时间序列分析方法及ARMA,GARCH两种常用模型

武 伟[1],刘希玉[2],杨 怡[2],王 努[1]
(1.山东师范大学 信息科学与工程学院,山东 济南 250014;
2.山东师范大学 管理与经济学院,山东 济南 250014)

摘 要:证券市场具有数据单一性(大量不需要经过特殊处理的数据)、分析手段多样性和隐蔽性的特点,且与其飞速发展不相称的是证券分析技术进展的缓慢。股市系统中时间序列的预测问题具有重要的理论及实际意义,时间序列的获取是通过对数据库中数据进行分类汇总分析而获得。获取时间序列数据以后可以对它进行预测分析,从而较准确地预见系统的演进。文中介绍了时间序列的基本知识,同时比较了ARMA和GARCH两种常用模型,得出对于中国股市,GARCH模型性能优于ARCH模型。

关键词:时间序列分析法;自回归移动平均模型;条件异方差模型
中图分类号:TP30 **文献标识码**:A **文章编号**:1673-629X(2010)01-0247-03

Analysis Method of Time Array and Two Models of ARMA and GARCH

WU Wei[1], LIU Xi-yu[2], YANG Yi[2], WANG Nu[1]
(1. Department of Information Science and Engineering, Shandong Normal University, Jinan 250014, China;
2. Department of Management and Economics, Shandong Normal University, Jinan 250014, China)

Abstract: The securities market has monotony of data (a large number of date does not need processed with special-method), characteristic of analyzing the means variety and disguise, and the technology of securities analysis has been developing so slowly that it can not be adapted to the high speed at which the securities market is going forward. The prediction question of the time array in system on the securities market has important theoretical and actual significance. People can analyze and predict the time array data after it has been obtained, and predict the systematic gradual progress more accurately. Provide the basic knowledge of time array, and make the comparison of the ARMA model and GARCH model, by which the conclusion can be easily found that the GARCH model is a little better than the ARMA model if it is utilized in the Chinese stock market.

Key words: analysis method of time array; ARMA; GARCH

0 引 言

由于股市内部规律非常复杂,变化周期无序,而我国资本市场个人投资者的比例高达99%,投资者个人心理状态不同,同时经济、政治等因素对其影响较大,使股价走势变化莫测,难以把握。特别是目前在受到金融危机影响的情况下,对股市进行合理分析和预测,对于指导投资者合理投资,维护证券交易市场稳定进而促进经济发展有重大意义。

目前常用的预测方法有证券投资分析法、神经网络预测方法和时间序列分析方法三种[1]。

证券投资分析方法是分析和预测股价变动方向和趋势的方法,可分为:技术分析法、基本分析法和组合分析法三大类。

神经网络是一种按照人脑的组织和活动原理而构造的一种数据驱动型非线性模型,它是由神经元结构模型、网络连接模型、网络学习算法等几个要素组成。神经网络预测方法主要包括前馈神经网络预测方法(FNN)、时间延迟神经网络预测方法(TDNN)和自回归神经网络预测方法(RNN)。

收稿日期:2009-04-28;**修回日期**:2009-07-19
基金项目:国家自然科学基金重大项目(60873058,60743010);山东省自然科学基金重大项目(Z2007G03)
作者简介:武 伟(1986-),女,硕士研究生,研究方向为流数据挖掘与人工智能;刘希玉,教授,博士生导师,研究方向为数据挖掘与人工智能。

文中重点介绍和比较时间序列分析方法的 ARMA 和 GARCH 两种常用模型。

1 时间序列预测方法

时间序列预测方法的基本思想是:通过时间序列的历史数据揭示现象随时间变化的规律,将这种规律延伸到未来,从而对该现象的未来作出预测[1]。

1.1 基本概念

时间序列就是一个变量在一定时间段内不同时间点上观测值的集合,如 $Y:\{y_1,y_2,\cdots,y_n\}$,这些观测值是按时间顺序排列的,时间点之间的间隔是相等的[2,3]。时间序列获取以后可以对它进行预测分析,预测方法可以从定性分析法和定量分析法两方面考虑。

长期趋势 T(Long term trend)、循环分量 C(Cyc-lical component)、季节分量 S(Seasonal component)、不规则分量 I(Irregular component) 是时间序列的四种成分。

长期趋势 T 表明在很长一段时间内总的走向趋势,也就是说是描述序列中长期运动趋势;循环分量 C 描述序列中不同幅度的扩张与收缩,趋势曲线所表现出来的是一种长期震荡,直线和曲线的振荡并不是周期的,这个循环并不遵循基于相等时间的规律,而是时间间隔不同的循环变动;季节分量 S 是有关时序数据在连续年份的各个相应月中所表现出相同或几乎相同的模式,描述的是序列中一定周期的重复变动,周期常为一年、一季、一周等;不规则分量 I 是由于一些突发的偶然事件而产生的,描述的是随机因素引起的变动,常带有偶然性由于各种因素引起变化相互抑制抵消,变动幅度常较小。时序变量 Y 可以是 $Y = T \times C \times S \times I$,也可以是 $Y = T + C + S + I$[4]。

1.2 时间序列分析的目的

时间序列分析的三个目的是预测(forecasting)、制模(modeling)、特征提取(characterization)[5]。预测的目的是较准确地预见系统的演进;制模的目的是给出能抓住系统的长期行为特征的描述;特征提取的目的是在没有先验知识的条件下确定序列的基本属性。时间序列的三个目的实际上是相互渗透、相互依存的,很难真正地把三者割裂开来。

当前对时间序列的处理主要集中在两个问题上,一是相似序列的搜索,另一个是时间序列的知识发现。严格的说,相似序列的搜索只是时间序列知识发现的一个重要组成部分,而时间序列的知识发现是一个更高层次的问题。

1.3 时间序列相似性搜索

时间序列表示方法的不同会严重地影响其距离度量对各种变形、扭曲的敏感程度,并决定相似性搜索的有效性。目前,已提出了一些时间序列的表示方法。其中,频谱表示法适合于局部稳定的时间序列,例如直接使用傅里叶系数[6]或参数频谱模型。但这些表示方法并不适用于有短暂行为的不稳定序列,同时从数据挖掘与知识发现的角度来看这种表示方法不直观,不易被人们理解和表达。Keogh 等人提出的逐段线性化表示法把复杂的曲线分段表示为直线段,不仅高度压缩了数据,同时较直观地反映了时间序列的变化形态。在此基础之上,Betty 等人采用区间离散化方法,提出了时间序列的符号表示方法[7]。这一方法直观新颖且符合人们的思维方式,但在离散化过程中,将本来相邻的数据硬性分割为不同的概念,分别表示为不同的符号,也会引起相似性判别的失误。

在时间序列相似性搜索领域,相似性的问题可以被分为子序列匹配和整序列匹配两类,子序列匹配是在查询序列较小且模式序列较大的情况下应用。

一般有四种相似性查询:完全匹配(给定的查询形状与数据库中序列的形状完全相同)、位移无关匹配(从数据库中选取相似的形状而不管其在坐标系中的位置)、尺度无关的匹配(既不关心位置也不关心形状的尺度)及位移和尺度都无关的匹配。

相似匹配的

定义 1:给定一个门限值 $\varepsilon \geqslant 0$ 和一个时间序列的尺度 D,当满足 $D(S_1,S_2) \leqslant \varepsilon$,称序列 S_1,S_2 相似。

定义 2:两个序列 S_1,S_2 之间的欧几里德距离:$D_E(S_1,S_2) = (\Sigma(S_1[i] - S_2[i])^2)^{1/2}$。

2 模型的介绍及样本的选择

常用的时间序列分析法主要是自回归移动平均模型和条件异方差模型。

2.1 ARMA 模型

ARMA 模型是描述平稳随机序列的最常用的一种模型,有三种基本形式:自回归模型(AR:Auto - Regressive);移动平均模型(MA:Moving - Average);混合模型(ARMA:Auto - Regressive Moving - Average)。也可以这样认为:ARMA = AR + MA。

(1) AR(p) 模型。

如果时间序列 $\{y_t\}$ 满足 $y_t = \sum_{i=1}^{p} \alpha_i y_{t-i} + \varepsilon_t$,其中 $\{\varepsilon_t\}$ 是独立同分布的随机变量序列,且满足:$E(\varepsilon_t) = 0$,$\mathrm{Var}(\varepsilon_t) = \sigma_\varepsilon^2 > 0$,则称时间序列 $\{y_t\}$ 服从 p 阶自回归模型。其平稳条件为:滞后算子多项式 $\alpha(B) = 1 - \sum_{i=1}^{p} \alpha_i B^p$ 的根均在单位圆外,即 $\alpha(B) = 0$ 的根大于 1。

(2)MA(q) 模型。

如果时间序列$\{y_t\}$满足$y_t = \varepsilon_t - \sum_{i=1}^{q}\beta_i\varepsilon_{t-i}$,则称时间序列$\{y_t\}$服从 q 阶移动平均模型,且在任何条件下都平稳。

(3)ARMA(p,q) 模型。

如果时间序列$\{y_t\}$满足$y_t = \varepsilon_t + \sum_{i=1}^{p}\alpha_i y_{t-i} + \sum_{i=1}^{q}\beta_i\varepsilon_{t-i}$,则称时间序列$\{y_t\}$服从($p,q$) 阶自回归移动平均模型,或者记为:$\alpha(B)y_t = \beta(B)\varepsilon_t$。

显然,$q = 0$,ARMA(p,q) 模型即为 AR(p);$p = 0$,模型即为 MA(q)。

2.2 GARCH 模型

许多实际问题中随着时间 t 的变化,序列$\{y_t\}$的随机扰动项的条件方差也在变化,即序列具有变方差的特性。Engel 在 1982 年首先提出了 ARCH 模型对方差进行建模,来描述股票市场的波动聚类性和持续性。

ARCH(q) 模型的条件方差函数为[8]:

$\alpha_t^2 = \alpha_0 + \sum_{i=1}^{p}\alpha_i\varepsilon_{t-i}^2$ 或$\varepsilon_t^2 = \alpha_0 + \alpha(B)\varepsilon_{t-i} + v_t$,其中:$\alpha_0 > 0, \alpha_i \geqslant 0$。

ARCH 模型通过对过去 p 期非预期回报(ε_t)的平方的平方的移动平均来捕获回报序列的条件异方差。但是 ARCH(q) 模型在实际应用中为得到较好的拟合效果需要很大的阶数 q,这增大了待估参数的个数,还会引发诸如解释变量的多重共线性等其他问题。另外,对于大数 q,非限制估计通常会违背 q 为负数的限定条件。

1986 年 Bollerslev 将 ARCH 模型推广发展成 GARCH 模型,GARCH 模型考虑了异方差本身的自回归。GARCH 模型可以描述大多数金融报酬时间序列,所以在波动性研究中被广泛采用。

GARCH(p,q) 过程的条件方差函数为[8]:

$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q}\alpha_i\varepsilon_{t-i}^2 + \sum_{i=1}^{p}\beta_i\sigma_{t-i}^2 = \alpha_0 + \alpha(B)\varepsilon_t^2 + \beta(B)\sigma_t^2$

其中:$p \geqslant 0, q \geqslant 0, \alpha_0 > 0, \alpha_i \geqslant 0(i = 1,\cdots,q), \beta_i \geqslant 0(i = 1,\cdots,p)$

GARCH(p,q) 过程是平稳过程的充要条件是:

$\alpha(1) + \beta(1) < 1$

其中:$\alpha(1) = \sum_{i=1}^{q}\alpha_i; \beta(1) = \sum_{i=1}^{p}\beta_i$。

3 实证分析

文中采用上证综合指数 1098 个交易日的收盘价,数据(来自 http://q.stock.sohu.com/)的跨度从 1999 年 2 月 3 日至 2003 年 10 月 5 日。上证综合指数的日收益率采用对数差分进行计算(由于对数收益值很小,故均扩大 100 倍,以减少由计算精度引起的误差)[9]:$y_t = 100 \times (\ln(P_t) - \ln(P_{t-1})), (t = 1,\cdots,1100)$。

3.1 收益率序列的特性

通过表 1 的观察可以看出:出现正收益的机会要大于出现负收益的机会,具有明显的尖峰厚尾现象,也就是出现较大波动的可能性较大,且收益序列明显不为正态分布。

表 1　收益序列的基本统计情况

样本数	均值	方差	偏度
1100	0.017312	2.176365	0.6536
峰度	Jarque－bera 统计量		Approx Prob
6.162769	490.7556		0.0001

3.2 建立 ARMA 模型

利用 SAS 软件计算出上证综指收益率序列的自相关系数、偏自相关系数和逆自相关系数,结果如表 2 所示。从表中看出这些系数都接近于零,序列$\{y_t\}$平稳,因此考虑建立关于$\{y_t\}$的 ARMA 模型。

表 2　自相关系数、偏自相关系数和逆自相关系数

滞后系数	自相关系数	偏自相关系数	逆自相关系数
1	0.02015	0.02015	−0.04620
2	−0.03337	−0.03379	0.03117
3	0.05175	0.05321	−0.03591
4	0.03862	0.03537	−0.03597
5	−0.01604	−0.01414	0.02317
6	−0.02295	−0.02269	0.00099

具有最小的 AIC 或 SBC 的模型为最佳模型,经过计算决定建立 AR(2,3,12)模型,估计结果如表 3 所示。可以看到,残差自相关检验无法拒绝其为白噪声的原假设,故建立的模型恰当。

表 3　AR(2,3,12)的参数估计

参数	估计值	标准误差	T 统计量	残差的自相关检验		
				Lag	自由度	Prob
α_1	−0.01529	0.02969	−1.06	6	9	0.406
α_2	0.05408	0.01970	1.82	12	9	0.831
α_3	0.48301	0.02974	3.53	18	15	0.586

3.3 建立 GARCH 模型

经过反复筛选,对 AR 模型的残差建立了存在滞后 12 阶自回归的 GARCH(1,1)模型:

$y_t - 0.078264y_{t-12} = \varepsilon_t(-3.614)$

$h_t = 0.092197 + 0.200997\varepsilon_{t-1}^2(5.325)(9.742) + 0.775696\ h_{t-1}(36.154)$

(下转封三)

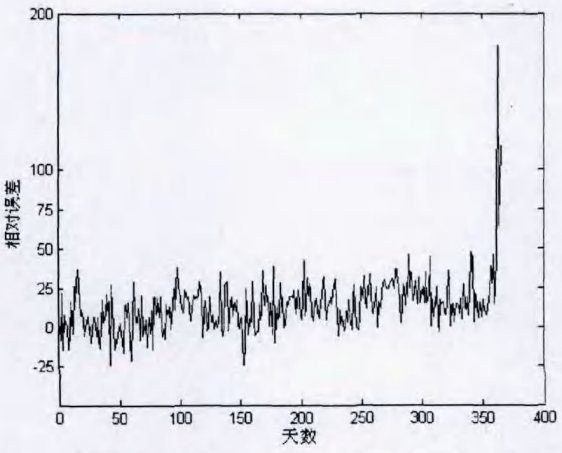

图 3　2002 年逐日 SO_2 监测值与预测值相对误差图

3　结束语

(1)由于支持向量机的优良特性,特别适合于那些模糊、随机、不确定性、样本数有限和非线性的复杂问题,支持向量回归模型能很好捕捉大气污染物浓度与其影响因子的非线性关系,通过实例证实支持向量机对大气污染的预测具有预测精度高和训练速度快的优点,并有广泛的应用前景。

(2)支持向量机预测大气污染模型的核函数以及惩罚参数 C 的选择将直接影响到支持向量机的学习效率和推广能力。文中核函数取 RBF 函数,通过对核函数核惩罚参数 C 的交叉测试, C 取 1000。

(3)SVM 方法属于参数预测方法,其预测精度在很大程度上依赖于预测模型的输入和输出参数的代表性。大气污染物浓度预测的可靠性和准确性,依赖对其各种影响因素的准确分析。

参考文献:

[1] 邓乃扬,田英杰.数据挖掘中的新方法——支持向量机[M].北京:科学出版社,2006:224-235.

[2] Nello C,John S T.支持向量机导论[M].李国正,王　猛,曾华军译.北京:电子工业出版社,2005:98-105.

[3] 白　鹏,张喜斌,张　斌,等.支持向量机理论及工程应用实例[M].西安:西安电子科技大学出版社,2008:41-55.

[4] 许建华,张学工,李衍达.支持向量机的新发展[J].控制与决策,2004,19(5):482-493.

[5] 谈建国,邵德民,黄家鑫.上海城市空气质量预报(日报)业务系统探讨[J].气象,2001,27(6):33-35.

[6] 金　龙.人工神经网络技术发展及在大气科学领域的应用[J].气象科技,2004,32(6):12-13.

[7] 张学工.关于统计学习理论与支持向量机[J].自动化学报,2000(1):32-42.

[8] Vapnik V N.统计学习理论的本质[M].张学工译.北京:清华大学出版社,2000:96-98.

(上接第 249 页)

$e_t \sim \mathrm{IID}(0,1)$

3.4　模型比较

利用上面得到的两个模型分别对上证综指的 10 月 13 日,14 日,15 日三天的收盘价作个短期预测,得到结果如表 4 和表 5 所示。

表 4　AR 模型预测上证综指

日期	03.10.13	03.10.14	03.10.15
真实值	1399.66	1388.17	1383.10
AR 预测值	1405.40	1391.50	1421.69
绝对误差(%)	0.41	0.24	2.79

表 5　GARCH 模型预测上证综指

日期	03.10.13	03.10.14	03.10.15
真实值	1399.66	1388.17	1383.10
GARCH 预测值	1405.19	1390.47	1421.13
绝对误差(%)	0.395	0.166	2.75

4　结束语

文中简要介绍了时间序列预测方法,分析了其在证券市场分析中的应用的可能和效果。并通过对上证综指的日收益率进行的实证研究,发现市场波动存在集群性,总体而言,GARCH 模型能比 ARCH 模型更好地反映这类特征。

参考文献:

[1] 谷　赫.时间序列的数据挖掘在证券预测分析中的应用研究[D].长春:吉林大学,2005.

[2] Mills T C.金融时间序列的经济计量学模型[M].北京:经济科学出版社,2002:217-355.

[3] 叶　峰.汇率的可预测性实证分析[J].管理工程学报,2001(3):41-45.

[4] 孔淑慧.流数据时序模式依赖挖掘在股市行情分析中的应用[D].北京:北京交通大学,2008.

[5] Weigend A S. Time Series Analysis and Prediction[D]. Colorado: Univerisity of Colorado, 1994.

[6] Faloutsos C, Rangsnathan M, Manoloppoulos Y. Fast stubsequence matching in time-series databases[C]//In: SIGMOD Proceedings of Annual Conference. Minneapolis: [s. n.], 1994:419-429.

[7] Xia B B. Similarity search in time series data set[D]. Canada: Simon Fraser University, 1997.

[8] 佚　名. GARCH 模型对沪市行业指数的实证研究[EB/OL]. 2008-10-12. http://www.govyi.com/lunwen/2008/200810/262942.shtml.

[9] 王　珏,秦伟良,钱海荣.上海股市的时间序列模型研究[J].理论新探,2004(11):11-23.