

基于正态分布离群点算法的反窃电研究

为了解决一直以来供电企业所面临的窃电预防及查处问题,结合用电信息采集系统反窃电功能发展的现状,通过对现存的窃电手段和反窃电技术进行分析,提出了一种在远程集抄海量数据的基础上利用基于正态分布的离群点算法进行数据挖掘的反窃电方法。利用Matlab对电压、电流海量数据进行曲线拟合,建立数学模型并引入基于正态分布的离群点算法,依据拉依达准则对海量数据进行数据挖掘找出窃电嫌疑户。通过实际案例验证了提出的离群点算法和窃电户查找流程能够高效筛选出窃电用户,为传统反窃电研究和高效反窃电的实现提供了新的思路。

王新霞¹ 王珂² 焦东翔² 魏新宇² / 1.西南交通大学电气工程学院
2.国网冀北电力有限公司秦皇岛供电公司

0 引言

我国每年因电能盗窃损失达200亿元,且这种经济损失呈现递增的态势。窃电现象越来越多、不同的窃电手段也层出不穷。窃电行为不仅扰乱了正常的供用电秩序,损害了供电企业的合法权益,而且威胁了电网的安全,因此如何实现高效反窃电一直是当前的热点问题^[1]。目前常见的窃电手段有电流型、电压型、扩差法和移相法四种^[2]。反窃电方法包括管理措施和技术措施两方面^[3-4]。除了以计量装置的升级和改造为主的主流技术措施外,文献[5]中提出了一种基于无线通信技术的防窃电实时监控系統,它能够准确实时地检测出窃电行为,但需要搭建470 MHz微功率无线传输通道,且无法保证与主站建立稳定的数据通信。文献[6]提出了一种基于距离离群点算法的反窃电方法,其中检测参数的设置会使得反窃电检测效果有很大的不确定性。文献[7]中介绍了一种新型防窃电

装置,增加了成本且存在设备兼容性问题。文献[8-9]中提出利用用电信息采集系统中的线损管理功能,但该方法只可以知道实施窃电用户所在的具体线路,却无法定位至某一具体用户,检查范围广,查处难度大,并且在进行电量追补的时候缺乏有效的工具和数据。

为了解决以往依靠台区线损异常、用户每月电量变化确定窃电嫌疑户或采用防窃电装置^[10]等传统方法准确率低、时效性差,通过现场排查发现窃电行为又效率低的问题,本文提出了一种在远程集抄海量数据的基础上,利用基于正态分布的离群点算法进行数据挖掘的反窃电方法。

1 窃电手段分析

窃电现象很普遍,常常发生在10 kV和380 V供电线路上^[11]。10 kV以上的用户由变电站出线直接供电,没有偷电,也很少出现漏电。而后者是复杂的树状结构,用户很多,是偷漏电的多发区。虽然380 V用户多,但10 kV的偷漏电损失严重,其电量损失约为前者的3倍,故本文将高压用户为主要研究群体。

统计秦皇岛地区2012~2015年4年间查处核实的窃电事件213起,将窃电手段按照电流型窃电、电压型窃电、移相法窃电及扩差法窃电分类,如图1所示。

由图1可知,在实际工作中,电流型和电压型窃电法是主要的窃电方式。故本文将采用Matlab对高压用户的日负荷电压、电流曲线数据进行曲线拟合,观察海量数据所体现出的规律性并建立数



王新霞/硕士研究生

关键词/Keywords

正态分布·
离群点算法·
拉依达准则·
反窃电·

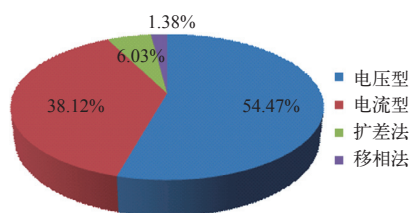


图1 窃电手段占比饼图

学模型, 进而引入基于统计的离群点算法, 对高压用户的电流型及电压型窃电进行反窃电分析研究。

2 高压用户海量数据曲线拟合

高压用户计量装置接线采用高供高计和高供低计两种方式。前者采用三相三线制两元件计量, 计量装置额定电压为 3×100 V; 后者采用三相四线制三元件计量, 计量装置额定电压为 $3 \times 220/380$ V。

本文随机抽取秦皇岛市直供直管区域台区线损值在正常范围内, 且没有出现窃电现象的30个台区1 000户用户2016年3月21日的电压、电流日负荷数据作为样本。目前用电信息采集系统已经能够实现每隔15 min对用户计量装置端口电压进行一次采样, 即一天中可以采集96点电压、电流数据。一般窃电发生时间都会维持几个小时, 因此在进行窃电分析时, 可对每天的96个数据点进行简化处理, 只分析每小时一次的电压、电流数据, 即每天24个数据点即可。

2.1 高压用户电压数据曲线拟合

对于高供低计的高压用户, 其采用了三相四线制接线, 计量装置额定电压为220 V。考虑到线路损耗, 在正常用电情况下采集上传至用电信息采集系统的三相电压值会比220 V更大一些。由于受到计量装置的计量准确度和供电线路的实际参数影响, 电压值会在某定值左右一定范围内波动, 但不超过一定的阈值范围。以A相电压为例进行曲线拟合, 仿真结果如图2所示。

从图2可以看出, 高供低计用户电压数据服从正态分布。进而画出该组数据的概率密度函数分布图如图3所示。

图3中, 蓝色为拟合概率密度函数曲线, 红色为实际数据函数曲线, 可以看出高供低计用户A相

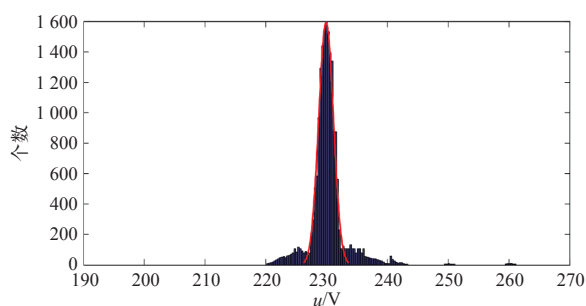


图2 高供低计用户电压数据带正态拟合的分布直方图

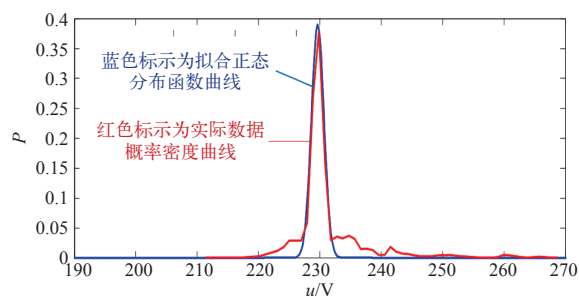


图3 高供低计用户电压数据拟合曲线

电压在229 V左右的一定范围内波动, 且基本符合正态分布。

对于高供高计的高压用户, 其采用了三相三线制接线, 计量装置额定电压为100 V, 考虑到线路损耗, 在正常用电情况下采集上传至用电信息采集系统的三相电压会比100 V更大一些。由于受到计量装置的计量准确度和供电线路的实际参数影响, 电压值会在某定值左右一定范围内波动, 但不超过一定的阈值范围。同样以A相电压为例进行曲线拟合, 仿真结果如图4所示。

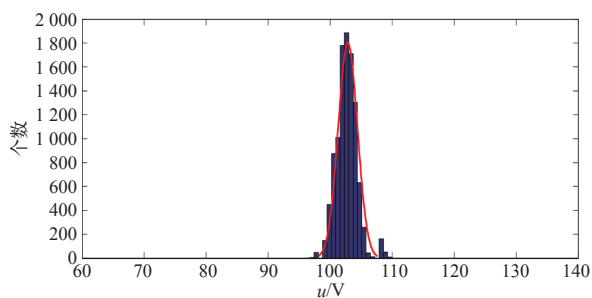


图4 高供高计用户电压数据带正态拟合的分布直方图

从图4可以看出, 高供高计用户电压数据分布服从正态分布的概率分布。进而画出该组数据的概率密度函数分布图如图5所示。

图5中, 蓝色为拟合概率密度函数曲线, 红色为实际数据函数曲线, 可以看出高供高计用户A相

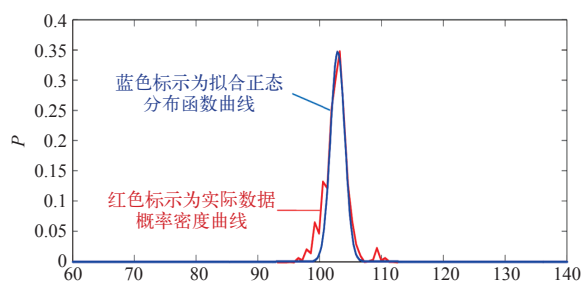


图5 高供高计用户电压数据拟合曲线

电压在103 V左右的一定范围内波动且基本符合正态分布。

2.2 高压用户电流数据曲线拟合

随着用户负荷变化负荷电流在不断变化,因此不能直接拿来作为研究对象。根据用户正常用电时的三相电流基本平衡,仅有极小波动,本文将采用两相电流差值作为研究对象。高供高计用户采用两元件计量,仅计量A、C两相电流,因此本文统一采用 $I_i = I_A - I_C$ 作为研究值。

高供低计用户由于受到计量装置计量准确度和实际负荷的平衡情况影响,在正常用电情况下,计量装置采集到并上传至用电信息采集系统的A、C两相电流基本大小相当,但会在一定范围内波动,如图6所示。按照 $I_i = I_A - I_C$ 作为研究值对高供低计A、C两相电流进行预处理仿真结果如图7所示。

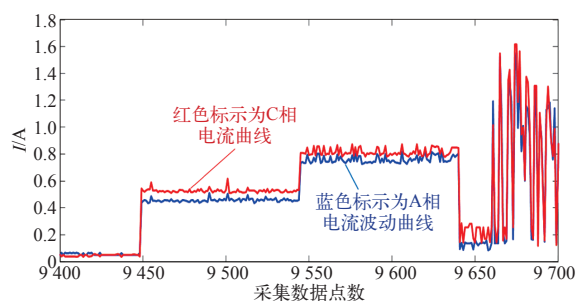


图6 高供低计用户A、C相电流曲线局部放大图

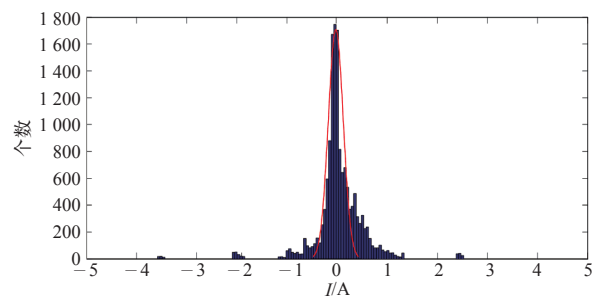


图7 高供低计用户电流数据带正态拟合的分布直方图

图6中蓝色为A相电流值,红色为C相电流值,可以看出A、C两相电流基本上保持一致但会在一定范围内波动。

从图7可以看出,高供低计用户电流数据分布服从正态分布的概率分布。进而画出该组数据的概率密度分布图如图8所示。

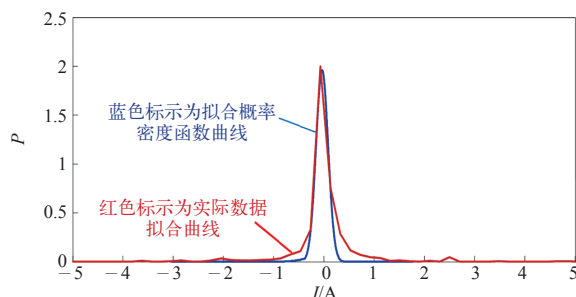


图8 高供低计用户A、C相电流预处理仿真图

图8中,蓝色为拟合概率密度函数曲线,红色为实际数据函数曲线,可以看出高供低计用户研究值 $I_i = I_A - I_C$ 在-0.2 A左右的一定范围内波动,且基本符合正态分布。

高供高计用户由于受到计量装置计量准确度和实际负荷的平衡情况影响,在正常用电情况下,计量装置采集到并上传至用电信息采集系统的A、C两相电流基本大小相当,但会在一定范围内波动,如图9所示。按照 $I_i = I_A - I_C$ 作为研究值对高供高计A、C两相电流进行预处理仿真结果如图10所示。

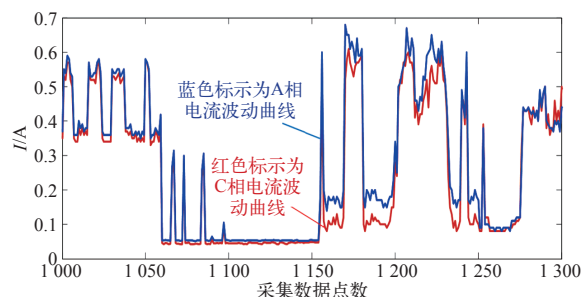


图9 高供高计用户A、C相电流曲线局部放大图

图9中蓝色为A相电流值,红色为C相电流值,可以看出A、C两相电流基本保持一致,但会在一定范围内波动。

从图10可以看出,高供高计用户电流数据分布服从正态分布的概率分布。进而画出该组数据的概率密度分布图如图11所示。

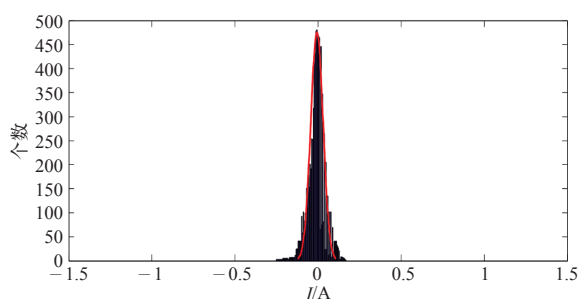


图10 高供高计用户电流数据带正态拟合的分布直方图

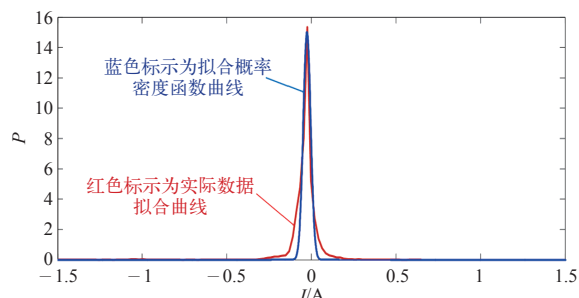


图11 高供高计用户A、C相电流预处理仿真图

图11中,蓝色为拟合概率密度函数曲线,红色为实际数据函数曲线,可以看出高供高计用户研究值 $I_i = I_A - I_C$ 在 -0.1 A左右的一定范围内波动,且基本符合正态分布。

3 窃电判定算法及筛选流程

3.1 离群点算法综述

在数据集中,如果某些数据对象表现的特性与很多对象不同,那么可以认为该数据对象可能是来自另外一种完全不同的机制,一般称这种数据为离群数据,或异常点、孤立点等^[12]。Hawkins等给出了离群数据的定义^[13]:离群数据在数据集中与其他数据完全不同,使人怀疑它们不是由于非随机偏差而产生,而是由于完全不同的机制。离群点检测目的是清除噪声或发现有意义的、潜在的重要的信息,离群点检测方法主要分为基于距离的、基于密度的、基于聚类的、基于规则的、基于偏离的、基于深度的和基于统计的7类^[14-18]。

基于距离的方法具有较高的时间复杂度,且难以挖掘局部离群点;基于密度的方法多应用于在局部离群点检测中;基于聚类的算法对离群点检测的质量很依赖于聚类的效果,每个聚类算法

都仅适用于一类数据,聚类算法的选取很重要;基于规则的算法仅适用于数据属性间有逻辑关系的一类数据;基于偏离的算法过于理想,选相异度函数也极为复杂,该函数的选取直接影响结果的准确性;基于深度的方法对高维数据处理效率低^[19-20]。

因正常用电时,电力用户的三相电压及电流值为某定值附近的微小波动,故窃电发生后电压、电流出现的异常就使得窃电户成为全局的离群点。受计量装置计量准确度和供电线路的实际参数情况影响,其值会在一定范围内波动。由上文中采用Matlab对海量电压、电流数据进行曲线拟合的仿真结果可知,用户电压、电流研究值服从正态分布。综合考虑各方面因素,本文将基于正态分布的离群点检测算法应用于窃电判定。

3.2 基于正态分布的离群点检测方法

该检测方法是为数据集构建一个正态分布概率统计模型,然后根据该模型采用不和谐检验识别出离群点。不和谐检验需要作出两个假设:工作假设 H_0 和备用假设 H_1 。工作假设 H_0 是一个命题,用来描述数据集的总体性质。整个数据集都服从一个初始的分布模型 F ——正态分布概率密度函数模型。即

$$H_0: X_i \in F \quad i=1,2,3,\dots,n \quad (1)$$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (2)$$

式中, μ 为期望值; σ 为均方差。

不一致性检验验证 X_i 与模型 F 相比差异是否显著的大或者小,来判断是否可以接受这个假设。如果分布上没有显著的证据来否认这个假设,就接受 H_0 。如果样本数据中的某个统计量相对于其他数据分布概率特别小,就可以认为该样本点是不和谐的,该工作假设不成立。此时可采用备用假设,它表明该样本点来自于另外一个完全不同的分布模型。总之,假如某样本点不符合工作假设,就可以认为它是离群点。如果它又符合另外一个备择假设,那么它就是符合备用假设分布的离群点。

3.3 窃电用户筛选流程

由Matlab仿真结果可知高供高计、高供低计用户电压和电流研究值均符合正态分布概率密度函数模型。可以求得高供高计、高供低计电压正态分布概率密度函数参数分别为： $\mu=102.736\ 5$ ， $\sigma=6.598\ 0$ 、 $\mu=229.660\ 5$ ， $\sigma=11.863\ 7$ ，电流正态分布概率密度函数参数分别为 $\mu=-0.097\ 0$ ， $\sigma=0.921$ 、 $\mu=-0.197\ 0$ ， $\sigma=1.943$ 。依据拉依达准则^[22]确定置信区间，置信概率为99.7%。当满足式(3)时认为该值为含有粗大误差的异常值，即该户为窃电嫌疑对象，即

$$X_i - \mu > 3\sigma \quad i=1,2,3,\dots,n \quad (3)$$

查找窃电户的算法流程图如图12所示。

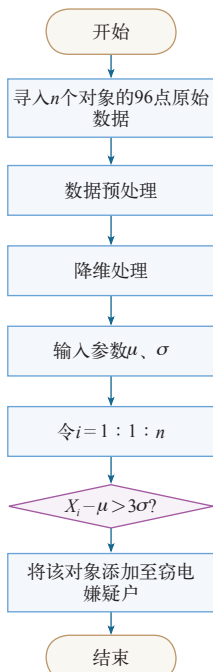


图12 窃电用户查找流程

4 案例分析

以秦皇岛地区窃电现象较为严重、台区线损率较高的30个台区1 000户用户作为分析对象，构成测试数据样本。分别利用电压型、电流型窃电的基于正态分布的离群点判定算法对窃电嫌疑用户进行筛选，并派用电稽查人员去现场检查核实情况，结果如下表所示。

为了测试本文所提算法的准确性，安排了6名

表 样本算法检测及现场核实情况

表计资产号	窃电方式	窃电相	现场核实	实际窃电相
13403012640176346	电压型	C相	窃电	C相
13403012537882309	电压型	A相	窃电	A相
112001800880003531477	电流型	A相	窃电	A相
163000100000008760712	电流型	C相	窃电	C相
30K24M05942400062642	电压型	A、C相	窃电	A、C相
13403012640177546	电流型	C相	窃电	C相
30K24M05942400064384	电流型	A相	窃电	A相
112002000620006429335	电压型	C相	窃电	C相
112002000620006429628	电流型	A相	窃电	A相

用电稽查人员对该测试数据样本中的窃电嫌疑户9户以外的991户用电户进行了为期两周的逐户现场用电稽查，并未发现有其他的窃电用户。可见，本文所提的算法及筛选条件完全能够找出窃电用户。在实际工作应用中，为了保证能够有效地打击窃电行为，要优先确保该算法能够完全不遗漏地筛选出窃电户，为此可以一定程度上牺牲准确率，故可适当缩小置信区间以保证所有窃电户均被检测出。

5 结束语

本文针对长期以来供电企业所面临的窃电预防及查处难题，结合用电信息采集系统反窃电功能发展现状，通过对现存的窃电手段和防窃电技术进行分析，提出了一种在远程集抄海量数据的基础上，利用正态分布离群点算法进行数据挖掘的反窃电方法。利用Matlab对电压、电流海量数据进行曲线拟合，建立数学模型，并引入分布式离群点算法，依据拉依达准则对海量数据进行数据挖掘。通过对秦皇岛地区1 000户用电户的测试数据样本进行离群点算法判定、窃电嫌疑户甄别及现场检查核实确认，结果表明本文所提的离群点判定算法和窃电户筛选流程完全能够找出窃电用户，为监控人员利用用电信息采集系统海量数据及时准确地进行反窃电分析提供了新的思路。

参考文献

- [1] 王珏昕, 孟宇, 殷树刚, 等. 用电信息采集系统反窃电功能现状及发展趋势 [J]. 电网技术, 2008

- (增刊2): 177-178.
- [2] 陈腾飞. 基于用电信息采集系统的窃电在线稽查装置的开发应用[D]. 北京: 华北电力大学, 2013.
- [3] 吴小美. 防丢窃电的技术措施与管理[J]. 电测与仪表, 2003(2): 8-10.
- [4] 郭立才, 彭志炜, 范强. 电能计量及反窃电方法综述[J]. 高压电器, 2010(5): 86-88, 91.
- [5] 王辉, 刘斐. 无线通信技术在防窃电工作中的应用[J]. 电测与仪表, 2015(1): 124-128.
- [6] 程超, 张汉敬, 景志敏, 等. 基于离群点算法和用电信息采集系统的反窃电研究[J]. 电力系统保护与控制, 2015(17): 69-74.
- [7] 赵兵, 吕英杰, 邹和平. 一种新型防窃电装置的设计[J]. 电力系统保护与控制, 2009(23): 116-118, 184.
- [8] 胡江溢, 祝恩国, 杜新纲, 等. 用电信息采集系统应用现状及发展趋势[J]. 电力系统自动化, 2014(2): 131-135.
- [9] 范洁, 陈霄, 周玉. 基于用电信息采集系统的电能计量装置异常智能分析方法研究[J]. 电测与仪表, 2013(11): 4-9.
- [10] 韩松林, 尚德祖, 高杰, 等. 反窃电、防窃电及防窃电产品[J]. 电测与仪表, 2002(9): 10-14, 29.
- [11] 辛洁晴, 夏正侃, 高亦凌, 等. 基于事故树的窃电损失要因分析[J]. 电力系统及其自动化学报, 2014(3): 47-51.
- [12] 郑敏. 大型数据集中离群数据挖掘算法研究及应用[D]. 重庆: 重庆大学, 2011.
- [13] Hawkins D. Identifications of outliers [M]. London: Chapman and Hall, 1980.
- [14] Chandarana D R, Dhamecha M V. A survey for different approaches of Outlier Detection in data mining [C]. Visakhapatnam: Electrical Electronics Signals Communication and Optimization (EESCO) 2015 International Conference, 2015, 1-5.
- [15] Rizk Y H, Elgokhy S, Sarhan A. A hybrid outlier detection algorithm based on partitioning clustering and density measures [C]. Cairo: Computer Engineering & Systems (ICCES), 2015 Tenth International Conference, 2015, 175-181.
- [16] Bhaduri Y K, Stefanski M D, Srivastava A N. Privacy-preserving outlier detection through random nonlinear data distortion [J]. IEEE Transactions on Systems, 2011, 41(1): 260-272.
- [17] Purwar Y A, Singh S K. Issues in data mining: A comprehensive survey [C]. (ICCIC) 2014 IEEE International Conference on Computational Intelligence and Computing Research, Coimbatore, 2014, 1-6.
- [18] Shaikh S A, Kitagawa H. Continuous outlier detection on uncertain data streams [C]. 2014 IEEE Ninth International Conference on Intelligent Sensors Sensor Networks and Information Processing (ISSNIP), Singapore, 2014, 1-7.
- [19] 杨福萍. 离群点检测及其应用研究[D]. 济南: 山东师范大学, 2013.
- [20] 刘靖. 复杂数据类型的离群检测方法研究[D]. 广州: 华南理工大学, 2014.
- [21] Zhang Liang, Qin Yongyuan, Zhang Jinliang. Study of polynomial curve fitting algorithm for outlier elimination [C]. 2011 International Conference on Computer Science and Service System (CSSS), Nanjing, 2011, 760-762.
- [22] Dao-wen C, Dong H, Wen-yong W, et al. Study on the fast judgment of abnormal value with Excel [C]. 2012 2nd International Conference on Computer Science and Network Technology (ICCSNT), Changchun, 2012: 652-655.

(收稿日期: 2016-04-11) **EA**

欢迎订阅 欢迎投稿 欢迎刊登广告