第 26 卷 第 22 期 中 国 电 机 工 程 学 报 Vol.26 No.22 Nov. 2006
2006 年 11 月 Proceedings of the CSEE ©2006 Chin.Soc.for Elec.Eng.

文章编号：0258-8013 (2006) 22-0023-06 中图分类号：TM77 文献标识码：A 学科分类号：470·40

基于 GSA 的肘形判据用于电力系统不良数据辨识

吴军基[1]，杨 伟[1]，葛 成[2]，赵 彤[3]

（1. 南京理工大学动力工程学院，江苏省 南京市 210094；2. 安徽电力设计院，安徽省 合肥市 230022；
3. 江苏电力公司，江苏省 南京市 210024）

Application of GSA-based Elbow Judgment on Bad-data Detection of Power System

WU Jun-ji[1],YANG Wei[1],GE Cheng[2],ZHAO Tong[3]

（1. College of Power Engineering, Nanjing University of Science and Technology, Nanjing 210094, Jiangsu Province, China;
2.Anhui Electric Power Design Institute，Hefei 230022, Anhui Province, China;
3.Jiangsu Power Company，Nanjing 210024, Jiangsu Province, China）

ABSTRACT: Based on bad data detection using GSA (Gap Statistic Algorithm) data mining method in power system, this paper propose the elbow criterion to estimate optimal clustering number. The criterion analyzes the relation between the degree of clustering dispersion and clustering number k of the data set firstly, then calculates the elbow angle at k and obtain the optimal clustering number based on the least elbow angle. Combined the criterion with GSA, bad data detection could be implemented efficiently. Computer results show that the integrated method not only can avoid residual pollution and residual submersion which would appear using traditional state estimate detection, but also is more accurate and rapid than GSA method. In the case of huge system and large amount of data, this method is a rapid and efficient algorithm, and has potential of good application.

KEY WORDS: power system; identification of bad data;elbow criterion; gap statistic algorithm; data mining; cluster

摘要：在分析 GSA (gap statistic algorithm)数据挖掘技术应用于电力系统不良数据辨识的基础上，提出一种判断最佳聚类个数的肘形判据，该判据通过分析数据集的聚类离散度与聚类个数 k 的关系，按照各个 k 点的聚类离散度计算 k 处的肘形折角，并以最小肘形折角判断最佳聚类个数。将该判据与 GSA 相结合用于电力系统不良数据辨识。仿真结果表明：该方法不仅可以避免状态估计方法辨识的残差污染和残差淹没现象，而且可以克服单纯 GSA 辨识法在计算速度和辨识准确性方面的缺陷。对于大系统、数据量巨大的情况，该方法是一种快速高效的算法，具有很好的应用前景。

关键词：电力系统；不良数据辨识；肘形判据；间隙统计算法；数据挖掘；聚类分析

0 引言

电力系统不良数据的检测与辨识一直是电力系统状态估计的重要功能之一[1-3]。不良数据检测与辨识的常用方法主要是基于状态估计的方法，但这类方法的缺点是可能会出现残差污染和残差淹没现象，从而造成漏检或误检[4-5]。

GSA 方法是一种强化聚类效果的数据挖掘算法，它可以估计数据集最佳的聚类个数[6-7]，在电力系统不良数据辨识中，可以将良好数据和不良数据所在的聚类准确地区分进而检测和辨识不良数据[8]。但在聚类个数较多时其计算量较大，在数据量越来越大的现代电力系统中，期望能够有一种更加快速的算法来满足这种要求。另外，在①多个不良数据同时出现；②相互关联不良数据出现的情况下，单纯 GSA 方法进行电力系统不良数据辨识会出现误判[9]。

为了提高计算速度和减少误判，本文将提出一种估计聚类个数的肘形判据，并将此判据与 GSA 方法相结合，形成基于 GSA 的肘形判据用于电力系统不良数据辨识。仿真表明，该方法可准确地检测和辨识不良数据，未出现误判现象，并可以显著提高计算速度。

1 不良数据的 GSA 辨识法

一组样本$\{x_i\}$，假设样本集被聚类成 k 个聚类

$G_1, G_2, \cdots, G_k$，对于任何聚类 G_a，聚类内每个样本围绕聚类均值的距离平方和 D_a 按照下式计算：

$$D_a = \sum_{x_i \in G_a} (x_i - c_a)^2 \tag{1}$$

式中 c_a 是聚类 G_a 的中心。

对应于聚类个数 k 的聚类离散度为

$$W(k) = \sum_{a=1}^{k} D_a \tag{2}$$

GSA 算法的核心就是将聚类离散度的自然对数与一个参考值进行比较，进而确定最佳的聚类个数。GSA 算法使用自然对数对聚类离散度进行处理，其目的是让离散度曲线更加线性化，使样本聚类和参考数据聚类之间的间隙值就更容易被确定。这里定义

$$g_{ap}(k) = E\ln[W_r(k)] - \ln[W(k)] \tag{3}$$

式中：E 表示参考数据的数学期望，下标 r 表示参考数据。

当随着 k 的变化，首次出现某个较大 $g_{ap}(k)$的时，k 就被认为是最合适的聚类个数，而当随着 k 的变化 $g_{ap}(k)$无明显变化时，认为最佳的聚类个数为 1[8-11]。

参考数据集的选取一直是 GSA 方法的重要问题，研究表明可以在待检测数据集的观察值范围内以均匀分布方式产生参考数据集。该方法中共需产生 F 组参考分布数据集，取其均值作为 $E[\ln W_r(k)]$ 的估计。每组值通过参考分布的样本值计算而得。

$$E[\ln W_r(k)] = \frac{1}{F}\sum_{j=1}^{F} \ln W_{r,j}(k) \tag{4}$$

GSA 算法用于电力系统不良数据辨识的具体实现过程见文献[8-10]。

2 基于 GSA 的肘形判据

肘形判据[9]是一种通过分析数据集的聚类离散度与聚类个数 k 的关系，按照各个 k 点的聚类离散度计算 k 处的肘形折角，并以最小肘形折角判断最佳聚类个数的判断依据。

分析 GSA 算法可知，其核心是比较聚类离散度的对数值与它的参考值，进而确定最佳的聚类个数。这样做的实质是建立一个合适的依据，寻找聚类离散度与聚类个数关系中隐藏的最佳聚类个数的信息。在参考曲线与实际曲线的对比中，最大间隙 gap 值的出现实质上是对应着 $\ln W(k)$–k 曲线从某个 k 值开始明显下降平缓的“最小肘形折角”位置。

图 1 是一聚类个数是 2 的 $\ln W(k)$–k 的典型曲线。现在根据图形求曲线在各个 k 处的肘形折角，即各

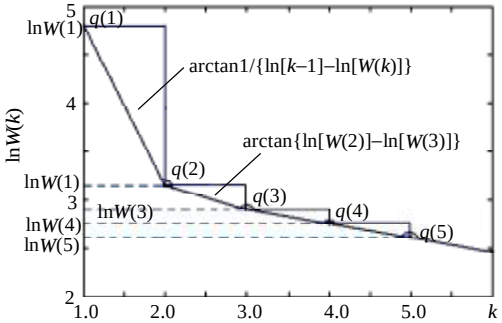

图 1 肘形判据原理分析图

Fig. 1 Theoretical figure of elbow judgment

个折点处 2 条直线段夹成的折角$q(k)$，它由 3 部分组成，q_u、q_r和π/2。q_u与 $W(k-1)$和 $W(k)$有关，q_r 可由 $W(k)$和 $W(k+1)$得到求取的数学公式如下：

$$q(k) = \pi/2 + q_u + q_r \tag{5}$$

其中：

$$q_u = \arctan\left\{\frac{1}{\ln[W(k-1)] - \ln[W(k)]}\right\};$$

$$q_r = \arctan\{\ln[W(k)] - \ln[W(k+1)]\}。$$

由于必须已知 $\ln[W(k-1)]$ 和 $\ln[W(k+1)]$ 值才能求得 $q(k)$，而求取 $q(1)$ 就必须已知 $\ln[W(0)]$，也就是说必须已知一组数据聚类成 0 个聚类的聚类离散度才能对聚类个数为 1 的情况进行有效估计，而一个数据集聚类成零个聚类的聚类离散度没有一个客观的计算依据。

关于 GSA 的研究表明[9]：GSA 在聚类个数为 1 的情况下比以往任何估计聚类个数的方法都优秀。但当聚类个数大于 1 时，由于需要产生 F 组均匀参考分布分别进行聚类计算，因此 GSA 的计算量较大，而肘形判据在聚类个数大于 1 时的计算量小，判断准确度高。针对这种情况可以考虑将 GSA 方法与肘形判据相结合，形成基于 GSA 的肘形判据。

3 基于 GSA 的肘形判据的系统建模[9]

与 GSA 方法的仿真类似，基于 GSA 的肘形判据方法也要将经过神经网络计算得到的平方误差数据作为其程序的输入数据，进而得出辨识结果。

基于 GSA 的肘形判据的主要计算步骤如下：

（1）判断不良数据的存在性。

对一组待检测数据 $(e_i-O_i)^2$，令 k=i, i=2。将待检测数据和参考数据分别聚类，得到聚类数据的聚类离散度。并计算 $g_{ap}(1)$和 $g_{ap}(2)$。

$$g_{ap}(k)=\frac{1}{F}\sum_{j=1}^{F}\ln W_{r,j}(k)-\ln W(k) \tag{6}$$

如果式(6)满足

$$g_{ap}(1)\geq g_{ap}(2)-s_2 \tag{7}$$

式中：$s_2=s_{d2}\sqrt{1+1/F}$；

$$s_{d2}=\sqrt{(1/F)\sum_{j=1}^{F}\left[\ln W_{r,j}(2)-\frac{1}{F}\sum_{j=1}^{F}\ln W_{r,j}(k)\right]^2}\ 。$$

则最合适的聚类个数应该为 1。则说明所有的数据均是良好数据。如果式(6)不满足式(7)，则进入下一步。

（2）应用肘形判据计算肘形折角。

置 k=1，并令 k=k+1，分别求取待检测数据的聚类离散度 ln$W$(k)，进而计算聚类离散度曲线在各个聚类点处的肘形折角为

$$q(k)=\pi/2+q_u+q_r \tag{8}$$

式中：$q_u=\arctan\left\{\dfrac{1}{\ln\left[W(k-1)\right]-\ln\left[W(k)\right]}\right\}$；

$q_r=\arctan\{\ln\left[W(k)\right]-\ln\left[W(k+1)\right]\}$。

完成对待检测数据聚类离散度曲线在各个聚类点处的肘形折角的求取之后，进入下一步。

（3）确定最合适的聚类个数。

在这个步骤中，寻找最小的 k 使之满足

$$q(k)<q(k+1) \tag{9}$$

k 即为最佳的聚类个数。

（4）检测和辨识不良数据。

如果最佳的聚类个数为 1，则表示所有的量测数据都是正常数据，否则就表示存在不良数据，要计算每个聚类内数据的平均值。具有最小平均值的聚类被认为是正常数据的聚类，其余的都被认为是不良数据组成的聚类，这样不良数据就可以被对应地检测和辨识出来。

基于 GSA 的肘形判据用于不良数据辨识的流程如图 2 所示。

4　仿真分析[9]

4.1　仿真系统简介

论文的仿真数据取自江苏省电力公司调度通信中心实时运行数据。数据取自镇江发电厂、谏壁发电厂、五洲变电站、官塘变电站和上党变电站 2 个发电厂和 3 个变电站的局部电网的实时运行数据。从此系统中共获取 88 个量测值，取自 2005 年 4 月 21 日的运行情况，其中包括 12 个节点电压值，10 对发电机组出力的有功和无功，15 对负荷潮流值，3 对变压器输入有功无功以及 10 对线路潮流值，对于各个量测值，都分配各自的标号以方便进一步分析。共从省调度通信中心采样 180 组实时量测数据，其中的 160 组作为训练样本对神经网络进行训练，余下的 20 组用来对神经网络进行测试。

系统量测数据先要经过神经网络的测试[12-13]。网络将系统的理论计算值 z^+作为目标值[14-15]。但对于大系统而言，由于不良数据辨识实时性的要求，在训练和测试阶段可将量测值分为几组分别计算[15]，也就是说，每组数据都可以有它相应的网络，这样，对于大量数据，就可以通过多机系统进行分析[15-16]。神经网络被训练好之后，将待检测的量测值经过神经网络测试，得出的输入输出差的平方值作为聚类分析的数据。

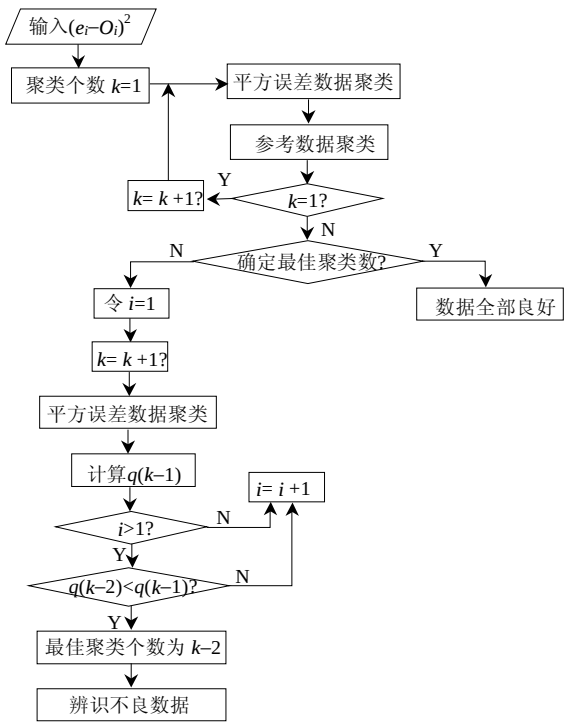

图 2　基于 GSA 的肘形判据用于不良数据辨识的流程图
Fig. 2　Flowchart of GSA-Based Elbow Judgment

4.2　正常情况

此种情况下，得出的最佳聚类个数是 1，程序计算到 GSA 部分即终止，表明所有的数据都是良好数据。当运行到聚类个数为 2 的时候，式(6)得到满足，最佳聚类数为 1，计算得到的结果见表 1。

表 1 正常情况下结果分析

Tab. 1 Clustered results for normal condition

k	ln$W(k)$	E[ln$W_r(k)$]	$g_{ap}(k)$	s_k
k=1	–2.8853	–2.5647	0.29063	0.0596
k=2	–11.068	–10.946	0.12206	0.0821

4.3 单个不良数据情况

这种情况下假设第 27 号量测数据(官塘变 2935 号联络线有功)超过正常值 25%，为不良数据。计算结果见表 2，图 3 和 4 分别表示的是 $g_{ap}(k)/q(k)$的分析以及 ln$W(k)$曲线。

由图 3 的 $g_{ap}(k)$可知， $g_{ap}(1)< g_{ap}(2)$，说明数据中存在不良数据；然后用肘形判据进行分析得到最佳的聚类个数为 2。进一步计算 2 个聚类内元素的平均值，得到较大平均值聚类内的元素为 27 号数据。仿真结果验证了算法的准确性和有效性。

表 2 单个不良数据情况下辨识结果分析

Tab. 2 Clustered results of single bad data

k	ln$W(k)$	E[ln$W_r(k)$]	$q(k)$	$g_{ap}(k)$	s_k
1	–2.1164	0.06979	6.2832	2.1862	0.0599
2	–9.292	–5.6637	2.3718	3.6282	0.0798
3	–10.072		3.2017		
4	–10.954				

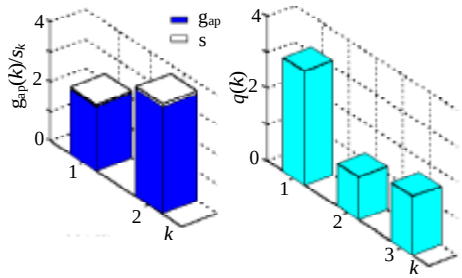

图 3 单个不良数据情况下 $g_{ap}(k)$与$q(k)$的分析

Fig. 3 $g_{ap}(k)$ and$q(k)$ under the single bad data scenario

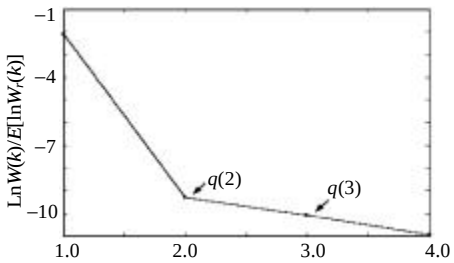

图 4 单个不良数据情况下的 ln$W(k)$曲线

Fig. 4 Curve of ln$W(k)$ under the single bad data scenario

4.4 多不良数据情况

电力系统中不良数据出现的机率是比较小的，一般约为 0.27%；多不良数据出现的可能性则更小。虽然这种情况不易出现，但这种情况用常规状态估计方法通常难以解决。因为在多个不良数据相互作用的影响下，可能导致在部分甚至全部数据点上的残差接近于正常残差，这样不良数据点不呈现残差特性，从而出现残差淹没现象，造成漏检。基于 GSA 的肘形判据由于采用自动聚类分析，寻找数据间的相似关系来自动查找不良数据，可有效地避开残差淹没现象。这种情况下假设量测数据共出现 6 个不良数据，编号分别是 6、19、26、36、52 和 78，其值与正常值相差在 15%~30%之间。计算结果见表 3，$g_{ap}(k)/q(k)$以及 ln$W(k)$的分析见图 5 和图 6。由图 5 可知：$g_{ap}(1)< g_{ap}(2)$，说明数据中存在不良数据；然后用肘形判据进行分析得到最佳的聚类个数为 3。图 6 的 ln$W(k)$曲线显示最小肘形折角位置是 k=3。程序将不良数据分为两类，良好数据聚成一类。进一步计算 3 个聚类内元素的平均值，得到 2 个较大平均值聚类内的元素为 6 个假设的不良数据。仿真结果表明：该方法可以有效地避免残差淹没现象，正确辨识不良数据。

表 3 多不良数据情况基于 GSA 的肘形判据的计算结果

Tab. 3 Clustered results for multiple bad data

k	ln$W(k)$	E[ln$W_r(k)$]	$q(k)$	$g_{ap}(k)$	s_k
1	–0.8586	0.4736	6.2832	1.3322	0.0516
2	–6.7039	–4.8374	2.9912	1.8666	0.0828
3	–9.7237		2.5877		
4	–10.561		2.9277		
5	–11.086				

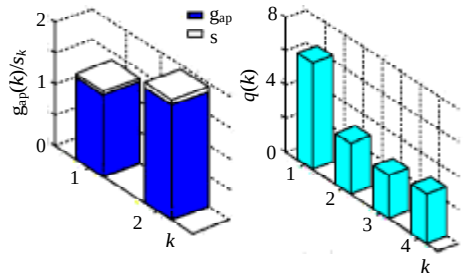

图 5 多不良数据情况下 $g_{ap}(k)$与$q(k)$的分析

Fig. 5 $g_{ap}(k)$ and $q(k)$ under the multiple bad data scenario

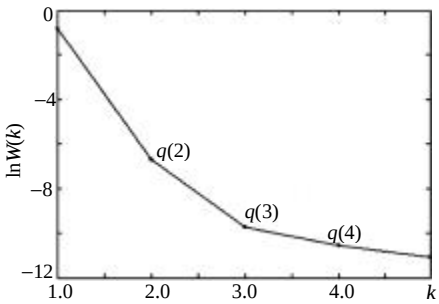

图 6 多不良数据情况下的 ln$W(k)$曲线

Fig. 6 Curve of ln$W(k)$ under the multiple bad data scenario

4.5 相互关联不良数据情况

系统一般出现不良数据时，不良数据往往相互关联。下面分析相互关联的不良数据情况的仿真。此情况下假设不良数据取自官塘变母线电压、官塘

变1号变压器的有功无功以及官塘变2935号联络线有功和无功共 5 个量测值。假设其不良数据的值为超过正常值 15%~20%。计算结果如表 4。$g_{ap}(k)/q(k)$以及 ln$W(k)$的分析见图 7 和 8。此情况得到的最佳聚类个数为 2，不良数据与良好数据被很好地区分开来。

表 4 相互关联不良数据情况的计算结果

Tab. 4　Clustered results of correlated bad data scenario

k	ln$W(k)$	E[ln$W_r(k)$]	$q(k)$	$g_{ap}(k)$	s_k
1	–1.3276	0.1639	6.2832	1.4906	0.0538
2	–7.0595	–5.4635	2.5067	1.5960	0.0943
3	–8.0161		2.9321		
4	–8.6344				

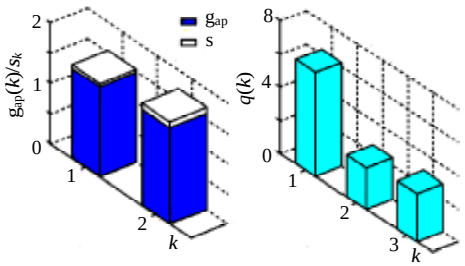

图 7 相互关联不良数据情况的 $g_{ap}(k)$与$q(k)$

Fig. 7　$g_{ap}(k)$ and $q(k)$ under the correlated bad data scenario

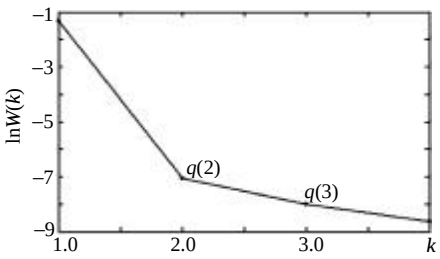

图 8 相互关联不良数据情况下的 ln$W(k)$曲线

Fig. 8　Curve of ln$W(k)$ under the correlated bad data scenario

4.6　GSA 与基于 GSA 的肘形判据的性能比较

本文在对 GSA 方法研究的基础上提出了基于 GSA 的肘形判据，现将 2 种方法通过仿真进行对比分析得各自的优缺点如下：

（1）基于 GSA 的数据挖掘可以自动确定最佳聚类数量，在聚类个数为 1 的情况下比以往方法优秀。

（2）基于 GSA 的肘形判据避开了参考数据分布的要求，使得计算的过程以及结果更加准确客观，在多个不良数据和相互关联不良数据出现时不会误判。

（3）在计算量方面，基于 GSA 的肘形判据的计算量集中在聚类个数为 1 和 2 的情况(GSA 算法部分)。在聚类个数 k>3 的时候只需对待检测数据进行一次聚类，计算量比 GSA 方法大为减少，实际运行测试中，最佳聚类个数不同时，2 种方法所用时间如表 5 所示。

表 5　GSA 方法与基于 GSA 的肘形判据计算速度比较

Tab. 5　Comparison of computing time of GSA and GSA-based elbow judgment

方法	计算时间/s	
	k=2	k=3
GSA	10.60~10.80	18.80~19.00
GSA 的肘形判据	4.00~4.10	4.20~4.30

5　结论

论文在对 GSA 数据挖掘技术应用于电力系统不良数据辨识的研究基础上提出了一种估计最佳聚类个数的快速判据——肘形判据。将之与 GSA 方法相结合形成基于 GSA 的肘形判据

文中将该方法应用于江苏局部电网实时数据的不良数据检测和辨识中，通过对几种不良数据情况的仿真可以发现：与传统的状态估计方法相比，该方法不受残差的影响，可以很好地避免残差污染和残差淹没。与 GSA 方法相比，该方法计算客观准确，未出现误判，而且计算速度大为提高。对于大系统、数据量巨大的情况，该方法是一种快速高效的算法，具有很好的应用前景。

参考文献

[1] 于尔铿．电力系统状态估计[M]．北京：水利电力出版社，1985．

[2] 杜正春，牛振勇，方万良．基于分快 QR 分解的一种状态估计算法 [J]．中国电机工程学报，2003，23(8)：50-55．
Du Zhengchun，Niu Zhenyong，Fang Wanliang．A block QR based power system state estimation algorithm[J]. Proceedings of the CSEE，2003，23(8)：50-55(in Chinese)．

[3] 刘广一，胡锡龙，于尔铿，等．快速正交变换阻尼最小二乘法在电力系统状态估计中的应用[J]．中国电机工程学报，1991，11(6)：34-40．
Liu Guangyi，Hu Xilong，Yu Erkeng，et al．Application of fast orthogonal transformation with damping factor to power system state estimation[J]．Proceedings of the CSEE，1991，11(6)：34-40(in Chinese)．

[4] 李钊年．电力系统状态估计中的不良数据辨识 [J]．青海大学学报(自然科学版)，2001，19(1)：49-51．
Li Zhaonian．Identification of bad data of electric power system state estimation[J]．Journal of Qinghai University，Science and Technology，2001，19(1)：49-51(in Chinese)．

[5] 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)：69-72，47．
Zhang Xingmin，Mao Yuhua，Zhu Jianfeng，et al．Detection and identification of multi-bad data using graph theory[J]．Proceedings of the CSEE，1997，17(1)：69-72，47(in Chinese)．

[6] Tibshirini R，Walther G，Hastie T．Estimating the number of cluster in a dataset via the gap statistic[R]．Unpublished Technical Report：Stanford University，2000：1-18．

[7] YuHui Luo，Jonathon C. Active source selection using gap statistic for underdetermined blind source separation[C]. Signal Processing and Its Applications 2003 Proceedings，Seventh International Sympsium，Paris，France， 2003：137-140.

[8] S J Huang，Jeu Min Lin. Enhancement of power system data debugging using GSA-based data-mining technique[J] . IEEE Transactions on power system，2002，17(11)：1022-1029.

[9] 葛成. 基于 GSA 的电力系统不良数据辨识方法研究[D]. 南京：南京理工大学，2005.
Ge Cheng. Study of power system data debugging using GSA-based data-mining technique[J]. Nanjing ：Nanjing University of Science and Technology，2005(in Chinese).

[10] 张斌. 基于 GSA 的数据挖掘在电力系统不良数据辨识中的应用[D]. 南京：南京理工大学，2003.
Zhang Bin. pplication of gsa-based data mining for identifying bad data of power system[D]. Nanjing ：Nanjing University of Science and Technology，2003(in Chinese).

[11] 史光荣，黄世杰，林矩民. 应用间隙统计法为辅之资料探勘技术于不良资料之检测[R]. 国立成功大学技术报告，2003.
Shi Guangrong，Huang Shijie，Lin Juming. Application of gap statistic in detecting material for bad material[R]. Technology Report of National Success University ，2003(in Chinese).

[12] 韩富春，王娟娟. 基于神经网络的电力系统状态估计[J]. 电力系统及其自动化学报，2002，14(6)：49-52.
Han Fuchun，Wang Juanjuan. State estimation in power system based on neural network[J]. Proceedings of the CSU-EPSA，2002，14(6)：49-52(in Chinese).

[13] 张国江，邱家驹，李继红. 基于人工神经网络的电力负荷坏数据辨识与调整[J]. 中国电机工程学报，2001，21(8)：104-107，113.
Zhang Guojiang，Qiu Jiaju，Li Jihong. Outlier identification and justification based on neural network[J]. Proceedings of the CSEE，2001，21(8)：104-107，113(in Chinese).

[14] Souza J C，Silva A P. Data visualization and identification of anomalies in power system state estimation using artificial neural networks [J]. Eng. Gen. Transm. dist.，1997，44 (9)：445-455.

[15] S J Huang，Jeu Min Lin. Artificial Neural Network Enhanced by Gap Statistic Algorithm Applied for Bad Data Detection of a Power System[C]. Transmission and Distribution Conference and Exhibition 2002，Tokyo，Japan，2002：764-768..

[16] Salehfar H， Zhao R. A neural network pre-estimation filter for bad-data detection and identification in power system state estimation [J]. Electric power system research. 1995，34 (8)：127-134.

收稿日期：2006-05-02。

作者简介：

吴军基(1955—)，男，教授，博士生导师，从事电力系统运行调度与控制、电力市场等方面的教学与研究工作；

杨 伟(1965—)，男，副教授，从事电力系统运行调度与控制方面的教学与研究工作，weiyang@mail.njust.edu.cn。

（责任编辑 喻银凤）