基于ICA的异常数据挖掘算法研究

王莉君[1,2,3]，何政伟[1]，冯平兴[3]

(1. 成都理工大学地质灾害防治与地质环境保护国家重点实验室 成都 610059；
2. 成都理工大学地球物理学院 成都 610059；3. 电子科技大学成都学院 成都 610051)

【摘要】在传统的独立成分分析方法中，没有考虑异常数据值对分离性能的影响。该文提出了一种基于影响函数的检测方法，通过该方法可以发现隐藏在观测数据中的异常成分。利用影响函数对数据进行投影分析，对混入脉冲噪声的观测信号进行盲源分离，从而实现对脉冲噪声的消除。实验仿真结果表明，该方法可以有效且可靠地检测出所观察信号中的异常数据。

关 键 词 异常数据挖掘； 盲源分离； 脉冲噪声； 独立分量分析； 信号处理
中图分类号 TP391，TN911.7 **文献标志码** A doi:10.3969/j.issn.1001-0548.2015.02.009

Study of Outlier Data Mining Algorithm Based on ICA

WANG Li-jun[1,2,3], HE Zheng-wei[1], and FENG Ping-xing[3]

(1. State Key Laboratory of Geohazard Prevention and Geoenvironment Protection, Chengdu University of Technology Chengdu 610059;
2. College of Geophysics, Chengdu University of Technology Chengdu 610059;
3. Chengdu College of University of Electronic Science and Technology of China Chengdu 610051)

Abstract In the traditional study of independent component analysis (ICA), the outlier data had not been considered. This paper proposes a method based on influence function to find the outliers from the observed data in ICA. General, outliers have a significant influence on the separation performance of ICA. Using the influence functions to project the observed data, the impulsive noisy components which mixed in the observed data can be eliminated from the normal data. The experimental results demonstrate the effectiveness of proposed method.

Key words abnormal data mining; blind source separation; impulse noise; independent component analysis; signal processing

独立成分分析(independent component analysis, ICA)是文献[1]提出的一种重要的盲源分离方法。该方法基于反馈神经网络，仅能用于两个混迭源信号的分离。文献[2]提出了一种解决非线性混迭信号盲分离问题的算法，文献[3-6]进一步研究了非线性混迭信号盲分离。传统的ICA方法没有考虑异常数据的影响[7-17]。而异常数据检测在信号诊断、财务监控、网络入侵检测、贷款审批等很多领域有重要用途。

异常数据检测的方法有多种，目前常用的方法大致有以下4种。

1) 基于统计模型。通过数据的变异指标发现数据中的异常点，如：极差、均差、四分位数距离等。变异指标的值越大表示变异越大、散布越广；值越小表示离差越小，越密集。

2) 基于距离模型。该方法避免了过多的计算问题，不依赖统计检验，将不具有多个“邻居”的对象检测出来。基于单元、索引的算法和嵌套-循环算法都是属于目前比较成熟的基于距离模型的异常数据挖掘算法。

3) 基于密度模型。计算对象的局部异常因子越大，发生异变的可能性越大。

4) 基于偏离模型。该方法模拟人的思维方式，通过对一个连续序列的观察，发现其中个别数据与其他数据的不同。常采用序列异常技术和OLAP数据立方体技术。

在异常值的常规研究中，以上方法是检测异常数据集最重要的方法。基于统计学方法针对单个属性的数据，而数据挖掘问题要求在多维空间中发现异常点。当没有特定的分布检验时，检测出所有的异常点数据非常困难。而基于距离的异常数据挖掘

收稿日期：2014 – 10 – 09；修回日期：2015 – 01 – 12
基金项目：高等学校博士学科点专项科研基金(20095122110003)；地质灾害防治与地质环境保护国家重点实验室开放基金(SKLGP2011Z005)；四川省教育厅自然科学项目(12ZB233)
作者简介：王莉君(1983 –)，女，博士生，主要从事数据挖掘方面的研究.

方法要求用户多次试探设置参数。基于偏差的异常数据挖掘方法对实现复杂数据的效果不佳，这类方法往往不能检测误差较小的点。因此，上述方法不适合用于ICA数据流，特别是数据量大的多维数据流。

本文提出了一个有效的异常值检测技术，该方法主要基于影响函数并对观测数据进行投影分析，从而发现数据中的异常值。

1 算法原理

假设存在 n 个随机变量，这些随机变量是在不考虑噪声的情况下，用 n 个未知独立源信号与混合矩阵 $\boldsymbol{A}$ 经过线性混合而得到，如图 1 所示。ICA 实际是对非高斯数据的一种线性变换，使得输出的分量之间是统计独立或尽可能独立。即它是依赖于源信号 $\boldsymbol{s}$ 彼此独立的条件完成分离任务，并寻求一个分离矩阵 $\boldsymbol{W}$，将其作用于观测信号 $\boldsymbol{x}$，估计出源信号 $\boldsymbol{y}$，故有：

$$\boldsymbol{y}(\boldsymbol{t}) = \boldsymbol{WAs}(t) \tag{1}$$

式中，$\boldsymbol{W}$ 为正交的分离矩阵，被定义为：

$$\mathrm{E}[(\boldsymbol{Wx})\cdot(\boldsymbol{Wx})^{\mathrm{T}}] = \boldsymbol{I} \tag{2}$$

$$\boldsymbol{WW}^{\mathrm{T}} = \boldsymbol{I} \tag{3}$$

式中，E(·) 表示数学期望；T 表示转置。

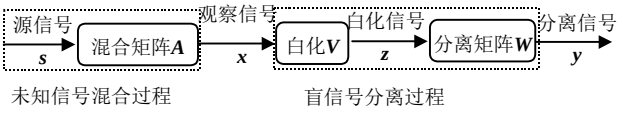


图1 盲信号分离系统模型

由于源信号 $\boldsymbol{s}(t)$ 和分离矩阵 $\boldsymbol{W}$ 都是不确定参数，在没有先验知识的情况下无法同时确定出这两个相关参数。优化分离矩阵 $\boldsymbol{W}$ 通过某种算法使得分离信号 $\boldsymbol{y}(t)$ 各分量之间相互独立性最大，以获得估计的源信号。

在 ICA 分析中，异常值可能出现在源信号或者观察到的信号中，噪声观测信号的模型为：

$$\boldsymbol{y}(t) = \boldsymbol{WA}[\boldsymbol{s}(t) + \boldsymbol{N}] \tag{4}$$

式中，$\boldsymbol{N} = [\boldsymbol{N}_1, \boldsymbol{N}_2, \cdots, \boldsymbol{N}_n]^{\mathrm{T}}$ 可以被视为固定的随机噪声的异常值。

如果 $\boldsymbol{N}$ 为脉冲噪声，它可以表示为：

$$\boldsymbol{N} = \sum_{i=1}^{N} b_i \delta(n_i) \tag{5}$$

式中，b_i 为脉冲噪声的振幅。脉冲噪声的特征具有近似平坦的频谱，作用时间短、幅值大。其产生原因主要由交流供电线路的干扰、继电器等引起的瞬时干扰。

然而，如果源信号中混入脉冲噪声信号，采用独立成分分析的传统方法不再有效。脉冲噪声可损害所观察到信号的统计学特性，此时，如果想让一个很小的残余点被显示出来，在使用传统的 ICA 方法之前，应采用以下方式对观测数据进行预处理操作。

为了分析 ICA 观测数据中的异常值，影响函数被定义为[7]：

$$\mathbf{IF}(\boldsymbol{z}, \hat{\theta}) = \boldsymbol{B}\psi(\boldsymbol{z}, \hat{\theta}) \tag{6}$$

$$\boldsymbol{B}\psi(\boldsymbol{z}, \hat{\theta}) = \boldsymbol{B}\begin{pmatrix} \boldsymbol{z}g(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{z}) + e\lambda\boldsymbol{w} \\ \|\boldsymbol{w}\|^2 - 1 \end{pmatrix} \tag{7}$$

式中，e 是一个常数；$g(\cdot)$ 是非多项式函数；$\boldsymbol{B}$ 是一个不依赖于 $\boldsymbol{z}$ 的不相干的可逆矩阵，$\boldsymbol{z}$ 是已预处理过的随机矢量：

$$\boldsymbol{z} = \boldsymbol{V}\overline{\boldsymbol{x}} \tag{8}$$

设 $\boldsymbol{C}_X = \tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^{\mathrm{T}}$ 是协方差矩阵，$\boldsymbol{E} = (e_1, e_2, \cdots, e_n)$ 是特征向量，它是包含 $\boldsymbol{C}_X$ 的 n 个最大特征值相对应的 n 个主特征向量，由其特征向量张成的子空间为信号子空间；$\boldsymbol{D} = \mathrm{diag}(d_1, d_2, \cdots, d_n)$ 为 $n \times n$ 维对角矩阵，包含有协方差矩阵 $\boldsymbol{C}_x$ 的 n 个最大特征值[7]。预白化操作本质上是一种去相关处理，因而可以考虑使用 PCA 算法[8-9]，因此白化矩阵可以写成：

$$\boldsymbol{V} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{E}^{\mathbf{T}} \tag{9}$$

$\tilde{\theta} = (\lambda, \boldsymbol{z})$ 和 λ 是与约束 $\|\boldsymbol{w}\| = 1$ 相关联的拉格朗日乘数[10]。

因此，有：

$$\left\|\mathbf{IF}(\boldsymbol{z}, (\lambda, \boldsymbol{w}))\right\|^2 = K_0\left[|\boldsymbol{z}|\, g(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{z})\right]^2 \tag{10}$$

式中，K_0 是一个常数。

为简化分析，本文考虑两个信号源的情况。假设 $\boldsymbol{z}_i = (z_{1i}, z_{2i})^{\mathrm{T}}$ 是含异常值的观测数据，可用非线性函数对其进行投影变换。因此，异常点检测可以通过分析 $\left\|\mathbf{IF}(\boldsymbol{z}, (\lambda, \boldsymbol{w}))\right\|^2$ 的值实现。在 ICA 的传统研究中，$\left\|\mathbf{IF}(\boldsymbol{z}, (\lambda, \boldsymbol{w}))\right\|^2$ 的值是用来测量当观测信号与异常值混合的估算值的鲁棒性。从式(10)可以看出，其值主要取决于[(·)g(·)]。然而，鲁棒性不仅仅是在异常值检测中要考虑的问题，所不同的是在异常值的检测中使用的非多项式函数的鲁棒性不应太好。

如果选用一个具有较强鲁棒性的非多项式函数，根据 $\left\|\mathbf{IF}(\boldsymbol{z}, (\lambda, \boldsymbol{w}))\right\|^2$ 的值将异常值从正常值中区分出来非常困难。

适合被选用的非多项式函数应该具有以下特征：1) 估算 $g(\cdot)$ 不应该有统计上的困难。2) 为了区分异常值和正常值，$g(\cdot)$ 应该具有适当的鲁棒性，但不应该对异常值有强烈的鲁棒性。3) $g(\cdot)$ 必须捕

捉到 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{z}$ 的分配中与使用预测数据计算 $\left\|\mathbf{IF}(\boldsymbol{z},(\lambda,\boldsymbol{w}))\right\|^2$ 相关的所有方面。

通过实验，本文选择一些适合的非多项式函数：

$$g_1(\xi)=a\xi \tag{11}$$

$$g_1(\xi)=b(1-\tanh^2(b\xi)) \tag{12}$$

$$g_3(\xi)\begin{cases}\beta\xi & \xi\leqslant C\\ \xi^2+\beta C-C^2 & \xi>C\end{cases} \tag{13}$$

式中，a 为常数；$\xi=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{z}$。通过实验发现，当 $1\leqslant a\leqslant 4$，$1\leqslant b\leqslant 2$，$0.5\leqslant\beta\leqslant 2$ 和 $1\leqslant C\leqslant 2$ 时能提供良好的近似值。

异常值的检测阈值定义为：

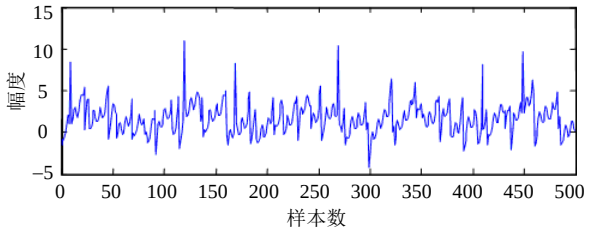
$$\text{threshold}=\sqrt{\sum_{i=}^{N}(\mathbf{IF}(\boldsymbol{z},(\lambda,\boldsymbol{w})))_i^4} \tag{13}$$

式中，$\mathrm{IF}(\boldsymbol{z},(\lambda,\boldsymbol{w}))_i$ 是影响函数使用实测数据作为参数得到的振幅。

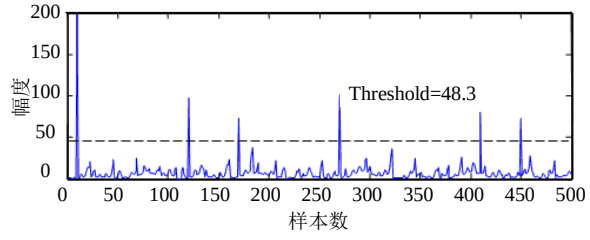
2　实验与结果分析

为了验证该方法的有效性，本文从两个方面进行验证。仿真实验和非多项式函数所选用的参数都相同，所有的信号都具有零均值和单位方差。将正弦波信号、三角波信号和锯齿波信号(各自幅度分别为1.5,1,1)进行混合，并在已混合的信号中加入脉冲噪声信号，得到观测信号。

考虑源信号未混入与已混入高斯噪声这两种情况，通过对观察的信号和预测的非多项式函数信号之间进行比较，判断影响函数是否能满足要求。通过仿真，发现基于影响函数的异常值检测方法可以有效地找出混合信号中的异常值，如图2、图3所示。

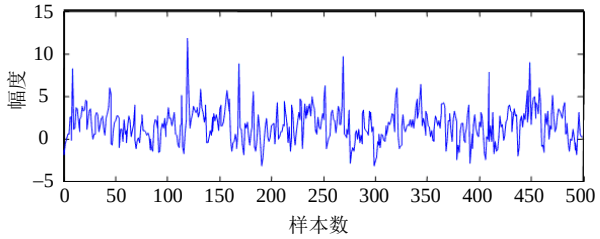


a. 无高斯噪声观测信号

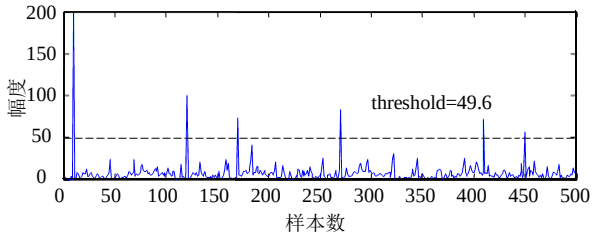


b. 利用函数对无高斯噪声观测信号进行投影的结果

图2　异常值检测的仿真结果1



a. 混入高斯噪声的观测信号



b. 利用函数对混入高斯噪声观测信号进行投影的结果

图3　异常值检测的仿真结果2

图2是观测信号不与高斯白噪声混合的情况下，利用 $g_2(\xi)=b(1-\tanh^2(b\xi))$ 对其进行投影，并计算出异常值的检测阈值 Threshold = 48.3。通过对比发现，该方法可以有效地找出混合信号中的异常值。通过这种方式，可以在分析观察到的混合信号中确定出异常值的位置。

图3是观测信号与幅度为1的高斯白噪声混合的情况下，利用 $g_2(\xi)=b(1-\tanh^2(b\xi))$ 对其进行投影，并计算出异常值的检测阈值 Threshold = 49.6，通过对观测信号是否混入高斯噪声的投影结果对比，在高斯噪声存在的情况下，仍然可以从观测信号中分离出异常值。

3　结 束 语

本文介绍了一种从观测信号中检测异常值的方法，该方法基于所观察到的信号的影响函数。不管在异常值检测中使用什么方法，主要目的是检测一组输入数据中的异常值，而不是预测新输入数据的异常值。与其他方法相比，本文所提出的异常检测方法可以避免计算的复杂性和先验知识的约束，采用非多项式函数查找观测数据的异常值，并通过实验结果证明了该方法能够有效地找到混合在ICA的观测数据中的异常值。

参　考　文　献

[1] HERAULT J, JUTTEN C. Space or time adaptive signal processing by neural network models[C]//AIP Conference Proceedings. [S.l.]: [s.n.], 1986: 151-206.

[2] BUREL G. Blind separation of sources: a nonlinear neural algorithm[J]. Neural Networks, 1992, 5(6): 937-947.

[3] PARRA L, DECO G, MIESBACH S. Statistical independence and novelty detection with information preserving nonlinear maps[J]. Neural Computation, 1996, 8(2): 260-269.

[4] YANG H H, AMARI S. Adaptive online learning algorithms for blind separation: maximum entropy and minimum mutual information[J]. Neural Computation, 1997, 9(7): 1457-1482.

[5] HYVÄRINEN A, OJA E. Independent component analysis: Algorithms and applications[J]. Neural Networks, 2000, 13(4): 411-430.

[6] DELORME A, MAKEIG S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis[J]. Journal of Neuroscience Methods, 2004, 134(1): 9-21.

[7] CARDOSO J F, SOULOUMIAC A. Blind beamforming for non-Gaussian signals[J]. IEE Proceedings F (Radar and Signal Processing), 1993, 140(6): 362-370.

[8] BELL A J, SEJNOWSKI T J. An information-maximization approach to blind separation and blind deconvolution[J]. Neural Computation, 1995, 7(6): 1129-1159.

[9] BELOUCHRANI A, ABED-MERAIM K, CARDOSO J F, et al. A blind source separation technique using second-order statistics[J]. Signal Processing, IEEE Transactions on, 1997, 45(2): 434-444.

[10] 史习智. 盲信号处理: 理论与实践[M]. 上海: 上海交通大学出版社, 2008.
SHI Xi-zhi. Blind signal processing-theory and practice[M]. Shanghai: Shanghai Jiao Tong University Press, 2008.

[11] HECKERLING P S. Parametric receiver operating characteristic curve analysis using mathematica[J]. Computer Methods and Programs in Biomedicine, 2002, 69(1): 65-73.

[12] 张兰勇, 刘繁明, 李冰. 基于聚谱分析的多通道盲信号自适应分离算法[J]. 电子与信息学报, 2014, 36(1): 158-163.
ZHANG Lan-yong, LIU Fan-ming, LI Bing. Multichannel blind signal adaptive separation algorithm based on polyspectra analysis[J]. Journal of Electronics & Information Technology, 2014, 36(1): 158-163.

[13] ANGIULLI F, FASSETTI F. Distance-based outlier queries in data streams: the novel task and algorithms[J]. Data Mining and Knowledge Discovery, 2010, 20(2): 290-324.

[14] BANERJEE R. Fair m-estimators as a cost function for FASTICA[C]//Signal Processing and Communication (ICSC), 2013 International Conference on. [S.l.]: IEEE, 2013: 445-448.

[15] ALI R, ZAHRAN O, ELKORDY M, et al. Blind source separation for different modulation techniques with wavelet denoising[J]. Digital Signal Processing, 2013, 5(12): 418.

[16] KE-LIN D U, SWAMY M N S. Neural networks and atatistical learning[M]. London: Springer, 2014.

[17] XU Bing-lin, LI Zhan-huai.An anomaly detection method for spacecraft using ICA technology[C]//International Conference on Advanced Computer Science and Electronics Information. Beijing: [s.n.], 2013: 50-54.

编 辑 税 红