基于一类分类的时间序列异常值检测

孙德山[1] 吴今培[2] 肖健华[2]
[1](辽宁师范大学数学系,大连 116029)
[2](五邑大学智能技术与系统研究所,江门 529020)
E-mail:sdeshan@163.com

摘 要 论文结合相空间重构理论与一类分类方法提出一种时间序列中的异常值检测方法。该方法首先将时间序列映射到相空间,然后对相空间中的点实行一类分类,最后,根据KKT条件进行异常值检测。仿真实验结果表明了所给方法的可行性和有效性。

关键词 异常值 一类分类 时间序列 相空间

文章编号 1002-8331-(2003)34-0011-03 **文献标识码** A **中图分类号** TP18

Outliers Detection in Time Series Based on One-class Classification

Sun Deshan[1] Wu Jinpei[2] Xiao Jianhua[2]
[1](Department of Math., Liaoning Normal University, Dalian 116029)
[2](Institute of Intelligent Technology & System, Wuyi University, Jiangmen 529020)

Abstract: A new method of outlier detection in time series is proposed in this paper, which is based on phase space reconstruction theory and one-class classification method. The method first maps time series to phase space, and then obtains a distinguish function by one-class classification in phase space, finally, outliers detection is proceed based on KKT conditions. The results of simulation experiments show the feasibility and effectiveness of the method.

Keywords: outlier, one-class classification, time series, phase space

1 引言

在实际工作中,我们观察的数据往往是一组相依有序的离散数据集合,称之为时间序列。时间序列通常包含着大量的信息,是建模和预测的主要依据。若序列中含有异常值,就会使传统的建模、估计及检验方法陷入困境,从而给不出准确的预测和控制。因此,近年来关于时间序列中的异常值检测问题受到统计学界的重视,传统的检测方法大多针对ARMA模型展开[1,2]。传统的模型方法在检测线性时间序列中的异常值时效果是很好的,并且模型具有很好的解释性,但这些方法还很难应用于较复杂的非线性时间序列的异常值检测中。

以统计学习理论为基础建立起来的支持向量机是目前人们研究的热点,它已广泛应用于解决分类和回归问题[3,4]。支持向量机通过引入核函数来克服维数灾难,并很好地解决了非线性以及局部极小等问题。另外,同样引入核函数的一类分类方法[5]在异常值检测中展现出广阔的应用前景,但该方法还无法直接应用于时间序列中的异常值检测。该文结合相空间重构理论[6]和一类分类方法提出一种时间序列中的异常值检测方法。

2 一类分类

异常值检测实际上可视为一类特殊的分类问题[5]。给定一个正类样本点集$\{x_i, i=1,\cdots,l\}, x_i\in R^d$,设法找到一个以$\alpha$为中心,以$R$为半径的能够包含所有样本点的最小球体。如果直接进行优化处理,所得到的优化区域就是一个超球体。为了使优化区域更紧致,这里采用支持向量机中的核映射思想。首先用一个非线性映射ϕ将样本点映射到高维特征空间,然后在特征空间中求包含所有样本点的最小超球体,这样能获得原空间中更紧致的优化区域。为了允许一定的误差存在,通常引入松弛变量ξ_i,同时用核函数来代替将高维空间中的内积运算,即找一个核函数$K(x,y)$,使得$K(x,y)=<\phi(x),\phi(y)>$。于是优化问题为:

$$\min\ F(R,\alpha,\xi_i)=R^2+C\sum_{i=1}^{l}\xi_i \tag{1}$$

约束为:

$$(\phi(x_i)-\alpha)(\phi(x_i)-\alpha)^T\le R^2+\xi_i, i=1,\cdots,l$$

$$\xi_i\ge 0, i=1,\cdots,l$$

该优化问题的对偶形式为:

$$\max\sum_{i=1}^{l}\alpha_iK(x_i,x_i)-\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_jK(x_i,x_i) \tag{2}$$

约束为:

$$\sum_{i=1}^{l}\alpha_i=1$$

$$0\le\alpha_i\le C, i=1,\cdots,l$$

求解该优化问题可以得到α的值,通常大部分α_i将为零,不为零的α_i所对应的样本称为支持对象。根据KKT条件,对应于$0<\alpha_i<C$的样本满足:

$$R^2-(K(x_i,x_i)-2\sum_{j=1}^{l}\alpha_jK(x_j,x_i)+\alpha^2) \tag{3}$$

基金项目:国家自然科学基金资助项目(编号:60075014);广东省自然科学基金资助(编号:021349)

作者简介:孙德山(1970-),男,博士生,主要研究方向:统计学习理论、时间序列分析。吴今培(1937-),男,教授,博士生导师,主要研究方向:人工智能及电子技术。肖健华(1970-),男,博士,从事人工智能及故障诊断等研究。

其中，$\alpha=\sum_{j=1}^{l}\alpha_j\phi(x_j)$。因此，用任意一个满足条件的支持对象可求出 R 值。对于新样本 z，求下面的值：

$$f(z)=(\phi(z)-a)(\phi(z)-a)^T=K(z,z)-2\sum_{i=1}^{l}\alpha_iK(z,x_i)+\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_jK(x_i,x_j) \quad (4)$$

若 $f(z)\leq R^2$，则 z 为正常点，否则 z 为异常点。

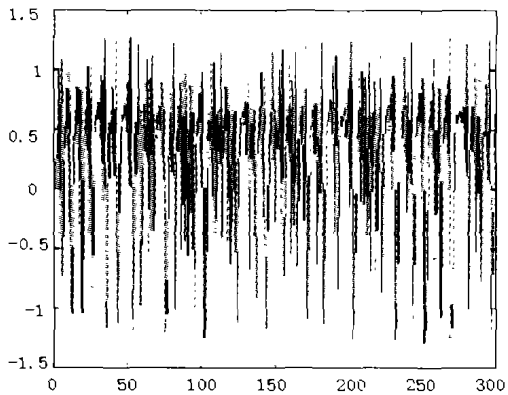
3 时间序列中的异常值检测

3.1 相空间重构

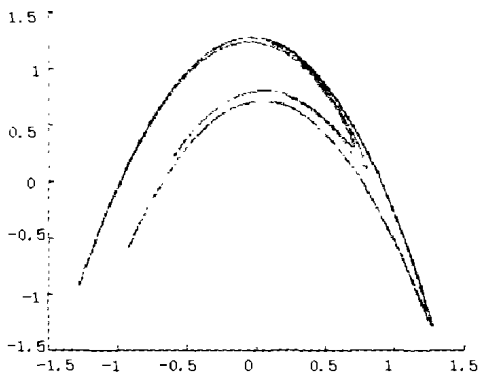
对于一个复杂的高维系统，能够得到的往往是一维的标量信息，即时间序列。根据相空间重构理论，对于时间序列 $x(t)$，$t=1,\cdots,N$，用一定的时间滞后 τ 和一定的嵌入维数 m，建立一个多维相空间 $X=\{X(t)|X(t)=[x(t),x(t+\tau),\cdots,x(t+(m-1)\tau)],t=1,\cdots,N-(m-1)\tau\}$，按照 Takens 的观点，只要嵌入维数 m 以及 τ 的选择恰当，重构相空间在嵌入空间的“轨线”就是微分同胚意义下的原系统的“动力学等价”。相空间重构可以使时间序列中的结构更清楚地表现出来。例如，Henon 混沌时间序列由下面的表达式式产生：

$$Y_t=1-1.4Y_{t-1}^2+0.3Y_{t-2},t\in Z \quad (5)$$

它的一维序列和二维相空间如图 1(a)和 1(b)，该序列的确定结构在相空间中清晰可见。



(a)一维轨迹

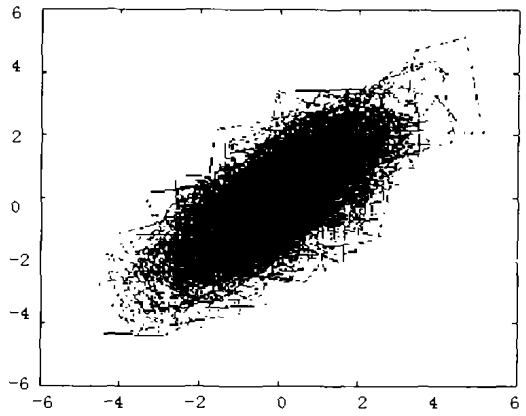


(b)相空间

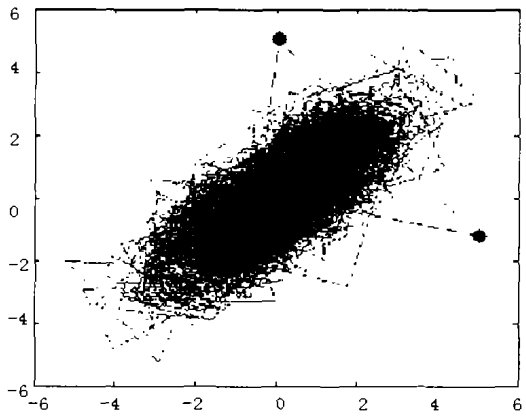
图 1 Henon 的一维轨迹和相空间

3.2 检测方法

既然相空间重构能够使时间序列中的结构显现出来，那么，同样也能够使异常值显现出来。也就是说，如果时间序列中含有异常值，那么经过相空间重构后，含有异常值的相点必然偏离正常区域。图 2 为一个 AR(1)过程的相空间，‘·’表示含有异常值的相点。时间序列中的异常值检测正是在此基础上进行的。



(a)不含异常值相空间



(b)含异常值相空间

图 2 AR(1)过程相空间

将时间序列转化成相空间中的点后，对相空间实行一类分类。根据任意一个支持对象，可以求出最小半径 R。这里要重点说明的是，优化方程中的松弛变量将起非常重要的作用，因为它的存在将允许一些点处在超球体外面，这也正是进行异常值检测的关键。超球体外面的点的个数由参数 C 决定，当它的值很大时，所有的样本点都将处于超球体内部。当参数 C 逐渐减小时，必将有样本点位于超球体外面，C 值越小，超球体外面的点越多，如图 3 所示，图(a)表示所有样本点都处于超球体内部，图(b)表示，当参数较小时，一些样本点跑到了超球体外面。

由于含异常值的相点通常远离正常轨道，所以只要选择较小的 C 值，该相点必位于超球体外面。当然，可能还有其它一些点也位于超球体外面。根据 KKT 条件，位于超球体外面的点所对应的 α 值一定满足 $\alpha=C$。

经以上分析，可给出时间序列中的异常值检测步骤如下：

(1)将时间序列重构到相空间。

(2)选择初始参数，包括核函数的类型和其中的参数，以及参数 C。

(3)采用一类分类方法求出 α 的值以及最小半径 R 的值。

(4)找出所有满足条件 $\alpha=C$ 的点，并将之代入(4)式得到这些点到超球体中心的距离 $f(z)$。

(5)计算 $f(z)/R$，如果某些值明显大于其它值，则可以判定

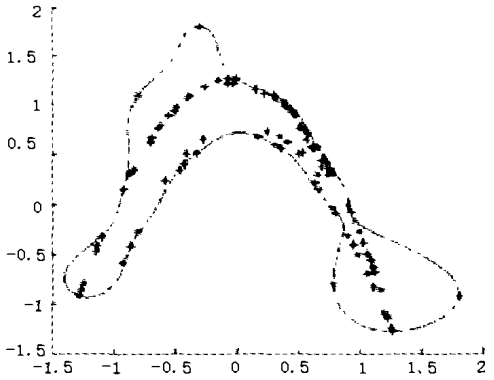
表 1 Henon 数据计算结果

序号	22	23	39	43	64	70	71	84	90	91	92	93
实值	0.9556	0.8366	0.7505	0.7708	0.7513	0.7510	0.7541	0.7524	0.7508	0.7538	1.0221	0.7785
比值	1.2734	1.1148	1.0000	1.0272	1.0012	1.0007	1.0048	1.0027	1.0006	1.0044	1.3621	1.0374

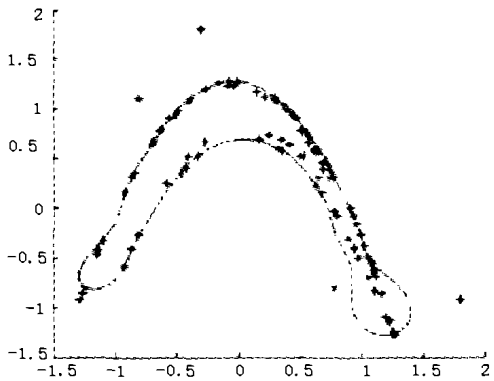
对应于该值的相点含异常值。这一步可根据具体情况事先选择一个标准值，大于该标准值的相点被判为含异常值。

(6)进一步判断异常值在时间序列中的位置。

由于时间序列中的一个异常值嵌入到相空间后将变成 m 个异常点，所以在相空间中检测出异常值后，还须进一步判断异常值在时间序列中的位置。只要时间序列中的异常值比较分散，这一步比较容易判断。但当异常值在时间序列中出现的间隔较近时，虽然能够找到相空间中的异常值，但往往不能清楚地判断异常值在时间序列中的位置。这时可以通过采用不同的延迟重新构造相空间，再进行一类分类运算，两种结果进行比较通常能检测出异常值在时间序列中的位置。



(a)参数 $C=\frac{1}{2}$ 时的封闭曲线



(b)参数 $C=\frac{1}{20}$ 时的封闭曲线

图 3 一类分类获得的原空间封闭曲线

4 仿真实例

(1)Henon 混沌时间序列

表达式如(5)式，取初值为[0,0]，产生 100 个一维数据作为研究对象，并且在其中加入两个异常值，第 23 个数据点用-1.2 代替，第 93 个数据点用 1.5 代替。在一维空间中，不易检测出这两个异常值。将该序列嵌入到二维空间中，采用该文的方法来进行检测。取参数 $C=\frac{1}{20}$，核函数取为：

$$K(x_i,x_j)=\exp\left(-\frac{\|x_i-x_j\|^2}{2\sigma^2}\right),\sigma^2=0.45$$

Matlab6.1 编程实现，所得结果为超球体半径 $R=0.7504$，其它结果如表 1 所示。其中，序号是指满足 $\alpha=C$ 条件的相点所处的位置；实值是对应相点到超球体中心的距离 $f(z)$；比值是 $\frac{f(z)}{R}$。

从表 1 中可以看到，第 22、23 个相点的比值，还有第 92、93 个相点的比值明显高于其它点的比值，从而可以判断时间序列中的第 23 个点和第 93 个点是异常值。

(2)NAR(2)过程

该过程由下面的式子产生：

$$y(t)=(0.8-0.5\exp(-y^2(t-1)))y(t-1)-(0.3+0.9\exp(-y^2(t-1)))y(t-2)+0.1\sin(y(t-1)\pi)+\varepsilon(t),\varepsilon(t)\sim N(0,0.09) \quad (6)$$

取初值为[0,0]，产生 100 个一维数据作为研究对象。第 50 个数据点用 2 代替，它可以作为一个异常值。将该序列嵌入到二维空间中，然后采用一类分类进行检测。核函数类型同前面，取参数 $\sigma^2=0.25$，$C=\frac{1}{25}$。计算结果为 $R=0.8255$，其它结果如表 2。

表 2 NAR(2)数据计算结果

序号	21	22	35	44	46	47	49
实值	0.8530	0.8443	0.8311	0.8334	0.8284	0.8375	1.0044
比值	1.0333	1.0227	1.0067	1.0095	1.0034	1.0145	1.2167
序号	50	51	70	72	75	76	79
实值	1.0641	0.8283	0.8260	0.8258	0.8274	0.8286	0.8369
比值	1.2890	1.0033	1.0006	1.0003	1.0023	1.0037	1.0138

从表 2 中可以看到第 49、50 个相点的比值明显大于其它值，因此可以判断时间序列中的第 50 个点为异常值。

5 结论

论文结合相空间重构理论和一类分类方法提出了一种时间序列中的异常值检测方法。由于采用了核函数，因此，所给方法在处理非线性时间序列异常值检测问题时具有传统方法不可比拟的优越性。为了便于说明，论文只以二维嵌入相空间为例来说明，所给方法对于高维嵌入相空间同样适用。

(收稿日期：2003 年 8 月)

参考文献

1.Tsay R S.Time series model specification in the presence of outliers [J].Journal of the American Statistical Association,1986;81:132~141

2.Chen C,L-M Liu.Joint estimation of model parameters and outlier effects in time series[J].Journal of the American Statistical Association,1993;88:284~297

3.Smola A J,Scholkopf B.A tutorial on support vector regression[R]. NeuroCOLT TRNCTR98030,Royal Holloway College University of London,UK,1998

4.Burges C J C.A Tutorial on Support Vector Machines for Pattern Recognition[R].Knowledge Discovery and Data Mining,1998;2(2)

5.Tax D.One-class classification[D].PhD thesis.Delft University of Technology,http://www.ph.tn.tudelft.nl/~davidt/thesis.pdf,2001

6.Cees Diks.Nonlinear Time Series Analysis[M].World Scientific Publishing Co.Pte.Ltd,1999