文章编号：1005-8451 (2004) 03-0005-03

ARIMA模型在HIS预测中的初步研究

叶明全
(合肥工业大学 计算机与信息学院，合肥 230009)

摘 要：时间序列预测分析在医院信息系统中具有广泛的应用前景，而ARIMA模型是挖掘时间序列模式的一个有效的方法。介绍利用ARIMA模型发现时间序列模式的方法，并应用于医院时序数据预测，为管理层提供决策依据，完善医院信息系统。

关键词：数据挖掘；医院信息系统；时间序列模式；ARIMA模型
中图分类号：TP39 文献标识码：A

Performance investigation of ARIMA model in HIS prediction

YE Ming-quan
(Institute of Computer & Information , Hefei University of Technology, Hefei 230009, China)

Abstract: Time series forecast analysis has a vast application in Hospital Information System, and ARIMA model is a useful method in mining time series models . It was introduced a method of time series model with ARIMA model .This model had been used in forecasting hospital time series, helped the predictors to make decision and improved the Hospital Information System .

Key words: data mining; Hospital Information System; time series models; ARIMA model

随着信息技术的高速发展，数据库应用的规模、范围和深度不断扩大。但数据库应用系统中缺乏挖掘数据背后隐藏知识的手段，导致了“数据爆炸但知识贫乏”的现象。数据挖掘就是在这种情况下产生并迅速发展起来的。数据挖掘模式的种类包括分类模式、回归模式、时间序列模式、聚类模式和关联模式

收稿日期：2003-11-03
作者简介：叶明全，在读硕士研究生。

3.4 控制

控制负责处理用户请求，调用相应的模型，更新模型的状态，刷新视图以及返回用户合理的页面。

3.4.1 请求处理器

请求处理器接收并处理用户的所有请求，调用请求事件转换器，把请求转换成预定义的事件，在事件处理完成后，进行视图更新。

3.4.2 客户控制器Web页面实施

客户控制器Web页面实施是调用EJB层的客户控制器的代理对象。

3.4.3 系统客户控制器

系统客户控制器是有状态的会话EJB，它为每个用户建立一个单独的实例，负责每个登陆帐号的生命周期，并负责处理事件；同时，它也控制状态机的生命周期。

3.4.4 状态机

状态机实现核心的业务逻辑，它负责改变模型的状态，包括处理每个业务事件的方法。

4 结束语

现场测试表明，该系统专家化的评价库，项目化的管理方式，工具化的简便应用，充分体现了设计时科学、公正的思想，而且操作简单易用。系统采用Java语言的MVC结构，其模块间关系清晰，调用途径明确，使系统具有很强的可维护性和可扩展性。

参考文献：

[1] Michael Girdley. J2EE应用与BEA Web Logic Server[M]. 北京：电子工业出版社，2002.
[2] Joseph L. Weber. Java，编程详解[M].北京：电子工业出版社，1999.
[3] 吴 俭.铁路货运技术[M].北京：中国铁道出版社，2000.
[4] 吴宗之，高进东，魏利军. 危险评价方法及其应用[M].北京：冶金工业出版社，2001.

等。时间序列模式就是从时间序列中寻找变化规律，并建立有关数学模型，从而获得系统的有关信息，预测将来的发展趋势，以揭示自然现象间的本质联系和内在规律性，从而有效地对客观现象及其变化规律进行预报和控制。

1 HIS时序数据预测

医院信息系统（Hospital Information System，简称HIS）是一项建立在计算机信息管理和医学基础上的交叉信息系统，是以提高医疗质量、经济效益、社会效益和工作效益为目的的信息管理系统。通过HIS可以高效地实现医疗和财务数据的录入、统计、查询等功能，但无法发现数据中存在的关系和规则，无法根据现有的数据预测未来的发展趋势，而隐藏在这些数据之后的更重要的信息是关于这些数据的整体特征的描述及对其发展趋势的预测，这些信息在决策生成的过程中具有重要的参考价值。在HIS中，时间序列数据是非常多见的，比如门诊量、急诊量、出院病人数、床位使用率、药材需求量和大型设备使用率等，通过对它们研究分析、预测可以提供医院管理层合理调配资源、动态调整工作计划的决策依据，从而具有重要现实意义。例如，门诊量数据的挖掘可能有助于医院制定门诊工作计划；根据医院床位使用率进行挖掘分析，可以预测将来床位使用率情况，从而为调配床位资源，方便患者住院治疗，为管理层提供科学决策依据。ARIMA模型是比较成熟的时间序列预测模型，已被作为一种理论比较完善、分析效果良好的工具，为此，可引入ARIMA模型对医院时间序列趋势预测进行数据挖掘，扩展HIS的功能和应用，为医院管理人员提供决策。

2 ARIMA模型建模原理

ARIMA（Autoregressive Integrated Moving Average，自回归求和平均）模型是由美国统计学家G. E. P. Box和G.M. Jenkins于1970年首次提出，该模型有3种基本模式：

（1）自回归模型（简称AR(p)模型）

$\phi(B)Y_t = e_t$，其中：B为后移算子，即$B^p Y_t=Y_{t-p}$；$\phi(B)=1- \phi_1 B- \phi_2 B^2-...- \phi_p B^p$

以上模型即为：$Y_t= \phi_1 Y_{t-1}+ \phi_2 Y_{t-2}+...+ \phi_p Y_{t-p}+e_t$，其中：$Y_t$，$Y_{t-1}$，...，$Y_{t-p}$分别是序列在t，t-1，...，t-p期的观测值。e_t是误差或偏差，表示不能用模型说明的随机因素，ϕ_1，ϕ_2，...，ϕ_p是待估计的参数。

（2）移动平均模型（简称MA(q)模型）

$Y_t = \theta(B) e_t$，其中：B为后移算子，即$B^q e_t=e_{t-q}$；$\theta(B)=1- \theta_1 B- \theta_2 B^2-...- \theta_q B^q$

以上模型即为：$Y_t= e_t- \theta_1 e_{t-1}- \theta_2 e_{t-2}-...- \theta_q e_{t-q}$，其中：$e_t$，$e_{t-1}$，...，$e_{t-q}$分别是序列在t，t-1，...，t-q期的误差或偏差，θ_1，..θ_q是待估计的参数。

（3）自回归移动平均模型（简称ARMA（p，q）模型）

$\phi(B)Y_t= \theta(B) e_t$

以上模型即为：$Y_t= \phi_1 Y_{t-1}+ \phi_2 Y_{t-2}+...+ \phi_p Y_{t-p}+e_t- \theta_1 e_{t-1}- \theta_2 e_{t-2}-...- \theta_q e_{t-q}$

ϕ_1，ϕ_2，ϕ_p，θ_1，θ_2，θ_q是待估计的参数。P，q称为自回归和移动平均的阶数。

很显然，这3种基本模型是ARIMA模型的特例，故ARIMA模型又常被作为这一族模型的总称。还有几种模型，它们是：

（1）自回归求和移动平均模型

当时间序列为非平稳序列，若通过d次差分可使序列平稳，采用的模型称作ARIMA(p，d，q)模型：

$\phi(B)(1-B)^d Y_t= \theta(B) e_t$

以上模型即为：

$(1-B)^d Y_t= \phi_1(1-B)^d Y_{t-1}+ \phi_2(1-B)^d Y_{t-2}+...+ \phi_p(1-B)^d Y_{t-p}+e_t- \theta_1 e_{t-1}-...- \theta_q e_{t-q}$

（2）季节性ARIMA模型

进行预测时考虑季节周期的因素，对于有季节性或周期性变动的数据，季节性ARIMA模型尤为适用。若sp，sq分别表示季节性自回归和季节性移动平均的阶数，sd为季节性差分的次数，则其一般形式为ARIMA$(sp, sd, sq)_S$，即：

$\delta(B^S)(1-B^S)^{sd}Y_t= \psi(B^S) e_t$，其中：S为周期中所包含的观察值数；

$\delta(B^S)=1- \delta_1 B^S- \delta_2 B^{2S}-...- \delta_{sp} B^{sp\,S}$，为季节性自回归算子；

$\psi(B^S)=1- \psi_1 B^S- \psi_2 B^{2S}-...- \psi_{sq} B^{sq\,S}$，为季节性移动平均算子。

（3）ARIMA乘积模型：ARIMA$(p, d, q)\times(sp, sd, sq)_S$模型

若一个季节性时间序列既含有季节性成分，又含非季节性成分，使用ARIMA乘积模型进行预测，模型为$\phi(B)\delta(B^S)(1-B)^d(1-B^S)^{sd}Y_t= \theta(B)\psi(B^S) e_t$

ARIMA模型是一种精确度较高的短期预测模型，

但其计算复杂，需借助计算机编程来完成。

3　应用实例

目前各大医院已建立医院信息系统，在医院管理及医学科研中，有时需对某项研究指标进行动态观察，可以从数据库中统计指标数据建立时间序列数据库，从而进行有效地挖掘原始时序数据中蕴藏的信息，进行趋势预测。如在医院药材管理系统中，若药材库存不足，容易发生供不应求，耽误病人治疗；若库存过剩，一方面药材积压，造成资金流通不畅，另一方面，药材质量下降，影响治疗效果。因此通过过去的药材月出库量构成时间序列，借助适当的预测手段，对近期或中期的药材需求量进行预测，为决策者制定采购计划提供了科学依据。

下面以某医院的每月中草药枸杞出库量（表1）为例，利用ARIMA模型短期预测将来每月枸杞出库量，从医院信息系统数据库中统计出1999年1月～2003年5月该院每月枸杞出库量，用1999年1月～2002年12月数据为时间序列建立模型，进行ARIMA模型拟合，对该院的剩余5个数据进行预测。

（1）绘制时间序列数据图：发现数据有明显的季节性变动规律，且年内呈现波动。

表1　某医院1999年1月～2003年5月枸杞出库量（单位：克）

年份	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
1999	32865	28643	33078	35538	37057	34197	30280	25706	28115	30435	36807	39776
2000	30690	27957	33814	37106	36933	35863	29079	24932	26501	28746	37730	41782
2001	33743	30969	35711	40762	41920	40840	32522	28257	31791	35799	42597	43905
2002	37710	33880	38919	40824	43228	37145	29688	27192	30986	36398	41319	45117
2003	38861	35086	41176	40229	44536							

（2）模型识别：对时间序列数据计算自相关函数和偏自相关函数，根据绘制的函数图形确定模型形式。具体步骤：先对时间序列数据作周期为12的季节性差分，并绘制序列的自相关函数图(ACF)和偏自相关函数图(PACF)，确定季节性ARIMA模型为$(0, 1, 1)_{12}$，再对季节性ARIMA模型$(0, 1, 1)_{12}$的残差序列作ACF图和PACF图，识别非季节ARIMA模型为(0, 1, 1)，最后求得模型形式为ARIMA乘积模型$(0, 1, 1)\times(0, 1, 1)_{12}$。

（3）参数估计和诊断检验：求得估计参数$\theta_1=0.2715$，$\phi_1=0.4006$，对ARIMA乘积模型$(0, 1, 1)\times(0, 1, 1)_{12}$的残差序列作ACF图和PACF图，残差的自相关函数和偏自相关函数均在可信限以内，不再包含可供建模的非随机成分，说明拟合效果好。

（4）预测：用求得的模型$(1-B)(1-B^{12})Y_t=(1-\theta_1B)(1-\phi_1B^{12})e_t$对剩余5个出库量进行预测，预测结果如表2。

表2　2003年药材出库量实际值与预测值对照表

月份	1月	2月	3月	4月	5月
实际值	38861	35086	41176	40229	44536
预测值	37785	34303	39321	42161	43951
相对误差	0.0277	0.0223	0.0451	0.048	0.0131

平均相对误差为0.0312，可见预测效果比较满意，说明ARIMA乘积模型对于有趋势性和周期性的数据预测具有一定的实用价值。

4　结束语

预测是HIS中数据挖掘的一个重要研究内容。时间序列数据预测是使用过去一个或多个时间序列的值发现将来某个序列的值。ARIMA模型克服了一般时间序列预测模型需对时间序列的发展规模作先验假设的局限，它先根据序列反复进行模型识别、参数估计和诊断检验，最终确定用于预测的最优模型。ARIMA可通过差分的方法将非平稳序列转变为零均值的平稳随机序列，以满足预测的前提。ARIMA使残差进入模型，提高预测模型的精度。但ARIMA建模法只考虑预测序列本身历史数据反映和包容的信息，几乎不直接考虑其他相关指标的信息，而时间序列是各种因素综合影响的结果，因此它主要适用于作短期的预测。在实际工作中若希望进行长期预测，可将ARIMA模型与其它模型（如神经网络模型）进行组合预测，综合利用各种方法所提供的信息，提高预测精度。

参考文献：

[1](美)G. E. P. Box, G. M. Jenkins. 时间序列分析预测与控制[M]. 顾　岚，等. 北京：中国统计出版社，1997.

[2]范　明，孟小峰. 数据挖掘—概念与技术[M]. 北京：机械工业出版社，2001.

[3]吕安民，柯　忠，等. 灰色系统模型在时间序列模式中的应用研究[J]. 微机发展，2002，(5)：41－43.

[4]张彦琦，黄　彦，等. SPSS在医院统计预测中的应用[J]. 中国医院统计，2002，(3)：131－134.