Operations Research and Fuzziology 运筹与模糊学, 2023, 13(1), 139-144
Published Online February 2023 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2023.131016

Hans汉斯

基于时间序列异常检测分析的方法

瞿杏元[1]，曹忠虔[2]

[1]四川建筑职业技术学院数学教研室，四川 德阳
[2]华为技术有限公司成都研究所，四川 成都

收稿日期：2022年11月11日；录用日期：2023年2月2日；发布日期：2023年2月9日

摘 要

不依赖于模型，基于累计变化量来实现异常点的检测，而无法检测出成片的异常点，并且很容易把正常点视为异常点，为了解决上述方法中所存在的问题，本文在所给方法的基础上重新定义了累计变化量，引入了推移算子，异常类型指示变量和异常惩罚量，并定义了两类异常类型，一种叫做高位异常，一种叫做低位异常，然后重新定义了异常点模型，引入了然后用2004年到2009年的沪市股票数据来进行数值实验，并对结果进行了对比证明了本文所给方法的有效性。

关键词

异常检测，异常点，高位异常，低位异常，时间序列

Method of Anomaly Detection and Analysis Based on Time Series

Xingyuan Qu[1], Zhongqian Cao[2]

[1]Sichuan College of Architectural Technology, Deyang Sichuan
[2]Chengdu Research Institute of Huawei Technologies Co., Ltd, Chengdu Sichuan

Received: Nov. 11th, 2022; accepted: Feb. 2nd, 2023; published: Feb. 9th, 2023

Abstract

About not dependent on the model and is relatively simple and easy to implement about the methods of the time series anomaly detection, but it can not detected a piece of outliers, and it is easy to make normal points as outliers. In order to solve the problems, on the basis of the method given, this paper redefines the cumulative change and introduces the transition operator, one indicator variable and unusual punishment are introduced in this paper and two exception types are defined, one is called the high anomaly, another is called the low abnormal. The effectiveness of

文章引用: 瞿杏元, 曹忠虔. 基于时间序列异常检测分析的方法[J]. 运筹与模糊学, 2023, 13(1): 139-144.
DOI: 10.12677/orf.2023.131016

the method given in this article is proved by using the data from Shanghai Stock Market between 2004 and 2009 be proved through numerical experiments. The results are compared to prove the effectiveness of the method presented in this paper.

Keywords

Anomaly Detection, Outlier, High Anomaly, Low Anomaly, Time Series

1. 引言

在异常检测研究领域，经过了20多年的发展，已经产生了很多有效的方法，如统计学方法[1]，基于密度的方法，基于距离的方法。以上都是针对无序序列的异常点检测，由于时间序列的最重要一个特征就是具有时间属性，序列值之间必须按照时间先后顺序进行严格的排序，因此上面介绍的方法都不适合用于时间序列。目前针对时间序列中异常检测的方法主要有生物学方法[2]，机器学习的方法[3] [4]，基于小波的方法[5] [6]，基于AR模型的方法[7]等。虽然在对时间序列异常检测研究领域里已经产生很多种方法，但是这些方法还不是很成熟。并且目前对于时间序列异常点检测的方法大都需要在预先知道所给数据满足哪一种时间序列模型的基础上针对相应的模型来进行异常点检测，但是在一般情况下很难知道数据到底满足哪一种时间序列模型，而且，一些针对时间序列异常检测的方法很复杂不太容易实现，比如基于贝叶斯方法的异常检测，理论看起来很复杂，不仅需要抽样还要经过反复的迭代。

通过前面对时间序列异常检测的研究，结合文献[8]中的异常点检测方法给出了一种改进的模型，为了解决文献[8]中的模型不能检测出成片的异常点还有把正常点误以为是异常点的情况，在文中定义两种异常形式，一种叫做高位异常，一种叫做低位异常，并引入一个表示异常类型的示性变量和惩罚函数，最后借助计算机将模型应用于生活中实际数据。

2. 基础知识

定义 1：设$\{\varepsilon_t\}$是$WN(0,\sigma^2)$，且实系数多项式$A(z)$和$B(z)$没有公共根，满足以下所给的条件

$$b_0=1, a_p b_q \neq 0$$

$$A(z)=1-\sum_{j=1}^{p} a_j z^j \neq 0, |z| \leq 1$$

$$B(z)=\sum_{j=0}^{q} b_j z^j \neq 0, |z|<1,$$

那么我们就把下面的差分方程

$$X_t=\sum_{j=1}^{p} a_j X_{t-j}+\sum_{j=0}^{q} b_j \varepsilon_{t-j}, t \in Z$$

称为自回归滑动平均模型，简称为$\mathrm{ARMA}(p,q)$模型。

定义2：给出一时间序列，$X=(x_1=(v_1,t_1),x_2=(v_2,t_2),\cdots,x_n=(v_n,t_n))$点$x_{t_i}=(v_{t_i},t_i)$表示时间序列在$t_i$

时刻的观测值为 v_{t_i}。用 $\left(N_1,N_2,\cdots,N_k\right)$ 表示点 x_{t_i} 的k个邻居点集合，其观测值集合记为 $\left(v_{N_1},v_{N_2},\cdots,v_{N_k}\right)$，给定阀值$T$，若点 x_{t_i} 与其k个邻居点的累积变化量(Accumulative Change)大于T，则判定点 x_{t_i} 为这段时间序列中的一个异常点，这一判定条件公式表示为

$$\text{Accumulative Change}=\frac{W_1\left|v_{t_i}-v_{N_1}\right|+W_2\left|v_{t_i}-v_{N_2}\right|+\cdots+W_k\left|v_{t_i}-v_{N_k}\right|}{W_1+W_2+\cdots+W_k}>T$$

式中的 $\left(W_1,W_2,\cdots,W_k\right)$ 为权值向量，赋予每个变化量不同的权重，一般来说，在时间轴上，越接近点 x_{t_i} 的邻居点赋予的权值越大；阀值 T 是用户给定的一个常数量，点 x_{t_i} 的累积变化量和阀值的大小关系，是判定 x_{t_i} 是否为一个异常点的依据[8]。

3. 异常点模型建立

文献[8]中利用上面定义 2 判定的异常点模型是通过时间序列的波动来得到的异常点，因此不需要知道时间序列具体是符合什么样的模型，而且这种异常点模型也容易在现实应用中实现，但是也存在一定的缺陷：

1) 由于是通过时间序列的波动来实现异常点的检测，因此，在波动量不是很大但却很频繁的时候，累积变化量也可能很大。

2) 在异常点成片出现的时候，因为不知道异常点成片出现时的数目，因此 T 的选择对异常点的检测很重要，这个异常点模型往往识别不了成片的异常点，因为可能成片的异常点之间的变化不大，从而造成了它们间的累积变化量不是大。

文献[8]中的方法在经过了一片异常点后，往往会把第一个出现的正常值视为异常点，因为在成片异常点后面的正常数据相对异常点的变化很大，所以累积变化量也很大。因此本文所定义的异常点思想是居于文献[8]中累积变化量的思想，然后试图重新对异常点模型进行定义，从而解决文献中对异常点定义所存在的缺点。

在定义异常数据点模型前,先给出高位异常点和低位异常点的定义。

定义 3 如果一个异常点的数据比它相邻的数据大很多，则称这个异常点为高位异常点。

定义 4 如果一个异常点的数据比它相邻的数据小很多，则称这个异常点为低位异常点。

为了解决文献[8]中的问题，引入 $\lambda_i\left(i=1,2,\cdots,k\right)$，称 λ_i 为数据的异常类型和异常惩罚量 C。其中 λ_i 的取值如下：

1) 当某一数据点为高位异常点时，那么 $\lambda_i=1$；

2) 当某一数据点为低位异常点时，那么 $\lambda_i=-1$；

3) 当数据点为正常数据点时，那么 $\lambda_i=0$。

本文中重新定义累积变化量为：

$$AC=\left|\frac{\sum_i^k W_i\left[\left(x_t-x_{t-i}\right)+\lambda_i\cdot C\right]}{W_1+W_2+\cdots+W_k}\right|$$

然后引入推移算子 B，则模型可写成：

$$AC=\left|\frac{\sum_i^k W_{i\,i}\left[x_t\left(1-B^i\right)+\lambda_i\cdot C\right]}{W_1+W_2+\cdots+W_k}\right|$$

异常点模型为：

$$AC=\left|\frac{\sum_i^k W_i\left[x_t\left(1-B^i\right)+\lambda_i\cdot C\right]}{W_1+W_2+\cdots+W_k}\right|>T$$

C 和 T 需要用户确定，当 $AC>T$ 时，则认为 x_t 数据点为异常点，为了知道异常点是偏大还是偏小，定义如下：当

$$\frac{\sum_i^k W_i\left[x_t\left(1-B^i\right)+\lambda_i\cdot C\right]}{W_1+W_2+\cdots+W_k}>0$$

则，异常点为高位异常点，并且该点的 $\lambda_i=1$。当

$$\frac{\sum_i^k W_i\left[x_t\left(1-B^i\right)+\lambda_i\cdot C\right]}{W_1+W_2+\cdots+W_k}<0$$

则，异常点为低位异常点，并且该点的 $\lambda_i=-1$。当

$$\left|\frac{\sum_i^k W_i\left[x_t\left(1-B^i\right)+\lambda_i\cdot C\right]}{W_1+W_2+\cdots+W_k}\right|<T$$

则，该点为正常数据点，并且该点的 $\lambda_i=0$。

4. 模型分析

本文中给出的模型相比较于文献[8]中的模型缺陷，有如下优势：

$$\left|\frac{\sum_i^k W_i\left[x_t\left(1-B^i\right)+\lambda_i\cdot C\right]}{W_1+W_2+\cdots+W_k}\right|<T$$

1) 因为文献[8]中的累积变化通过 $\sum_i^k W_i\left|v_t-v_{t-i}\right|$ 的值来表示的，因此，即使波动量不大，但是波动得很频繁的时候 $\sum_i^k W_i\left|v_t-v_{t-i}\right|$ 的值都会很大，而且，本文认为那些异常点莫非是相对于邻近点过大或者是过小的数据点，因此用 $\sum_i^k W_i\left(v_t-v_{t-i}\right)$ 来描述某一数据点对于邻近点的变化会更好一些。

2) 对于惩罚量的引入：因为文献[8]中的异常点模型是通过 $\sum_i^k W_i\left|v_t-v_{t-i}\right|$ 的值来确定的，因此在出现成片异常点的地方 $\sum_i^k W_i\left|v_t-v_{t-i}\right|$ 会很小，从而导致检测不出异常，本文引入的惩罚量就是当 t 时刻的数据点是异常点时它会对后续的数据产生影响，因此即使在出现成片异常点的情况下，当引入异常惩罚量之后，即使异常点之间真实的波动量并不大，但是本文定义的累积变化也会很大，这样就可以检测出成片的异常点。

3) 对于指示变量的引入，由于上面加入了异常惩罚量，但是如果每个异常所加的惩罚量都是正或是负的话也会出现问题，假设惩罚量全部都取为正数，那么，当遇到成片的异常点时并且这些异常点会比正常的数据点数值要低，则在成片异常点后面出现的正常点往往就会被认为是异常点，再假设惩罚量全部都取为负数，当遇到成片的异常点时并且这些异常点会比正常的数据点数值要高，则在成片异常点后面出现的正常点往往也会被认为是异常点，并且，引入具有三个取值的指示变量还有一个好处，就是不仅知道某个数据点是否异常而且还可以知道异常点是偏高了还是偏低了。

5. 数值试验

结合实际数据进行试验，所用的数据是 2004 年到 2009 年的沪市股票数据一共有 302 条数据。为了了解由于异常点的存在对时间序列数据建立模型的影响，下面将会对所给的数据在对异常点进行修正前

后来进行拟合模型的对比，主要看所拟合的模型的参数在修正异常点前后的差异，因为一般的金融数据都会有异方差的，在这先对数据进行异方差的检验，然后用数据进行 GRACH 模型的检验。

首先对数据进行了异方差检验，异方差检验结果如图 1 显示：

Q and LM Test for ARCH Disturbances				
Order	Q	Pr>Q	LM	Pr>LM
1	284.282	<0.0001	281.8075	<0.0001
2	542.3052	<0.0001	282.6534	<0.0001
3	769.1152	<0.0001	283.2137	<0.0001
4	966.3574	<0.0001	283.2164	<0.0001
5	1140.1721	<0.0001	283.5321	<0.0001
6	1294.0647	<0.0001	283.5321	<0.0001
7	1433.0334	<0.0001	283.5823	<0.0001
8	1560.0976	<0.0001	283.5939	<0.0001
9	1678.1196	<0.0001	283.6304	<0.0001
10	1785.1553	<0.0001	284.0491	<0.0001
11	1881.1165	<0.0001	284.0514	<0.0001
12	1965.1592	<0.0001	284.0622	<0.0001

Figure 1. Heteroscedasticity test results
图 1. 异方差检验结果

从上面图中可以看到，Q 统计量从 1 到 12 的时滞窗体现出了方差随时间的变化，这些检验都显示数据存在着异方差。因此，用 GRACH 模型对数据进行建模，下面得到了 GRACH 模型的参数估计，如图 2。

The AUTOGER Procedure					
GRACH Estimates					
	SSE	797869020	Observat Ions	302	
	MSE	2641950	Uncond Var	3371050.42	
	Log LikeLihood	-2656.2166	Total R-Square	0.9889	
	SBC	5346.69579	AIC	5324.43334	
	Normality Test	2363.0905	Pr>ChiSq	<0.0001	

Variable	DF	Estimate	Standard Error	t Value	Approx Pr>|t|
Intercept	1	221.5	772.0174	2.93	0.0034
date_n	1	137.438	42.9008	3.2	0.0014
AR1	1	-1.1321	0.1056	-10.72	<0.0001
AR2	1	0.1413	0.105	1.35	0.1783
ARCH0	1	2606174	0.2593	1.00E+07	<0.0001
ARCH1	1	0.2269	0.0603	3.76	0.0002
GARCH1	1	5.10E-20	2.77E-09	0	1

Figure 2. Parameter estimation of GRACH model
图 2. GRACH 模型的参数估计

用本文给出的异常检测方法对数据进行异常点的检测，然后与文献[8]中的方法进行对比。最后在异常点进行修正后对数据进行建模。我们的异常点判断模型为：

$$AC = \left| \frac{\sum_i^k W_i \left[x_t \left(1 - B^i\right) + \lambda_i \cdot C \right]}{W_1 + W_2 + \cdots + W_k} \right| > T$$

异常修正后的参数估计结果如图 3 所示：

The AUTOGER Procedure

GRACH Estimates

	SSE	274722753	Observat Ions	302	
	MSE	1360014	Uncond Var	.	
	Log LikeLihood	-1674.0209	Total R-Square	0.9938	
	SBC	3379.8914	AIC	3360.04179	
	Normality Test	4353.7799	Pr>ChiSq	<0.0001	

Variable	DF	Estimate	Standard Error	t Value	Approx Pr>|t|
Intercept	1	-2652	793.9498	-3.34	0.0008
N	1	216.9053	288.4898	0.75	0.4521
AR1	1	-1.3482	0.0881	-15.3	<0.0001
AR2	1	0.3459	0.0897	3.86	0.0001
ARCH0	1	933686	0.3051	3.06E+06	<0.0001
ARCH1	1	1.7015	0.5009	3.4	0.0007
GARCH1	1	3.24E-18	2.39E-10	0	1

Figure 3. Parameter estimation after anomaly correction
图 3. 异常修正后的参数估计

从图中可以看出，修正前与修正后的参数估计结果图相比，对异常点修正前后就行的建模，参数差别很大，因此，更加说明了异常检测的重要性。

6. 结语

在本文给出的异常点模型中引入了一个异常惩罚量，而且在数值试验时把它取值为滑动窗口数据的平均值，本文给出一个比较简单易于实施而且有效的异常点发现方法，该方法是基于文献[8]的基础，然后通过引入指示变量和惩罚函数，在最后的数值试验中也证明该方法的有效性。

参考文献

[1] 张保稳, 何华灿. 时态数据挖掘研究进展[J]. 计算机科学, 2002, 29(2): 124-126, 103.

[2] 钱昱, 郑斌. 基于时序模式的异常检测[J]. 微机发展, 2004, 14(9): 53-55.

[3] 杨虎, 王会琦, 程代杰. 基于时间序列异常数据挖掘[J]. 计算机科学, 2004, 31(4): 117-119.

[4] 向旭, 蒋静坪. 时间序列的符号化方法研究[J]. 模式识别与人工智能, 2007, 20(2): 154-161.

[5] 李爱国, 覃征, 贺升平. 时间序列数据的相似模式抽取[J]. 西安交通大学学报, 2002, 36(12): 275-1278.

[6] Ester, M., Kriegel, H.P. and Sander, J., *et al.* (1996) A Density Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. AAAI Press.

[7] Agrawal, R., Imielinski, T. and Swami, A. (1993) Mining Association Rules between Sets of Items in Large Databases. Management of data.

[8] 林森. 时间序列异常检测的研究与应用[D]: [硕士学位论文]. 南京: 河海大学, 2008.