基于预测的序列异常数据挖掘

杨 虎 王会琦 程代杰
(重庆大学 重庆400044)

摘 要 本文中,我们分析了给定的股票时间序列。首先,基于稳定化时间序列,我们通过模型识别和估计,给出了一个初始模型,用以预测股票价格。然后,我们可通过股票检测来发现股票时间序列的异常点。最后,通过修正这些异常点,便可完善模型,逐步提高股票的预测精度。

关键词 异常挖掘,序列数据挖掘,预测,AR 模型

Series Outlier Data Mining Based on Forecastment

YANG Hu WANG Hui-Qi CHENG Dai-Ji
(Chongqing University,Chongqing 400044)

Abstract In this paper, we analyze given stock-time-series. Firstly, based on steadied-time-series, we get an initial model by model identification and estimation, and use the model to forecast the price of stock. Then we can use the Score test to find the outliers in stock time series. Finally, by revising the outliers we can improve the model and increase the accuracy of forecastment step by step.

Keywords Outlier mining, Series date mining, Forecastment, AR model

1 引言

数据挖掘[1](Data Mining,简记为 DM)是近年来随着数据库和人工智能发展起来的一门新兴的数据库技术,由于其前沿性和多学科的综合性,引起学术界广泛的重视,其中异常挖掘(Outlier Mining)和序列数据挖掘(Series Data Mining)是比较困难的研究分支之一,尤其在金融数据的挖掘中显得特别困难[2,3],因为预测问题本身就是金融数据分析中非常困难的研究课题,很多时候不得不借助于计算工具进行实际模拟,这给挖掘算法的设计带来很大的困难。比如对于股票数据,面对大量的数据,如何寻找对投资者有利的信息?为了有良好的预期,本文尝试了一些新的挖掘方法。统计学家对于数据挖掘研究的重视和投入近年来显得很突出[4]。而且很多国内外的学者都把重点放到了时间序列异常点的挖掘上[5],对于证券研究,一般是通过找出股票日线序列中的异常点来构筑投资策略以提高投资的成功率,避免盲目性,然而,我们不应该忽视异常点的另一类作用,它也可以用于改进模型,提高预测精度。本文就是针对上述情况,从另外一个角度挖掘异常点所提供给我们的有用信息。

2 基于自回归 AR 模型的挖掘算法

2.1 引入 AR 模型的原因

时间序列的模型类型很多,但对于股票时间序列,我们可以只考虑 AR 模型,其原因主要有以下几点:

(1)在证券市场出现的股票时间序列,差分平稳化后一般都符合 AR 模型;

(2)ARMA(p,q)可以用 AR(p+q)很好地近似;

(3)对于 AR 模型容易进行参数估计,而对于其他模型要涉及非线性方程组的迭代算法,很难控制它的收敛性,而且初值的选取也很讲究,效率不高。对于涉及大量数据且以效率优先的股票市场,AR 模型相对于其他模型更实用。

2.2 AR 模型介绍

首先对于一般的 AR(p)模型:

$$X_t=\varphi_1X_{t-1}+\varphi_2X_{t-2}+\cdots+\varphi_pX_{t-p}+a_t$$

其中,$\{X_t\}$为零均值(即中心化处理后的)平稳序列,对于非平稳时间序列我们可以通过差分化为平稳序列;φ_i 表示 X_t 与 X_{t-i} 之间的依赖程度,a_t 为随机扰动(白噪声序列)。这个模型也就是说第二天的预测值是其前 p 天值的某种加权平均。

2.3 算法的设计过程

根据平稳序列的自相关和偏自相关函数形式及其统计特性,我们可以得出:对于零均值平稳时间序列要么拖尾,要么截尾,而且会迅速衰减。否则就说明它存在某种趋势或周期性。在 AR(p)模型中有:自相关函数拖尾;偏自相关函数截尾。这不仅为我们提供了平稳化序列判断的方法,还可以用来进行 AR 模型的识别[6]。下面以青岛海尔这支股票2002.6.17到2003.1.21每天的开盘价所构成的时间序列为例(图1)。

通过零均值化、二阶差分后得到了一个时间序列(图2),算法如下(Matlab 6.1):

```
%求自相关系数
    sum1=1/148*x0*x0';//x0是原始数据经过二阶差分和零
                        均值化后得到的行向量
    for k=1:147
        sum3=0;
        for m=1:147-k
            sum3=sum3+1/148*x0(m)*x0(m+k);// 求样本
                                自协方差并存储在 sum3中
        end
        r0(k)=sum2(k)/sum1;//求样本自相关函数并存储在 r0
                            中
    end
%求偏自相关系数
    for k=2:147
        for j=1:k-1
            sum4=sum4+f0(k-1,j)*r0(k-j);
            sum5=sum5+f0(k-1,j)*r0(j);
        end
        f0(k,k)=(r0(k)-sum4)/(1-sum5);
        for j=1:k-1
```

杨 虎 教授,研究领域为数据挖掘技术与算法、统计诊断、金融统计。**王会琦** 信息与计算科学专业本科生。**程代杰** 博导,研究领域为计算机体系结构、并行处理、信息网络与数据挖掘。

```
    f0(k,j)=f0(k-1,j)-f0(k,k)*f0(k-1,k-j);
    end //利用递推方法求f0,并在此前赋f0(1,1)=r0(1)
end
```

```
pr0(i)=f0(i,i);//取f0对角线元素作为偏自相关系数并存储于pr0
```

图1

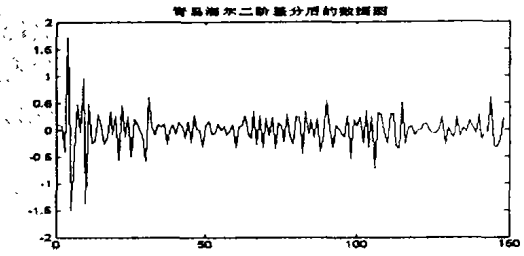

图2

它的自相关和偏自相关函数见图3、图4。图中可以看出经过二阶差分后，序列已经平稳，而且自相关函数拖尾，偏自相关函数截尾[7]。我们基本可以判断序列适应于 AR 模型。

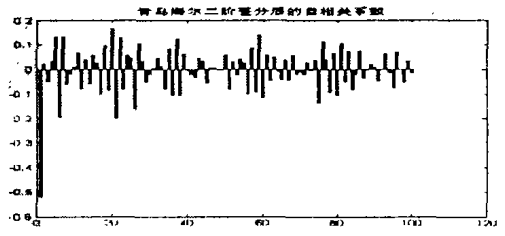

图3

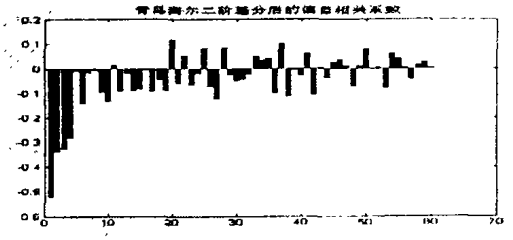

图4

在识别出 AR 模型后，我们的问题就转化为这个序列到底是几阶？也就是模型的定阶问题。我们可以综合考虑 AIC 和 BIC 定阶准则[1]，得到 AIC、BIC 函数如图5。由此可以确定模型的阶数是4。同时，此过程中还进行了模型的参数估计（可利用 Yule-Wolker 方程得到），AR(4)模型的各个参数值如下表1。

表1

模型	φ_1	φ_2	φ_3	φ_4	残差平方和
AR(3)	−0.9020	−0.7288	−0.5552	−0.2820	6.3818

模型定阶和参数估计的算法如下：

```
%定Ar(n)模型阶数并同时进行参数估计
    ruo=zeros(n,1);
    for i=1:n
        ruo0(i,i)=1;
        for j=i+1:n
            ruo0(i,j)=r0(j-i);
            ruo0(j,i)=ruo0(i,j);
        end
    end // 构造Yule-Wolker方程的系数矩阵
    f(1:n,n)=inv(ruo0(1:n))*r0(1:n); // 解方程求出AR(n)所对应的参数并存储在f中
    FPE(n)=(1+n/148)/(1-n/148)*(sum1-f(1:n,n)'*r0(1:n)*sum1);// 求FPE(n)函数
    cancha(k)=sum1-f(1:k,k)'*r0(1:k)*sum1; // 求不同阶数所对应的残差
    AIC(n)=log(cancha(n))+2*n/148; // 求AIC(n)函数
    BIC(n)=log(cancha(n))+n/148*log(148); // 求BIC(n)函数
```

至此我们已经建立了一个 AR(4)模型，下面我们分析挖掘效果。因为序列阶数为4，即 x(t)受到前四个点 x(t−1)、x(t−2)、x(t−3)、x(t−4)的影响而表现出一定的相关性；并由于是自回归模型，x(t)也将与 t 时刻以前进入的扰动 a(t−j)(j>0)独立。作散点图可直观地看出。

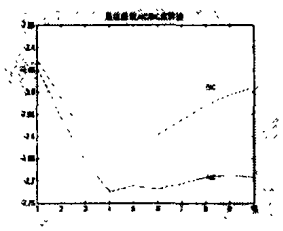

图5

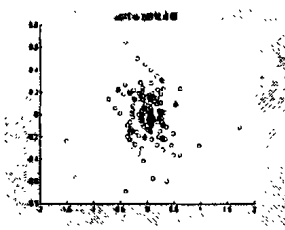

图6

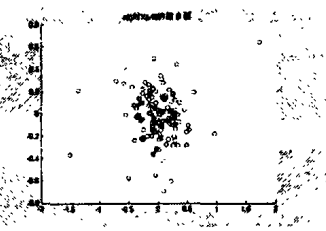

图7

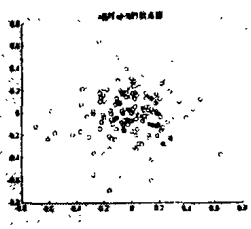

图8

从图中我们已经可以看出 x(t)与 x(t−3)还存在一定的负相关趋势（图6），而与 x(t−5)、a(t−1)看不出有相关趋势（图7、图8），所以大致判断 AR(4)模型是适应的，但为了更具说服力，我们再采用 χ^2 检验法[8]，得到

$$Q=N\times\sum_{k=1}^{L(N)}\rho_i^2(N,a_t)\approx 17.737<\chi^2_{0.01}(8)=20.090$$ （取显著性水平 $\alpha=0.01$）

其中：$\rho_k(N,a_t)$ 为 a_t 的自相关函数，$L(N)=[\frac{N}{10}]$ 或 $\sqrt{N}$，因而模型是显著的。

可用这个模型进行股价走势预测[7]，算法如下：

```
%一步预测
    w(t)=f(1,3)*x0(t-1)+f(2,3)*x0(t-2)+f(3,3)*x0(t-3);// 平稳序列的一步预测
    x'(t)=w(t-2)+2*x(t-1)-x(t-2);
        // 化平稳前原始序列的一步预测(其中x存储原始序列)
%二步预测
    ww(t+1)=f(1,3)*w(t)+f(2,3)*x0(t-1)+f(3,3)*x0(t-2);// 平稳序列的二步预测
    xx2(t)=ww(t-2)+2*x2(t-1)-x(t-2);
        // 化平稳前原始序列的二步预测(其中x存储原始序列)
```

以“青岛海尔”2003.1.21以后的数据作为预测效果的评价准则，可以看出预测值和真实值之间的比较（图9）。

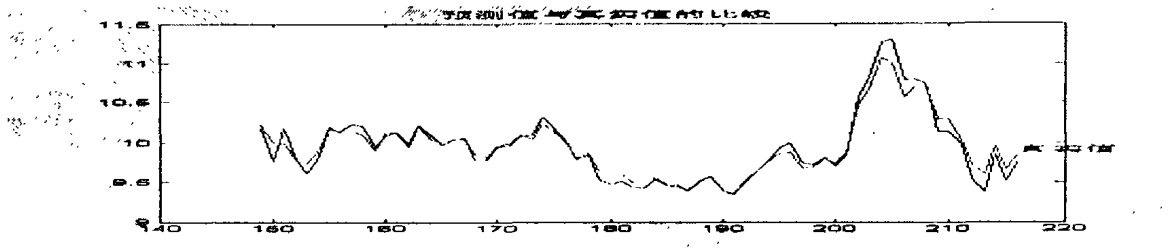

图9

图中的总体走势基本上一致，误差不是很大。要评价预测效果的好坏，需要确定的仅仅是根据模型预测出的值和真实值之间的涨跌趋势是否一致。

(1)如果真实值和预测值之间的趋势是同时涨或同时降，则认为是成功的。因为同为涨势时，可以盈利；同为跌势时，可以规避风险。

(2)如果真实值无涨势和跌势(即实际股价不变)，不管预测值如何变动，都认为是无效的，因为，在这种情况下，由于股价不变，也就没有进行任何操作的必要。

(3)其余情况认为是失败的。

(4)同时，我们定义平均收益率为：$\bar{p}=\frac{1}{m}\sum_{i=1}^{m}\frac{\Delta P_i}{P_i}$

其中，ΔP_i 为某一时间间隔内实际股票价格的涨幅(可正可负)；P_i 为这一时期内实际股票的平均价格；m 为预测可以赢利的次数。

通过以上定义，我们就可以判断模型预测效果的好坏，以青岛海尔2003.1.21以后的68个数据为例，效果见表2。

表2

效果	一步预测		二步预测	
	次数	百分比	次数	百分比
成功	58	85.30%	55	82.09%
失效	5	7.35%	0	0%
失败	5	7.35%	12	17.91%
平均收益率	1.541%		2.406%	

误差	一步预测		二步预测	
	小于1%	1%-3%	小于1%	1%-3%
次数	51	17	15	26
百分比	75.00%	25.00%	22.39%	38.81%
平均误差	0.675%		1.278%	
备注：平均误差大小可反映预测风险大小				

可见投资者如果根据这个模型的一步预测值来进行投资，失败的几率不会超过7.35%；二步预测不会超过17.91%，所达到的平均收益率也基本可以接受。

上面是基于预测所设计的序列挖掘算法，但此时我们的预测任务并没有真正完成，在这个基础上，我们还可以通过异常挖掘设计出更加精确可靠的算法来。

3 算法的改进和精度的提高

由于序列中异常数据的存在，往往会影响算法的有效性，对于证券历史数据，序列中也会因为各种原因存在很多异常点，所以如果根据含有异常数据的序列设计模型和相应的挖掘算法，预测的精度就会受到一定程度的影响。为了进一步提高预测精度，有必要先进行异常挖掘，找出历史数据中的异常数据，然后对异常数据进行适当的处理，以消除这种影响，重新设计序列挖掘算法，使其达到更精确的预测效果。

通常时间序列中的异常数据被分为加性异常(AO)和创新异常(IO)，往往被人们所重视的是 IO 异常，认为 IO 作用时间长，具有投资价值。然而却忽视了 AO 异常，其实它和 IO 异常一样，在改进模型、提高预测精度方面都起着重要作用，这是因为 AO 异常点是非本质的，剔除它之后完全不会影响时间序列内在的相关结构。下面我们的任务就是找出时间序列中的 AO、IO 异常点[8]。通过引入 Score 检验统计量：

$$SC_k=\frac{1}{\sigma^2}a_k^2+(\sum_{i=1}^{p}\varphi_i a_{k+i})^2/(\sigma^2\sum_{i=1}^{p}\varphi_i^2)$$

且 $SC_k\sim\chi^2(2)$，取 $\alpha=0.05$，通过下面的算法

```
%计算 p+1到 n-p 时刻异常点的检验统计量(p<k<n-p+1)
    SC1(k)=(a(k)-f(1:p,4)*a(1+k:p+k))^2/(var(a)*(1+
           f(1:p,4)'*f(1:p,4)));
           // AO 统计量 SC2(k)=a(k)^2/var(a);// IO 统计量
    SC(k)=a(k)^2/var(a)+(f(1:p,4)*a(1+k:p+k))^2/(var
          (a)*sum4);
%判断异常点位置
    L=max(abs(SC(k)) // 记取得最大值时的位置为 T
%判断异常点的类型
    u=var(a)*(SC1(T)-SC2(T));// 如果有 L 大于某一给定常
                             数时
    if u>=0
      input('AO 异常点!!');
    else
      input('IO 异常点!!');
    end
```

可以挖掘出“青岛海尔”股票在时间段2002.6.17-2003.1.21存在5个 IO 异常数据见表3。

表3

异常点位置	真实时间	统计量 SC	漂移量 Δ	类型
11	2002.07.01	14.101	0.643	IO
8	2002.06.26	11.106	-0.693	IO
108	2002.11.21	8.569	-0.606	IO
32	2002.07.30	8.349	-0.580	IO
15	2002.07.05	7.426	-0.555	IO

对挖掘出来的异常数据可以利用预测值代替真实值进行修正：(1)对于 AO 异常直接用预测值代替；(2)对于 IO 异常，除代替异常数据外，还要对后继的某些数据作适当修正，因为序列具有动态性。在修正过程中由于只涉及序列中极少数点，因此一般不会改变序列的平稳性。我们依然可以利用同样的方法对修正后的序列重新进行模型识别、定阶和参数估计。

通过以上的处理可以消去异常数据对挖掘算法的影响。首先我们获得了改进的模型参数(见表4)。

表4

模型	φ_1	φ_2	φ_3	残差平方和
AR(3)	-0.9691	-0.5271	-0.2391	6.2572447

利用这个模型建立的算法进行预测，我们得到了2003.1.21以后股价的真实数据与预测数据的比较图如图10，预测效果可见表5。

(下转第146页)

47 Landay J A, Myers B A. Sketching Interfaces: Toward More Human Interface Design. IEEE Computer, 2001, 34(3): 56～64
48 Hearst M A, Gross M D, Landay J A, Stahovich T F. Sketching Intelligent Systems. IEEE Intelligent Systems, 1998, 13(3): 10～19
49 Lin J, Newman M, Hong J, Landay J A. DENIM: Finding a Tighter Fit between Tools and Practice for Web Site Design. In: Proc. of ACM CHI'2000, 2(1): 510～517
50 Bederson B B, Meyer J, Good L. Jazz: an Extensible Zoomable User Interface Graphics Toolkit in Java. In: Proc. of UIST'2000, San Diego, CA, 2000. 171～180
51 Pedersen E R, McCall K, Moran T P, Halasz F G. Tivoli: An Electronic Whiteboard for Informal Workgroup Meetings. In: Proc. of the ACM INTERCHI'93 Conf on Human in Computing Systems, 1993. 391～398
52 Moran T P, Melle W V, Chiu P. Spatial Interpretation of Domain Objects Integrated into a Freeform Electronic Whiteboard. In: Proc. of UIST'98, San Francisco, CA, 1998. 175～184
53 Alvarado C, Oltmans M, Davis R. A framework for multi-domain sketch recognition. Sketch Understanding, Papers from the 2002 AAAI Spring Symposium, 1～8
54 Hammond T, et al. Multi-domain sketch recognition. MIT Student Oxygen Workshop, 2002
55 Veselova O. Perceptually based learning of shape descriptions from one example. MIT Student Oxygen Workshop, 2002
56 Sezgin M. Generating domain specific sketch recognizers from object descriptions. In: MIT Student Oxygen Workshop, 2002
57 Anthony J. Adaptive interfaces and agents. In: Julie A. Jacko, Andrew Sears, eds. Handbook of Human-Computer Interaction in Interactive Systems. Erlbaum, Mahwah, NJ, 2002. In press
58 Geoff W, Michael J P, Daniel B. Machine learning for user modeling. User Modeling and User-Adapted Interaction, 2001, 11: 19～29
59 Bruce K. Lifestyle Finder: Intelligent user profiling using large-scale demographic data. AI Magazine, 1997, 18(2): 37～45
60 Anthony J. Numerical uncertainty management in user and student modeling: An overview of systems and issues. User Modeling and User-Adapted interaction, 1996, 5: 193～251
61 Alfred K, Wolfgang P. The user modeling shell system BGP-MS. User Modeling and User Adapted Interaction, 1995, 4: 59～106
62 Anthony J. Barbara Grobmann-Hutter, Leonie March, Ralf Rummer, Thorsten Bohnenberger, and Frank wittig. When actions have consequences: Empirically based decision making for intelligent user interfaces. Knowledge-Based Systems, 2001, 1: 75～92

(上接第 119 页)

表 5

效果	一步预测		二步预测	
	次数	百分比	次数	百分比
成功	57	83.83%	60	89.55%
失效	5	7.35%	0	0%
失败	6	8.82%	7	10.45%
平均收益率	1.578%		2.815%	

误差	一步预测		二步预测	
	小于 1%	1%-3%	小于 1%	1%-3%
次数	53	14	15	25
百分比	77.94%	20.59%	22.39%	37.31%
平均误差	0.639%		1.385%	
备注：平均误差大小于可反映预测风险大小				

可见通过异常挖掘后再对序列挖掘算法加以改进，预测效果明显好转，尤其对于多步预测。实际中多步预测的效果往往更受投资决策者青睐。

我们用同样的挖掘算法测试其它的股票(部分结果可见表 6)，结果是一致的。

通过异常挖掘可以有效地改进序列挖掘算法以提高预测的精度，尤其对于多步预测。同时平均收益率也会有相应的增加。由此可认为这种改进设计是切实可行的，而且效果也不错。当然，我们的挖掘算法改进还有潜力，可以利用第二次建立的模型再次进行异常挖掘，如此循环逐步提高预测精度。当然这种提高是有限的，这是由于股票走势本身的不可预测性和受各种因素的干扰所致，通过有限次的循环，一般可以最大限度地剔除历史数据中的外部干扰因素对预测精度的影响。

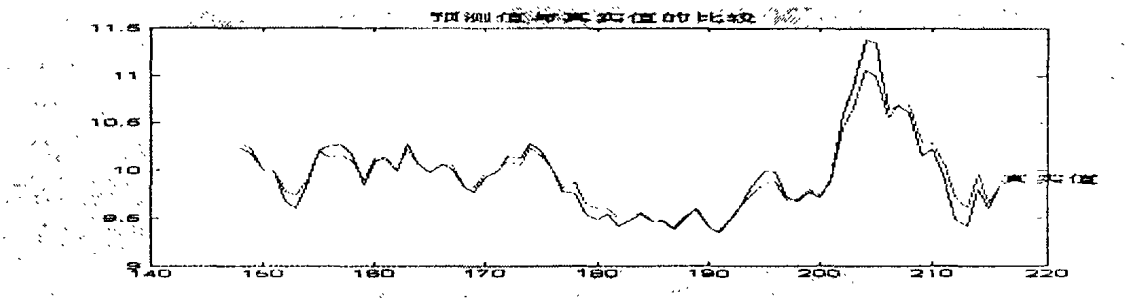

图 10

表 6

股票名称	初步模型				改进模型			
	一步预测		二步预测		一步预测		二步预测	
	成功率	收益率	成功率	收益率	成功率	收益率	成功率	收益率
青岛啤酒	83.82%	2.141%	89.55%	2.815%	89.71%	2.094%	92.54%	3.244%
深圳泰丰	88.24%	2.485%	89.55%	4.439%	88.06%	2.650%	92.54%	4.743%
涌鼎光辉	92.65%	1.759%	92.54%	3.562%	95.59%	2.020%	94.03%	3.806%

参 考 文 献

1 Han J, Kamber M. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, Inc., 2001
2 John G H, Miller P. Building long/short portfolios using rule induction. In: Computational Intelligence for Financial Engineering, Piscataway NJ: IEEE Press, 1996
3 John G H. Stock selection using rule Induction. IEEE Expert, 1996, 52～58
4 Glymour C, Madigan D, Pregibon D, Smyth P. Statistical themes and lessons for data mining. Data mining and knowledge discovery, 1997, 1: 11～28
5 LiuL-M, et al. Data mining on time series: an illustration using fast-food restaurant franchise data. Computational Statistics & Data Analysis, 2001, 37: 455～476
6 王振龙. 时间序列分析. 中国统计出版社, 2000
7 易丹辉. 统计预测—方法与应用. 中国统计出版社, 2001
8 韦博成, 等. 统计诊断引论. 东南大学出版社