基于时间序列的趋势性分析及其预测算法研究

吕林涛[1] 王 鹏[1] 李军怀[1] 吕 晖[2] 张 景[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)
E-mail:lvlintao@xaut.edu.cn

摘 要 文章通过时间序列分析研究,提出了基于时间序列的趋势性分析3类算法和随机性分析12类预测算法,以该算法实现的数据挖掘系统,经实际应用后其效果很好。因此,该算法在国民经济应用领域中具有较高的理论和实用价值。

关键词 时间序列 趋势性分析 预测算法 数据挖掘

文章编号 1002-8331-(2004)19-0172-03 **文献标识码** A **中图分类号** TP393

Research on the Trend Analysis and Predictive Algorithm Based on Time Series

Lv Lintao[1] Wang Peng[1] Li Junhuai[1] Lv Hui[2] Zhang Jing[1]
[1](Department of Computer and Engineering, Xi'an University of Technology, Xi'an 710048)
[2](College of Civil Engineering, Chongqing University, Chongqing 400044)

Abstract: Based on the analysis of time series, the three kinds of trend analysis algorithms and the twelve kinds of trend predictive algorithms based on time series are proposed in this paper. The data mining system using these algorithms has been proved to be highly effective, so it is very valuable in practice and academic study.

Keywords: time series, analysis of trend, predictive algorithm, data mining

近年来,随着信息技术的迅猛发展,许多领域搜索、积累了大量的数据,这些数据的背后隐藏着许多重要的信息,从而人们迫切地需要一种新技术从海量数据中自动、高效地提取、分析所需的有用知识[1,2]。数据挖掘就是适应这一要求迅速发展起来的一种数据处理新技术。

文献[3~8]给出数据挖掘的技术主要有序列分析、关联分析、分类分析、聚类分析、基于模糊集合的分析、基于神经网络的分析等等。时间序列是按时间顺序排列的、随时间变化又相互关联的数字序列。这样的例子在工程、经济等各个领域都广泛存在。

客观世界与工程实际中,会遇到各种各样的时间序列。要通过对这种时间序列的分析,达到认识事物、掌握事物的目的。所用的基本方法是对给定的时间序列选择合适的数学模型。这样的数学模型通常含有有限个未知参数,通过对这些参数的估计,最终完全建立起这个数学模型。当数学模型建立以后,就可以根据实际需要进行预报或控制。近年来时间序列分析发展非常迅速,在气象、天文、水文、机械、电力、生物、经济等各个领域已有广泛的应用,显示出强大的生命力。

1 时间序列的变化分解模型

时间序列分析的目的可以概括为四个方面:描述、推断、预报和控制。描述就是通过对数据的分析建立适当的数学模型来描述产生数据的随机机制。推断是根据某一随机机制所产生的数据,分析判断它是否具有某些指定的属性,或者由多个不同的时间序列,分析不同的随机机制是否具有相同的属性。为此,必须假设可用适当的一类模型来描述产生观测数据的随机机制。预报是利用时间序列的相关性,预报随机机制在未来时刻的取值。控制是对某一随机机制(或多个随机机制)的一段观测数据的分析,寻求对某些量的控制,以达到优化目的。

为了能根据时间序列信息推断对象之间的依赖关系或进行预测。例如,以"月"为单位是非常重要的,它确定了对象的顺序,故称之为顺序属性。所以必须保证对象的正确顺序,下面引入时间信息系统,目的是将对象的顺序(或次序)信息形式化。

1.1 时序信息系统TIS定义

定义1.时序信息系统TIS满足:

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

其中:$U\Rightarrow$对象集(案例,状态,疾病,观测……);$A\Rightarrow$属性(特征,变量,特点,条件……);$d\Rightarrow$决策属性,$d\notin A$;$t\Rightarrow$顺序属性,$t\notin A$;$<\Rightarrow$顺序属性t上的一个次序关系,且$<=\{(x,y):x,y\in N,x<y\}$。

1.2 时间序列分析的变化分解模型

现实中的时间序列的变化受许多因素的影响,有些起着长期的、决定性的作用,使时间序列的变化呈现出某种趋势和一定的规律性,有些则起着短期的、非决定性的作用,使时间序列的变化呈现出某种不规则性。

该文提出,时间序列分析具有以下三种变化分解形式:

(1)趋势变动。指现象随时间变化朝着一定方向呈现出持续稳定地上升、下降或平稳的趋势。

基金项目:国家863高技术研究发展计划资助(编号: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、数据挖掘。王鹏(1979-),男,硕士研究生,研究方向为数据挖掘算法研究。李军怀(1969-),男,博士,研究方向为分布式计算、CSCW。吕晖(1981-),男,硕士研究生。张景(1952-),男,博导,研究方向为Internet应用。

(2)周期变动(季节变化)。指现象受季节性影响,按某固定周期呈现出的周期波动变化。

(3)随机变动。指现象受偶然因素的影响而呈现出的不规则波动。

2 基于时间序列的趋势性分析算法

基于1.2节变化分解三种形式,该课题提出了三种基于时间序列的趋势性分析算法,即滑动平均算法、整体滤波算法和差分算法。由于篇幅所限,该文仅给出滑动平均算法和整体滤波算法的数学模型。

定义2.时间序列的加法模型:

$X_t=M_t+S_t+Y_t$

并满足 $S_{t+d}=S_t$, $\sum_{j=1}^{d} S_j=0$。

其中 X_t:原始数据项,M_t:趋势项,S_t:周期项(季节项),Y_t:随机项。

定义3.时间序列的乘法模型:

$X_t=M_tS_tY_t$

2.1 滑动平均算法及应用效果

2.1.1 滑动平均算法

定义4.设观测值是 $X_1,X_2,\cdots X_N$,设 q 是非负整数,在每个周期内趋势 M_t 项近似为常数,第一步选择可消除趋势项,并用噪声衰减的滑动平均滤波器估计趋势项:

(1)当周期 d 是偶数时,令 $d=2q$,则:

$$M_t=\frac{0.5X_{t-q}+X_{t-q+1}+\cdots+X_{t+q-1}+0.5X_{t+q}}{d},\ q+1\le t\le N-q$$

(2)当周期 d 是偶数时,令 $d=2q+1$,则:

$$M_t=\frac{X_{t-q}+X_{t-q+1}+\cdots+X_{t+q-1}+X_{t+q}}{d},\ q+1\le t\le N-q$$

定义5.当偏差满足 $\{(X_{k+jd}-M_{k+jd}):q+1\le k+jd\le N-q\}$ 时,则平均值 W_k:

$$W_k=\frac{\sum_{j=1}^{d}(X_{k+j\times d}-M_{k+j\times d})}{l_k},\ k=1,2\cdots d$$

其中 j:满足 $q+1\le k+jd\le N-q$ 的和,l_k:满足不等式的个数。

由此得季节项 S_k:

$$S_k=W_k-\frac{\sum_{i=1}^{d}W_k}{d},\ k=1,2\cdots N \text{ 且 } S_k=S_{k-d}\quad k>d$$

定义6.若需进一步估计趋势项时,则求解消除季节项后的数据 d_t:

$d_t=X_t-S_t\quad t=1,2\cdots N$

2.1.2 应用效果

已知某国际航线月度旅客数的统计数据(见表1),应用滑动平均算法其效果基本与实际测试的结果相吻合(如图1、图2、图3所示)。

2.2 整体滤波算法及应用效果

2.2.1 整体滤波算法

定义7.生成 M_t 的平滑滤波:

$$M_t=\sum_{j=-36}^{36}A_jX_{t+j}\quad t=37,\cdots,N-36$$

其中:$A_{-j}=A_j$,A_j 的值已知(略)。

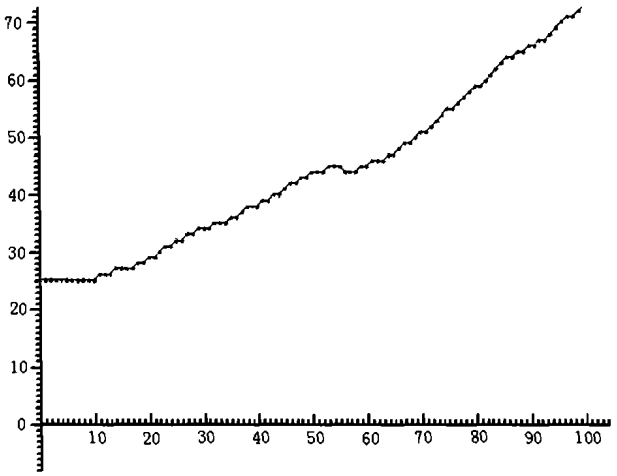

图1 趋势项

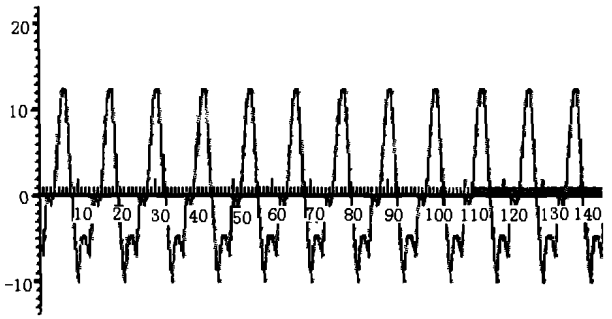

图2 周期项

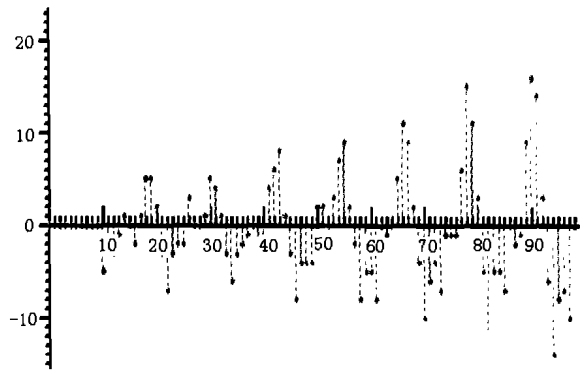

图3 随机项

注:图中表示10人/单位长度

注:若用平直延伸法扩充 X_t 的数据,可获得 $M_t(t=1,2\cdots N)$ 的数据。M_t 得到后,再配以合适回归,仍记为 M_t。

定义8.生成 S_t 的平滑滤波:

$$S_t=\sum_{j=-38}^{38}B_jX_{t+j}\quad t=39,\cdots,N-38$$

其中:$B_{-j}=B_j$,B_j 的值已知(略)。

注:若用平直延伸法扩充 X_t 的数据,可获得 $M_t(t=1,2\cdots N)$ 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 $l=1,2\cdots12$ 相应的点上的值平均,得标准图形 S_t(仍以 S_t 记之)按 $S_{t+12}=S_t$,延拓得到 $S_t,t=1,2\cdots N$。

定义9.随机项 $Y_t=X_t-M_t-S_t,t=1,2\cdots N$,这样可以对 X_t 进行预测。

2.2.2 应用效果

已知某国际航线月度旅客数的统计数据(见表 1),应用整体滤波算法其效果基本与实际测试的结果相吻合(如图 4、图 5、图 6 所示)。

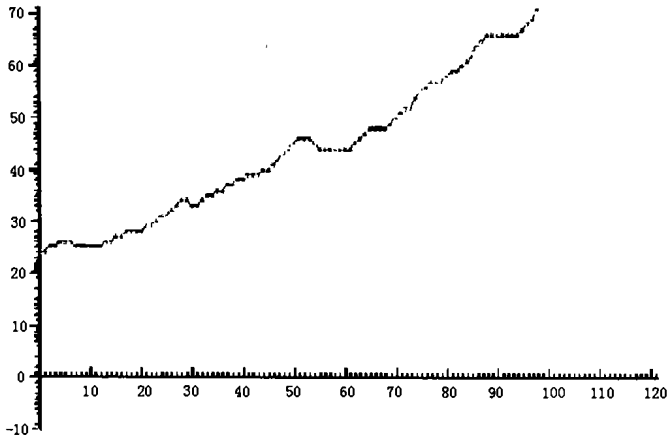

图 4 趋势项

图 5 周期项

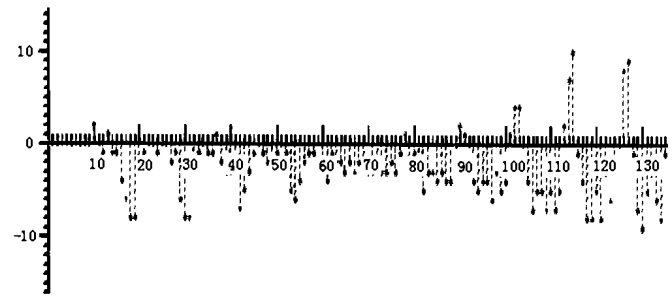

图 6 随机项

注:图中表示 10 人/单位长度

3 预测算法

该课题提出了基于时间序列的预测 12 类算法,由于篇幅所限,这里仅给出其中的一种。

定义 10.依据 2.1~2.2 的算法,时间序列分析的预测算法描述为:

(1)选取线性趋势、平方趋势、三次趋势和指数趋势曲线拟合中的一种方法,并遵循序列中的估计准则 AIC,得到一个合适的趋势项 $M_N(l)$;

(2)用延拓的方法向下推算得到季节项 $S_N(l)$;

(3)用自回归滑动平均模型 $ARMA(p,q)$ 进行预测得随机项 $Y_N(l)$;

(4)求解 $X_N(l)=M_N(l)+S_N(l)+Y_N(l)$,得最终预测值。

4 基于时间序列的趋势性分析及其预测算法在数据挖掘中的 B/S 模型

4.1 B/S 模型的定义

该课题提出的数据挖掘系统,是以数据库(或数据仓库)的数据、信息分析员的指导以及存储在挖掘系统中的知识和规则为基础,所选择的数据在各自挖掘模块中加以处理,并生成辅助模块和关系。然后进行评价,通过与分析员交互以期发现不同的模式。最后将这些发现加入知识库,以便后续的抽取和评价。因此,B/S 模型的定义如图 7 所示。

4.2 B/S 模型的功能定义

该课题提出的 B/S 模型的功能主要由 4 大模块组成:

(1)数据接口模块:提供本原型系统和数据库访问接口。

(2)数据预处理模块:该模块的目的是对原始数据进行处理,生成数据挖掘工具可利用的数据。

(3)时间序列分析模块:包括若干时间序列模型的工具。

(4)结果及可视化模块:对挖掘结果进行评估,给出各种直观的可视化图形显示方法,该文采用可视化的图形方式表示数据挖掘的结果。

5 基于时间序列的趋势性分析及其预测算法在数据挖掘中的应用效果

5.1 国际航线 1949 年~1956 年月度旅客总数的统计数字(见表 1)

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

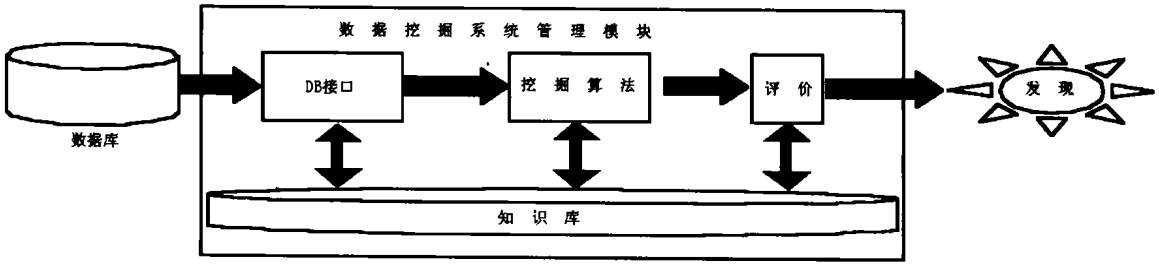

图 7 数据挖掘逻辑模型

(下转 208 页)

或标志字的标志位以及对临界区加锁等方法来实现。

主进程和子进程之间同样存在同步与互斥问题，主要是通过设置一些系统标志字来解决的。如：

```
PROGRAM TRK620  !TRK620处理程序
                ...
  FLAG(7)=2  !FLAG(7)同步信号量
  STATUS=SYS$HIBER()  !设置为睡眠状态
END PROGRAM TRK620
```

4.3 数据通信方式

在内存映像网卡的内存中指定一个区域（如64K~200K）作为共享数据区，只要定义数据区中字节或位代表的含义或数据，子系统就可以通过对同一共享数据区的操作来实现相互之间的通信。

主进程和子进程之间数据通信采用共享存储区方式，其定义如下：

```
MODULE MOD_GLOBAL_COMMON !系统全程公用区声明
  STRUCTURE /DI_SIGNAL/ !输入信号结构定义
    INTEGER DIGITAL_INPUT(54)
  END STRUCTURE
                ...
  !系统标志字
  INTEGER FLAG(10)
  INTEGER HMD_NO
                ...
  !系统公用区声明
  COMMON /TRAKING_INDEX_TABLE/ MTRKI  !输入信号结构定义
    COMMON /TRACKING_POINTER_TABLE/ MTRKP
  END MODULE MOD_GLOBAL_COMMON
```

4.4 C/S(Client/Server)结构模式的应用

采用了Visual C++、Visual Basic工具软件自行开发C/S结构模式网络数据库的MMI应用程序，这样节约了因购买昂贵的专用商业软件所需的费用，降低了系统的成本。

4.5 系统可扩展性

子系统B中的NT Server中的应用程序留有接口，可以接收上一级系统（如：生产管理系统）的数据，实现了系统的扩展。

5 结束语

采用上述拓扑结构和软件设计技术的中厚钢板计算机过程控制系统，实现了中厚钢板生产过程中的板坯位置和数据跟踪、顺序控制、轧制规程计算等功能。实践证明，整个系统功能完善，性能稳定，实时性强，扩展性好，数据通信迅速且准确无误，完全能满足中厚钢板生产过程控制的要求。

（收稿日期：2004年4月）

参考文献

1.孙一康.自行集成适用带钢热连轧的计算机控制系统[J].北京科技大学学报，1999；(5)

2.郑雪锋.轧钢计算机控制系统集成技术[J].冶金自动化，2001；(5)

3.VMIC PCI-5576 Reflective memory board product manual document. NO.500-85576-000A

4.VMIC products Instruction 1998—1999

（上接174页）

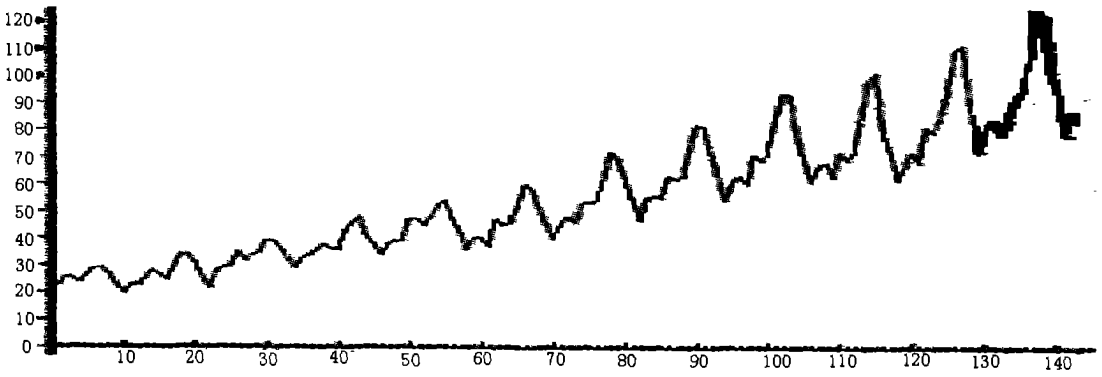

图8 应用效果

5.2 基于时间序列的趋势性分析及其预测算法在数据挖掘中的应用效果（如图8所示）

图8中表示10人/单位长度，灰颜色表示原始数据，黑颜色表示预测数据。其结果表明原始数据与预测数据吻合得很好，从而验证了算法的正确性。

6 结束语

该文提出的基于时间序列的趋势性分析及其预测算法，在电信系统数据挖掘中应用后表明，该算法具有很好的理论和实用价值，其效果很好。可广泛应用于金融、保险、商业、销售、电信等领域。（收稿日期：2004年1月）

参考文献

1.[美].George E P Box,[英]Gwilym M JenKins,[美] Gregory C Reinsel.时间序列分析预测与控制(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.中国科学院计算中心概率统计组编著.概率统计计算[M].北京：科学出版社，1979

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

6.刘同明.数据挖掘技术及其应用[M].北京：国防工业出版社，2001

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

8.Zaki m J.Scalable Algorithms for Association Mining[J].IEEE Trans on Knowledge and Data Eng，2001