一种基于 Spark 模型的电力异常数据检测方法

朱昌敏,岳　东
(南京邮电大学 先进技术研究院,江苏 南京 210023)

摘　要:电网的信息化与智能化程度不断的提升使得电力数据量越来越大,给数据的处理和分析带来很大的困难。在智能电网大数据应用处理的过程中,数据的实时性存储、高效处理、多源异构数据的融合以及数据的可视化方面面临着严峻的挑战,需要深入对这些方面开展研究,切实发挥大数据在保障电网安全稳定运行的作用。这些异常数据的存在对现代电力系统状态的估计结果的影响是不容忽视的。现有的电力数据异常检测方法未能充分挖掘数据特征,存在计算复杂、灵活性差、精度较低等缺点。目前已有的预测算法无法满足预测速度和精度的要求,因此基于大数据计算平台,提出一种基于 Spark 的改进 ISODATA 聚类算法对异常数据进行检测与修正。实验结果表明,该方法对异常数据的检测和修正有很好的效果,降低了检测时间,有效提高了状态估计结果的准确性。

关键词:电力系统;Spark;异常数据检测;ISODATA 算法
中图分类号:TP301　　**文献标识码**:A　　**文章编号**:1673-629X(2019)01-0140-05
doi:10.3969/j.issn.1673-629X.2019.01.029

A Method for Identifying Bad Data of Power System Based on Spark

ZHU Chang-min, YUE Dong
(Institute of Advanced Technology, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

Abstract: The continuous improvement of the informatization and intellectualization of the power grid makes the power data more and more large, which brings great difficulties to the data processing and analysis. In the application and processing of smart grid big data, there exists severe challenges in real-time data storage, efficient processing, multi-source heterogeneous data integration and data visualization. It is necessary to conduct in-depth research on these aspects and give full play to the role of big data in ensuring the safe and stable operation of the power grid. The influence of these abnormal data on the state estimation of modern power system cannot be ignored. The existing abnormal detection methods of power data fail to fully exploit the data features, and have the disadvantages of complex computation, poor flexibility and low accuracy. At present, the existing prediction algorithms cannot meet the requirements of prediction speed and accuracy. Therefore, based on big data computing platform, we propose an improved ISODATA clustering algorithm based on Spark to detect and correct abnormal data. Experiment shows that the proposed method has a better effect on the detection and correction of abnormal data, reduces the detection time and effectively improves the accuracy of state estimation.

Key words: power system; Spark; abnormal data detection; ISODATA algorithm

0　引　言

传统的数据异常检测方法计算复杂,精度低,灵活性差,电力系统中的异常数据影响电力系统状态估计结果的准确性,而传统机器学习算法在处理海量高维度多类型数据时计算资源不足,虽然 Map-Reduce 框架能并行处理数据,但在迭代计算方面性能不足。对此,文中从电力异常数据检测与修正方面阐述了电力大数据与智能电网的深度融合,基于大数据 Spark 计算平台,提出一种基于 Spark 的改进 ISODATA 聚类算法对异常数据进行检测。对某一节点电力负荷数据提取出日负荷特征曲线,通过特征曲线对比,对其中异常数据进行检测并修正。以大航杯“智造扬中”电力 AI 大赛提供的扬中市高新区 1 000 多家企业的历史用电量数据进行实验,对该方法的性能进行验证[1]。

收稿日期:2018-01-16　　**修回日期**:2018-05-16　　**网络出版时间**:2018-11-15
基金项目:国家自然科学基金(51507084)
作者简介:朱昌敏(1992-),男,硕士研究生,研究方向为云计算与物联网;岳　东,教授,博导,长江学者,研究方向为智能电网大数据分析与协调控制、复杂系统与多智能体理论等。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20181115.1046.006.html

1　基于 Spark 改进的 ISODATA 聚类算法

1.1　Spark 简介

Spark 是一个通用的并行计算框架,其分布式计算框架的实现基于 MapReduce 算法模式,因此拥有 MapReduce 拥有的所有优点,并在 Hadoop 的基础上对 MapReduce 做了大量优化:快速处理,易于使用,支持查询,支持流式计算,可用性高,丰富的数据源支持。

Spark SQL、Spark Streaming、MLLib 和 GraphX 等组件组成了 Spark 丰富的生态圈,Spark SQL 的即时查询、Streaming 的流式处理、MLLib 的机器学习和 GraphX 的图处理,能够完美集成并提供一站式解决平台方案。

引用弹性分布式数据集(resilient distributed dataset,RDD),Spark 在集群计算中可将数据集分布式缓存在各节点内存中,省去大量的磁盘 IO 操作,大大缩短访问延迟。作为 Spark 架构的核心模块,RDD 能将用户数据存储在内存,并支持显式缓存(cache)及持久化(persistence)存储,对于需要多次迭代使用的数据,省去了多次内存和磁盘的 IO 时间。RDD 创建之后,只支持两类操作:action 和 transformation。Spark 对 RDD 是惰性计算的,只有在行动操作(action)时,才会真正计算[2]。

1.2　改进的 ISODATA 算法

传统 ISODATA 算法能够在聚类过程中根据各个类所包含样本的实际情况动态调整聚类中心的数目[3]。如果某个类中样本分散程度较大(通过方差进行衡量)并且样本数量较大,则对其进行分裂操作;如果某两个类别靠得较近(通过聚类中心的距离衡量),则对它们进行合并操作。初始状态时聚类数目和中心选择的好坏,直接影响聚类过程中所需要的时间[4],并很有可能对聚类的效果产生影响。而传统 ISODATA 算法在聚类前随机决定聚类数目和选择初始聚类中心,不仅聚类时间不确定,还会造成聚类结果的偶然性,导致出现 ISODATA 算法得到的分类结果不一致的现象。

基于上述问题,对传统 ISODATA 聚类方法进行改进,改进后的算法能够在聚类算法执行前自动确定初始簇数目,并确定簇中心位置。相较传统 ISODATA 算法在聚类前随机决定聚类数目和选择初始聚类中心,改进后的 ISODATA 算法大大缩减了数据计算量,通过减少迭代次数优化算法执行时间。改进 ISODATA 算法的步骤如下:

(1)计算各样本点之间的距离,并将其存储在一个矩阵 $\boldsymbol{A}$ 中,$A(i,j) = \| I_i - I_j \|$,$i,j = 1,2,\cdots,N$。

(2)根据第一步生成的距离矩阵 $\boldsymbol{A}$,将每列上的距离数据累加求平均聚类距离 ad,取 $R_1 = \text{ad} * $(经验参数,一般取 0.5 ~ 2),$R_2 = 2 * R_1$。

(3)将 ad * 作为半径,以每个样本点为圆心画圆,将圆内的数据点作为一个分组。

(4)计算每组内的样本点数目,并按样本数目降序排列,将第一组的中心点作为第一个初始聚类中心 C_1,若第二组的中心点与第一个初始聚类中心 C_1 之间的距离大于 $R_2 = 2 * R_1$,就将其作为第二个初始聚类中心 C_2,否则继续判定下一组,若当前组与已确定的聚类中心的距离都大于 R_2,则将当前组的中心点作为一个新聚类中心,依照这个规则直到没有新的聚类中心产生。

2　负荷特征曲线提取

不同的季节,不同的天气情况,不同时间点的用电负荷曲线都是不同的[5]。不同用户类型的用电规律也是不同的,如居民用电,企业用电和商业用电等。因此文中算法的 K 取值肯定是大于 2 的。对异常数据的检测,本质上就是对存在异常数据的日负荷曲线的检测,将存在异常数据的负荷曲线与正常的负荷曲线模式分开,算法本质上是一个分类问题[6]。

负荷曲线定义:将某个节点连续 t 个时间点及对应的负荷量 y_t 线性拟合的曲线,记为 $Y_t = (y_1, y_2, \cdots, y_t)$。异常数据即为某个或多个时间点的负荷值偏离正常值过多,在负荷曲线上的表现为突变(突增或突减)。以某一节点负荷数据为研究对象,根据该节点历史用电数据,通过聚类分析提取出该节点的日负荷特征曲线,再利用负荷曲线两种特征来检测和辨识负荷曲线上是否存在异常数据及异常数据所在位置。

文中采用改进的 ISODATA 聚类算法对日负荷曲线进行分类,实验时将正常负荷曲线 Y_1 和含异常数据的负荷曲线 Y_2 分别进行处理,然后将和 Y_1 相似的其他不含异常数据的负荷曲线放到一个数据集中,对这类负荷曲线进行聚类,算法最后生成的一条代表这类负荷曲线的不包含异常数据的曲线,就是日负荷特征曲线[7]。日负荷特征代表这一类负荷曲线的数值特征和趋势特征[8]。

用纵向相似性和横向相似性来衡量日负荷曲线的相似性。纵向相似性是指相邻时间段内同一时间点的负荷量是近似相等的,实验以两个值进行衡量:一是待检测曲线与日负荷特征曲线上各采样点之间距离的最大值;二是待检测曲线与日负荷特征曲线上采样点的变化率。横向相似性是指一条曲线相邻时间段的负荷曲线的形状是相似的,指的是相邻采样点的负荷变化率是相似的。

3 异常数据检测与处理

利用提取出的日负荷特征曲线 Y_t 对待检测日负荷曲线 Y_d 进行辨识[9],从而确定日负荷曲线中是否含有异常数据。日负荷曲线拥有横向相似性和纵向相近性,这两个特征可以用来检测负荷曲线上是否存在异常数据。下面分别介绍利用负荷曲线两种特征来检测和辨识负荷曲线上是否存在异常数据及异常数据所在位置。

3.1 利用纵向相似性检测和辨识异常数据

纵向相似性的检测有两个参考标准:待检测曲线与日负荷特征曲线上各采样点之间距离的最大值;待检测曲线与日负荷特征曲线上采样点的变化率。

首先,假设 Y_d 表示待检测日负荷曲线,Y_t 表示日负荷特征曲线。考虑待检测日负荷曲线 Y_d 上的第 i 点,其中 $i \in \{1,2,\cdots,N\}$,N 为采样点数;它的负荷值为 $Y_d(i)$,将其与日负荷特征曲线 $Y_t(i)$ 上第 i 点对应的负荷值 $X_t(i)$ 进行比较,计算待检测负荷曲线 Y_d 与日负荷特征曲线 Y_t 上第 i 点的负荷值 $Y_d(i)$ 、$Y_t(i)$ 的大小,两条曲线间的距离 D 用待检测曲线与日负荷特征曲线上各采样点之间距离的最大值表示。

$$D = \max\{|Y_d(i) - Y_t(i)|\}, i = 1,2,\cdots,N \quad (1)$$

计算两条负荷曲线的负荷变化率 $\theta(i)$,以衡量待检测曲线与日负荷特征曲线上采样点的变化率。

$$\theta(i) = \frac{Y_d(i) - Y_t(i)}{Y_d(i)} \times 100\%, i = 1,2,\cdots,N \quad (2)$$

然后,统计历史上各日第 i 时刻的负荷差值的最大值 D_{max} 和变化率的正常范围 $[\theta_{min},\theta_{max}]$ 。对比待检测日的第 i 时刻的负荷值与日负荷特征曲线 Y_t 的距离 D 是否小于 D_{max} ,且变化率是否在正常范围 $[\theta_{min},\theta_{max}]$ 内,根据这两个条件判断该负荷曲线是否为异常负荷曲线。若不满足任一条件,则待检测曲线为异常曲线,第 i 时刻的负荷值即为异常数据。

3.2 利用横向相似性检测和辨识异常数据

横向相似性指的是相邻采样点的负荷变化率是相似的,在负荷曲线图上表现为一条曲线相邻时间段的负荷曲线的形状是相似的[10]。横向相似性是以相邻采样点的负荷变化率来衡量,负荷特征曲线代表了正常曲线的基本特征,具有很好的横向相似性即平滑性,这就决定了相邻时刻的正常负荷不可能突变,相邻时刻的正常负荷变化率应在某个范围内,不在正常范围内的负荷点就是异常数据点[11-12]。

判断正常负荷变化率的范围,是依据检测单位的历史用电负荷数据。针对实际情况,统计该日各采样点过去所有的负荷变化率 $\theta(i)$,将历史负荷变化率的最小值记为 θ_{min} ,将历史负荷变化率最大值记为 θ_{max} 。则该日负荷采样点的负荷变化率的正常范围是 $[\theta_{min},\theta_{max}]$ 。对比待检测日各采样点的负荷变化率是否在该范围内,以判断该点是否为异常数据点。若 $\theta(i) \notin [\theta_{min},\theta_{max}]$,则待检测曲线为异常数据曲线,第 i 时刻的负荷值即为异常数据点。

当根据日负荷曲线的横向相似性和纵向相似性判定待检测负荷曲线 Y_d 上某一个时刻点 i 的数据为异常数据后,可根据提取出来的特征曲线 Y_t 进行异常数据的修正,根据异常点前后数据的变化率的均值修正异常数据。修正公式为:

$$Y_c(i) = Y_t(i) \times \left[\frac{Y_d(i-1)}{Y_t(i-1)} + \frac{Y_d(i+1)}{Y_t(i+1)}\right]/2 \quad (3)$$

若待检测曲线 Y_d 的 i 点至 j 点为连续异常数据,则对应的修正公式为:

$$Y_c(i) = Y_t(i) \times \left[\frac{Y_d(i-1)}{Y_t(i-1)} + \frac{Y_d(j+1)}{Y_d(j-1)}\right]/2 \quad (4)$$

算法流程如图 1 所示。

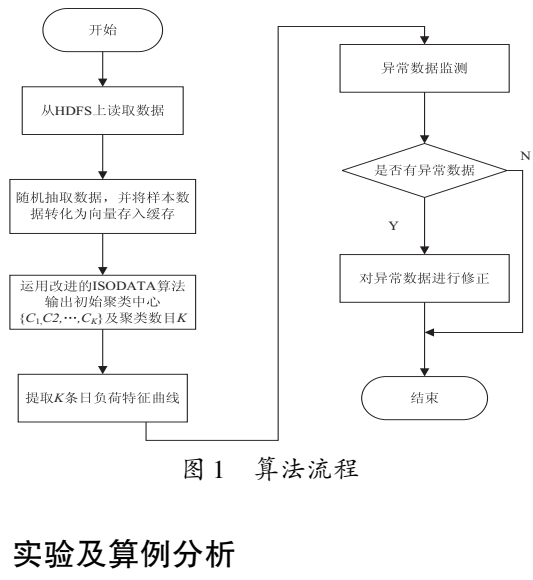

图 1 算法流程

4 实验及算例分析

4.1 实验环境与内容

实验数据采用的是阿里天池大航杯“智造扬中”电力 AI 大赛的比赛数据,开放扬中市高新区 1 000 多家企业的历史用电量数据,主办方提供 2015 年 1 月 1 日到 2016 年 11 月 30 号 1 454 家企业每日用电量。

实验环境:因具体的实验条件限制,平台配置为 5 个服务器节点,每个节点 16 GB 内存,Hadoop 版本为 2.6.4,JDK 版本 1.7,Spark 版本为 2.1.1,Scala 版本为 2.11.8。实验平台在 Intellij IDEA 开发环境上进行开发实现,以 Hadoop 的 HDFS 存储数据。

经数据可视化分析,所获取的测量数据存在异常数据,具有偶然性、分布不确定性,且不同性质的企业存在不同的用电模式[13],实验主要进行以下两点测试:基于 Spark 平台 ISODATA 算法改进前后企业用电类型聚类分析;基于 Spark 平台 ISODATA 算法改进前后异常数据检测性能测试。

4.2　算例分析

4.2.1　企业用电类型聚类分析

实验将传统的 ISODATA 算法,改进后的 ISODATA 聚类算法和基于 Spark 的并行 ISODATA 算法进行比较,测试三种算法对 1 454 家企业某月用电量进行聚类分析的结果,如表 1 所示。

表 1　不同算法聚类比较

算法	聚类数量	执行时间/min	迭代次数
传统 ISODATA	3	27	15
改进后的 ISODATA	5	19	9
基于 Spark 的改进 ISODATA	5	8	8

经测试分析,基于 Spark 改进后的 ISODATA 算法,在算法执行时间和迭代次数上,都优于其他两种算法。

4.2.2　异常数据检测

仍以阿里天池大航杯“智造扬中”电力 AI 大赛的比赛数据为研究对象,为了测试文中方法能否对连续多个异常数据进行准确辨识,选取 728 号企业五周(07.06-08.09)的用电量,将第三周 7 月 21 号,22 号,23 号三天的用电数据 235,258,270 分别增加 60% 的误差,变为 376,413,432。

实验以该企业每周的用电量作为一个数据集,五周正常负荷数据及聚类得到的负荷特征曲线如图 2 所示。

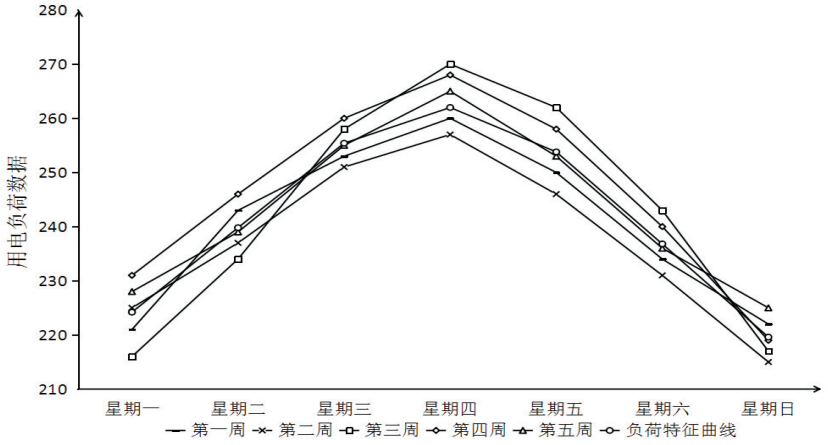

图 2　728 号企业五周负荷曲线及负荷特征曲线

将 728 号企业五周(07.06-08.09)包含三个异常数据的用电量数据集进行聚类得到新的特征曲线,然后将 728 号企业包含三个异常数据的第三周对应的负荷曲线与聚类得到的新负荷特征曲线做负荷变化率计算。含异常数据的负荷曲线和新构建的负荷特征曲线如图 3 所示。

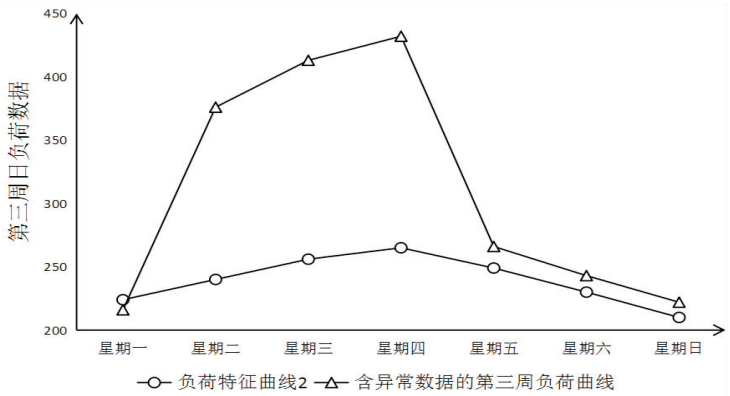

图 3　异常数据处理前的负荷曲线和负荷特征曲线

通过计算这三个异常数据测试点数据负荷变化率,发现都不在正常范围内,分别为 41.22%,42.56%,41.83%。因此验证了算法用纵向相似性和横向相似性来衡量日负荷曲线的相似性论断。判定这三个点为异常数据后,使用上述修正公式进行修正:

$$Y_c(21) = Y_t(21) \times \left[\frac{216}{224} + \frac{266}{249}\right]/2 = 243.9$$

$$Y_c(22) = Y_t(22) \times \left[\frac{216}{224} + \frac{266}{249}\right]/2 = 260.2$$

$$Y_c(23) = Y_t(23) \times \left[\frac{216}{224} + \frac{266}{249}\right]/2 = 269.3$$

修正后的曲线如图 4 所示。

修正后的数据与其实际值的误差百分比如表 2 所示。

方方数据

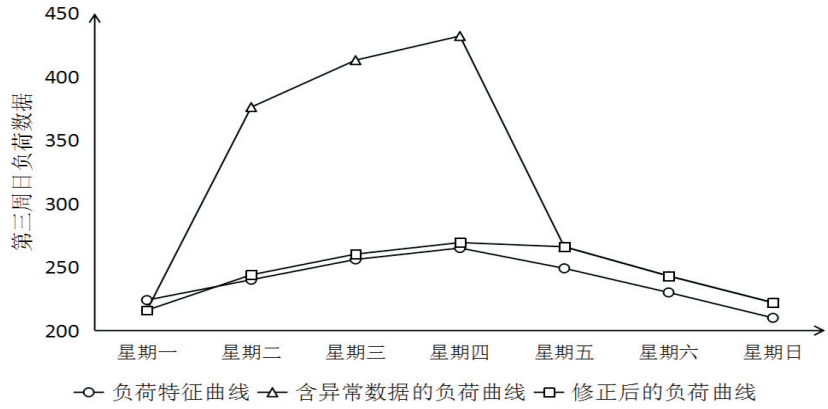

图 4 修正后负荷曲线及特征曲线

表 2 算法修正后误差率

实际值	异常值	修正值	误差率/%
235	376	243.9	3.78
258	413	260.2	0.85
270	432	269.3	0.26

实验结果表明,基于 Spark 的并行 ISODATA 算法对同一时间段的连续异常数据能进行准确检测和辨识,且修正后的数值误差率较小,在正常范围内。

5 结束语

对传统 ISODATA 聚类方法进行改进,改进算法能够在聚类算法执行前自动计算最佳初始簇数目,并确定簇中心位置。实验结果表明,该方法能有效提高算法聚类和异常数据检测的速度,提高异常数据修正的准确性,基于 Spark 的并行 ISODATA 算法对同一时间段的连续异常数据能进行准确检测和辨识,且修正后的数值误差率较小。未来电网中单机环境必定无法处理电力系统中的海量数据,结合 Hadoop HDFS 的分布式存储,基于 Spark 的并行化集群,运用大数据技术中的高性能挖掘算法,以解决智能电网中海量复杂数据的分析与挖掘的实际业务场景难题。在能源互联网和大数据的新时代,依托电力大数据的电网将迈进全景实时的电网时代,大数据将成为电力企业发展和进步的强力依托工具,大数据也因在电力系统中的广泛使用而得到更广阔的发展[14-15]。

参考文献:

[1] 刘 莉,翟登辉,姜新丽. 电力系统不良数据检测与辨识方法的现状与发展[J]. 电力系统保护与控制,2010,38(5):143-147.

[2] SAINI G,KAUR H. A novel approach towards k-mean clustering algorithm with PSO[J]. International Journal of Computer Science & Information Technology,2014,5(4):5978-5986.

[3] 吴军基,杨 一帅,葛 成,等. 基于 GSA 的肘形判据用于电力系统不良数据辨识[J]. 中国电机工程学报,2006,26(22):23-28.

[4] 刘 莉,王 刚,翟登辉. k-means 聚类算法在负荷曲线分类中的应用[J]. 电力系统保护与控制,2011,39(23):65-68.

[5] 王兴志,严 正,沈 沉,等. 基于在线核学习的电网不良数据检测与辨识方法[J]. 电力系统保护与控制,2012,40(1):50-55.

[6] 高彦杰. Spark 大数据处理技术、应用与性能优化[M]. 北京:机械工业出版社,2014:26-30.

[7] TIBSHIRIUL R,WALTHER G,HASTIE T. Estimating the number of clusters in a dataset via the gap statistic[R]. Stanford:Stanford University,2000.

[8] 司风琪,徐治皋. 基于自联想神经网络的测量数据自校正检验方法[J]. 中国电机工程学报,2002,22(6):152-155.

[9] MESSINA A R,VITTAL V. A structural time series approach to modeling dynamic trends in power system data[C]//Proceedings of 2012 IEEE power and energy society general meeting. San Diego,USA:IEEE,2012:1-8.

[10] 董永贵,孙照焱,贾惠波. 时间序列中异常值检测的负向选择算法[J]. 机械工程学报,2004,40(10):30-34.

[11] 黎 祚,周步祥,林 楠,等. 基于模糊聚类与改进 BP 算法的日负荷特性曲线分类与短期负荷预测[J]. 电力系统保护与控制,2012,40(3):56-60.

[12] 张兴民,毛玉华,朱剑峰,等. 利用图论方法进行多不良数据检测与辨识[J]. 中国电机工程学报,1997,17(1):69-72.

[13] SALEHFAR H,ZHAO R. A neutral network pre-estimation filter for bad data detection and identification in power system state estimation[J]. Electric Power Systems Research,1995,34(2):127-134.

[14] 孙雅明,王晨力,张智晟,等. 基于蚁群优化算法的电力系统负荷序列的聚类分析[J]. 中国电机工程学报,2005,25(18):40-45.

[15] 卫志农,张云岗,郑玉平. ISODATA 方法在配网状态估计不良数据辨识中的应用[J]. 河海大学学报:自然科学版,2002,30(2):97-100.