45-46.

[9] 鲍晓地,张芳芳. 大数据处理的关键技术研究[J]. 信息化建设,2013(10):49-54.

[10] 马　浩,黄　俊. 动态 k 近邻辅助多权值 Slope One 算法[J]. 计算机工程与设计,2020,41(11):3072-3077.

[11] 杨　浩,张　磊,何　潜. 基于自适应模糊 C 均值算法的电力负荷分类研究[J]. 电力系统保护与控制,2010,38(16):111-115.

[12] 单玉双,邢长征. 一种更有效的 K-means 聚类算法[J]. 计算机系统应用,2009(8):96-99.

[13] 孟海东,唐　旋. 基于类轮廓层次聚类方法的研究[J]. 计算机应用与软件,2011,28(11):119-121.