第30卷 第17期 计 算 机 工 程 2004年9月
Vol.30 №17 Computer Engineering September 2004
·基金项目论文· 文章编号：1000—3428(2004)17—0050—03 文献标识码：A 中图分类号：TP393

时间序列模式及其预测模型算法应用

吕林涛[1]，李军怀[1]，吕 晖[2]，王 鹏[1]，王志晓[1]
（1. 西安理工大学计算机学院，西安 710048；2. 重庆大学土木工程学院，重庆 400044）

摘 要：通过对时间序列模式分析研究，提出了时间序列的趋势性、季节性和随机性分析的应用模型及随机性12类预测数学模型算法，以该算法实现的数据挖掘系统经实际应用后效果很好。

关键词：时间序列模式；趋势性分析；季节性分析；随机性分析；预测模型算法；数据挖掘

Time Series Pattern and Its Application of Predictive Model Algorithm

LV Lintao[1], LI Junhuai[1], LV Hui[2], WANG Peng[1], WANG Zhixiao[1]
(1. Department of Computer and Engineering, Xi'an University of Science and Engineering, Xi'an 710048;
2. Department of Civil Engineering,Chongqing University,Chongqing 400044)

【Abstract】Based on the analysis of time series model, this paper proposes the application model of trend, seasonalness and randomness of time series, and also presents twelve algorithms of randomness. These algorithms are proved to be highly effective in data mining system.

【Key words】Time series pattern; Analysis of trend; Analysis of seasonalness; Analysis of random; Predictive model algorithm; Data mining

近年来,随着信息技术的迅猛发展，许多领域（政府、银行、税务、海关、企业等）搜索、积累了大量的数据，这些数据的背后隐藏着许多有用的重要信息。因而人们迫切需要一种新技术，能从海量数据中自动、高效地提取、分析所需的有用知识。

数据挖掘是一种新的信息处理技术，其主要特点是对企业数据库中的大量业务数据进行抽取、转换、分析和其他模型化处理，从中提取辅助企业决策的关键性数据[1-3]。

目前，国外关于基于时间序列模式分析及其预测模型算法在数据挖掘中的应用，取得了一批重要的成果，如SAS软件；国内中科院、复旦大学等在20世纪80年代对此类问题也作了理论研究，但应用性研究成果较少。

1 基于时间序列模式分析的数据挖掘算法

1.1 算法的定义

$$S_t=(U,\ A\cup\{d,t\},\ \prec)$$

其中：U为对象集（案例，状态，疾病，观测…）；A为属性（特征，变量，特点，条件…）；d为决策属性，$d\notin A$；t为顺序属性，$t\notin A$；$\prec$为顺序属性t上的一个次序关系，且$\prec=\{(x,y):x,y\in N,x<y\}$；$S_t$为t时刻预测数据。

1.2 时间序列模式分析及其预测模型算法的处理流程

经应用研究后本文提出，时间序列模式的分析及其预测模型算法的处理流程如图1所示。

2 时间序列模式分析及其预测核心算法的数学模型

经应用分析，已知时间序列数据存在分为两类：(1)具有趋势性和随机项性；(2)具有趋势性、季节性和随机性。因此，依据这两类分别提出，时间序列趋势性和随机性的分析及预测数学模型；时间序列趋势性、季节性和随机性的分析及预测数学模型。由于篇幅所限，本文仅给出时间序列趋势性、季节性和随机性分析的X-11方法数学模型。

定义1 若考虑具有季节性的月度数据序列，其周期d=12，设X_t的模型是：$X_t=M_t+S_t+Y_t$，其中M_t是趋势项、S_t是季节项、Y_t是随机项$S_{t+12}=S_t$，则X-11方法的实现步骤有下5点：

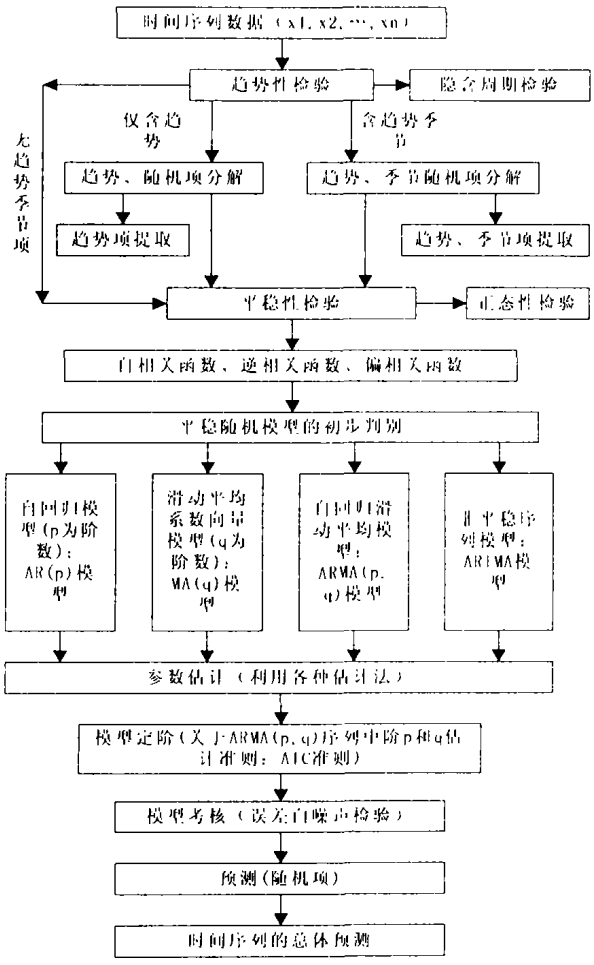

图1 时间序列的趋势性、季节性、随机性分析及其预测

基金项目：国家“863”计划基金资助项目（2001AA113182）
作者简介：吕林涛（1954—），男，副教授，研究方向为电子商务与网络安全；李军怀，博士；吕 晖、王 鹏、王志晓，硕士生
收稿日期：2003-09-03 **E-mail**：lvlintao@xaut.edu.cn

(1) 趋势的初估计

$$M_t^{(1)}=\frac{1}{12}[0.5X_{t-6}+X_{t-5}+X_{t-4}+X_{t-3}+X_{t-2}+X_{t-1}+X_t$$

$$+X_{t+1}+X_{t+2}+X_{t+3}+X_{t+4}+X_{t-5}+0.5X_{t+6}$$

(2) 季节项的初估计

记 $Z_t=X_t-M_t^{(1)}$ ，则季节项的初估计为

$$S_t^{(1)}=0.111Z_{t-24}+0.222Z_{t-12}+0.333Z_t+0.222Z_{t+12}+0.111Z_{t+24}$$

(3) 趋势项的估计

记 $U_t=X_t-S_t^{(1)}$ ，则趋势项的估计为

$$M_t^{(1)}=-0.019U_{t-6}-0.028U_{t-5}+0.066U_{t-4}+0.147U_{t-2}+0.214U_{t-1}+0.24U_t$$
$$+0.214U_{t+1}+0.147U_{t+2}+0.066U_{t+3}-0.028U_{t-5}-0.019U_{t+6}$$

(4) 季节项的估计

记 $G_t=X_t-M_t$ ，则季节项的初估计为

$$S_t=0.07G_{t-36}+0.13G_{t-24}+0.2G_{t-12}+0.02G_t+0.2G_{t+12}+0.13G_{t+24}+0.07G_{t+36}$$

(5) 随机项的估计

$Y_t=X_t-M_t-S_t$ ，这样得到 X_t 的分解为

$$X_t=M_t+S_t+Y_t$$

定义2 若用X-11作短期预测，则需要完成由平滑引起的缺失值的补齐、用回归延拓法估计S_t在[N-q+1，N]范围内的值、M_t的回归估计。即：

(1) 由平滑引起的缺失值的补齐

设观测数据是X_1，X_2，…，X_N，则定义

$$X_t'=\begin{cases}X_1 & t<1\\ X_2 & 1\le t\le N\\ X_1 & t>N\end{cases}$$

在X-11的滤波的每一过程都作这样的平直延伸处理。

(2) 设 $Y_t=\sum_{i=-q}^{q}a_iX_{t-i}$ ，$q+1\le t\le N-q$ 。用回归延拓法估计S_t在[N-q+1，N]范围内的值。设在[q+1，N-q]范围内S_t有L个周期$S_l(t)$，$1\le t\le 12$，$l=1,2,\cdots,L$。令

$$\hat{S}(t)=\frac{1}{L}\sum_{l=1}^{L}S_l(t+12(l-1))\,,\quad 1\le t\le 12$$

$\hat{S}(t)$ 作为理想的周期成分，按 $\hat{S}(t)$ 的周期延拓得到S_t在[N-q+1、N]范围的值。(同样可获得S_t在[1，q]的值)。这样，得到周期成分S_t的估计($1\le t\le N$)，仍记为M_t。

(3) M_t的回归估计

设M_t是由本节计算的M_t，用回归方法取线性趋势，平方趋势，三次趋势进行选择(用AIC准则)，最终得到趋势项，仍记为M_t。

定义3 若时间序列服从乘法模型 $X_t=M_tS_tY_t$，在用X-11方法得加法模型的趋势、季节、随机项分解及预测。再还原成X_t的分解及预测。

定义4 X-11的整体性滤波需进行5个步骤的运算，其中分解出S_t、M_t的步骤有4个。若考虑整体性滤波，将4个步骤合并成一个整体，直接计算出S_t、M_t。计算步骤描述如下：

(1) 生成Mt的平滑滤波

$$M_t=\sum_{i=-36}^{36}a_iX_{t+i}\qquad t=37,\cdots,N-36$$

其中：$a_{-i}=a_i$，a_i为Mt的平滑系数（略）。若用于平直延伸法扩充X_t的数据，可获得M_t，t=1，2，…，N的数据。M_t得到后，再配以合适的回归，仍记为M_t。

(2) 生成S_t的平滑滤波

$$S_t=\sum_{i=-28}^{38}b_iX_{t+i}\qquad t=39,40,\cdots,N-38$$

其中：$b_{-i}=b_i$，b_i为S_t的平滑系数(略)。若用平直延伸法扩充X_t的数据，可获得S_t，t=1，2，…，N的数据。若N=12m，则所得到的S_t近似为以12为周期的图形(共有m个这样的图形)。将这m个图形在l=1，2，…，12相应的点上的值平均，得到的图形S_t(仍以S_t记之)，按$S_{t-12}=S_t$研拓，得到S_t，t=1，2，…，N。

(3) 生成随机项

$Y_t=X_t-M_t-S_t$ ，t=1，2，…，N。这样，可以对 X_t 进行预测(如上所述)。

3 随机项预测核心算法的数学模型

本文提出了 $AR(p)$ 序列的ML建模等12类用于随机项预测的核心算法的数学模型。由于篇幅所限，本文仅给出利用逆转形式的 $ARMA(p,q)$ 序列预测算法的数学模型。

定义5 设 X_t 为$ARMA(p,q)$序列，$\varphi_1,\varphi_2,\cdots,\varphi_p;\theta_1,\theta_2,\cdots,\theta_q$。已知$x_t$的逆转形式是

$$x_t-\sum_{j=1}^{\infty}I_jx_{t-j}=\varepsilon_t$$

计算I_t的递推公式是

$$I_t=\begin{cases}\varphi_t+\sum_{j=1}^{t}\theta_t^*I_{t-j} & 1\le t\le p\\ \sum_{j=1}^{t}\theta_j^*I_{t-j} & t>p\end{cases}$$

其中：$\theta_t^*=\begin{cases}\theta_t & 1\le j\le q\\ 0 & j>q\end{cases}$

递推计算 $I_t^{(j)}$ 为

$$I_t^{(1)}=I_t$$

$$I_t^{(j)}=I_{t+1}+\sum_{i=1}^{t-1}I_iI_t^{(j-i)}$$

得

$$\hat{x}_k(l)=\sum_{t=1}^{r}I_t^{(l)}x_{k+1-t}\qquad j\ge 20$$

4 基于时间序列模式分析及其预测算法的数据挖掘系统

基于时间序列模式分析及其预测算法的数据挖掘系统B/S模型见图2。

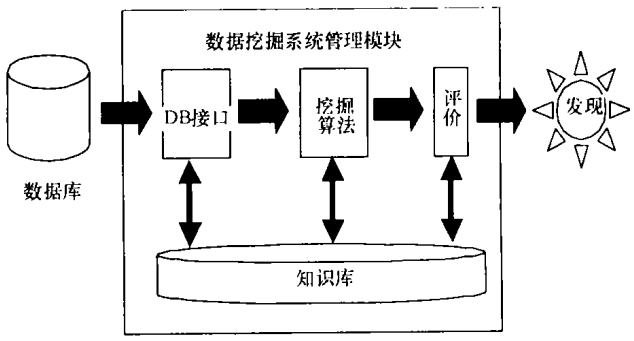

图2 数据挖掘逻辑模型

基于时间序列模式分析及其预测算法的数据挖掘系统模型实现效果见图3。

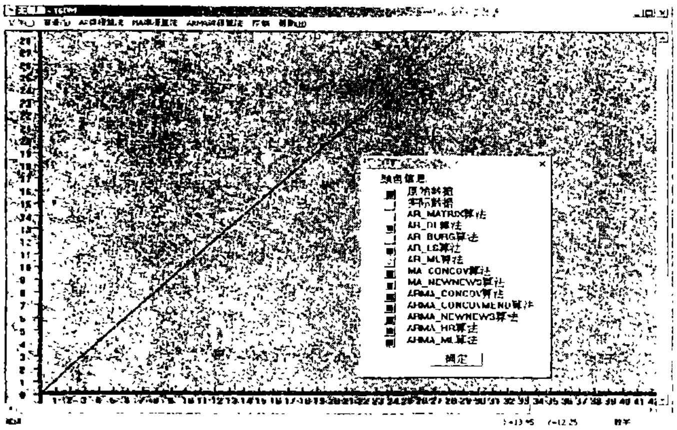

图3 系统B/S模型的实现效果

5 时间序列模式分析及其预测算法的应用效果

国际航线1949年～1956年月度旅客总数的统计数字如表1所示。

表1 国际航线月度旅客总数（1949.01-1956.12 单位：千人）

	1949	1950	1951	1952	1953	1954	1955	1956
1月	112	115	145	171	196	204	242	284
2月	118	126	150	180	196	204	242	284
3月	132	141	178	193	236	235	267	317
4月	129	135	163	181	235	227	269	313
5月	121	125	172	183	229	234	270	318
6月	135	149	178	218	243	264	315	374
7月	148	170	199	230	264	302	364	413
8月	148	170	199	242	272	293	347	405
9月	136	158	184	209	237	259	312	355
10月	119	133	162	191	211	229	274	306
11月	104	114	146	172	180	203	237	271
12月	118	140	166	194	201	229	278	306

国际航线月度旅客数的预测图见图4。

根据表1数字特征，图4选用的方法是时间序列趋势项、季节项和随机项的分解及预测中的滑动平均法，其中对随机项的预测使用的是AR中的ml方法，预测的月数为48个月。统计数据呈周期性地增加，根据预测，短期内乘客数将呈周期性增加，预测的结果与实际统计数据基本吻合。

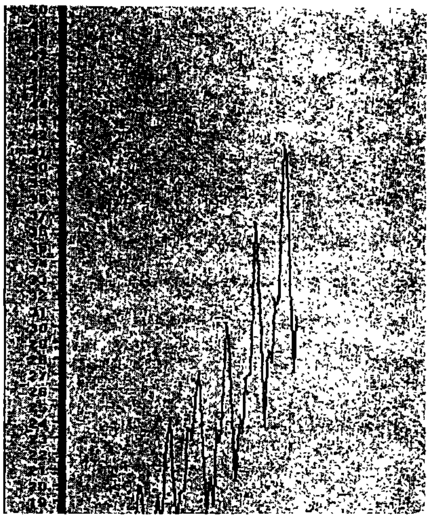

图4 国际航线月度旅客数的预测图

6 结束语

本文提出了时间序列模式分析的应用模型、随机性预测的12类数学模型算法，应用表明，它具有较高的理论和实用价值，可广泛应用于金融、保险、商业销售、电信等领域。

参考文献

1 Box G E P, JenKins G M, Reinsel G C. Time Series Analysis Forecasting and Control[M]. 北京：中国统计出版社, 1997

2 田 铮. 动态数据处理的理论与方法——时间序列分析[M]. 西安：西北大学出版社, 2001

3 马逢时, 何良材, 余明书等. 应用概率统计[M]. 北京：高等教育出版社, 1990

4 查特菲尔德 C. 时间序列分析导论[M]. 北京：宇航出版社, 1985

5 刘同明. 数据挖掘技术及其应用[M]. 北京：国防工业出版社, 2001

6 Ganti V,Gehrke J,Ramakrishnan R. Mining and Monitoring Evolving Data[J].IEEE Trans on Knowledge and Data Eng., 2001

7 Heikki D, Padhraic M. 张银奎, 廖 丽、宋 俊等译. 数据挖掘原理[M].北京：机械工业出版社, 2003

☆☆☆

（上接第13页）

$$\begin{cases} d_{(k,k+1)} = E_{k+1} - E_k \\ d_{(k-1,k)} = E_k - E_{k-1} \end{cases} \tag{16}$$

相应地计算出以两类标准偏差之和作比例系数下的单位距离分别为

$$\begin{cases} M_{(k,k+1)} = d_{(k,k+1)}/(\sigma_k + \sigma_{k+1}) \\ M_{(k-1,k)} = d_{(k-1,k)}/(\sigma_{k-1} + \sigma_k) \end{cases} \tag{17}$$

在考虑拒绝域的情况下，根据具体情况取$0<q<1$，则有$[E_k - qM_{(k-1,k)}\sigma_k, E_k + qM_{(k,k+1)}\sigma_{k+1}]$，这即为考虑拒绝域情况下被划分到第$k$类的特征矢量的平均频率所在的区间。

4 结语

通常情况下，平均频率用来对信号的频率进行估计，而多个宽带信号的分类使用的特征向量是N维的信号的频谱或功率谱；本文对两种方法进行了结合，阐述了以平均频率作为线性判别函数加权方法的特例的理论，并给出了基于标准偏差的学习方法，是对平均频率的一个新的观点，此方法可以用于以信号功率谱作为特征矢量的模式分类之中。它具有以下特点：(1)通过以信号的能量来进行模式的标准化，并且以信号的频率分量作为加权矢量，N维的模式分类问题简化为一维情况，减少了分类的复杂性。(2)平均频率的计算量少，而且物理意义清晰明确。(3)本文给出学习训练过程和特征矢量的分类过程简单，可以减少计算量。

作者将此方法应用于非线性信道数据传输系统中的码元检测，已经取得较好的结果。

参考文献

1 孙即祥. 现代模式识别[M]. 长沙：国防科技大学出版社, 2002

2 (美)科恩著. 白居宪译. 时-频分析：理论与应用[M].西安：西安交通大学出版社, 1998

3 盛 骤, 谢式千、潘承毅. 概率论与数理统计[M] .北京：高等教育出版社, 1990