文章编号:1000 - 582X(2005)10 - 0091 - 03

回归分析在连续型数据目标预测中的应用*

蔡章利[1],石为人[1],刁 绫[2]
(1. 重庆大学 自动化学院,重庆 400030; 2. 重庆电子职业技术学院 计算机系,重庆 400043)

摘 要:数据挖掘能从已有的大量数据中抽取隐含的、以前未知的、具有潜在应用价值的信息或模式. 如何从数据仓库中提取知识辅助用户决策是开发决策支持系统必须解决的问题. 针对所开发工业企业市场营销决策支持系统时遇到的连续型数据目标预测问题,选用回归分析方法,系统地探讨了如何建立挖掘模型和设计挖掘算法等问题,并将其用于销售预测. 对模拟数据进行处理,结果表明算法能实现预期效果.

关键词:决策支持系统;回归分析;数据挖掘;销售预测
中图分类号:TP182 **文献标识码**:A

数据挖掘是把人工智能、机器学习、数据库等技术结合起来,由计算机自动从已有的大量数据中抽取隐含的、以前未知的、具有潜在应用价值的信息或模式的过程,目的是为了解决数据量大而知识贫乏的矛盾,能完成关联分析、时序模式、分类、聚类、偏差监测、预测6个方面的任务[1].

目前,用于数据挖掘的方法和技术已很多,有归纳学习法、仿生物技术、公式发现、统计分析方法、模糊数学方法和可视化技术六大类[2]. 其中,能用于连续型数据目标预测的,有神经网络法和回归分析法. 神经网络是一种仿生物技术,回归分析则是统计分析方法中的一种.

由于销售预测这种连续型数据目标预测在用户开发决策支持系统时会经常遇到,因此在讨论多元线性回归分析原理的基础上,以二元线性回归分析为例,设计了数据挖掘算法,并将其用于笔者开发的工业企业市场营销决策支持系统(MDSS).

1 回归分析原理

回归分析是将数理统计规则用于研究自变量与因变量之间关系形式的方法,目的是希望根据已知的自变量来预测因变量. 将回归分析方法用于数据挖掘的预测任务,实际上就是利用大量的历史数据,以时间为变量建立线性或非线性回归模型. 预测时,只要输入任意的时间值,利用建立的回归模型就可求出该时间的状态.

1.1 多元线性回归模型

设研究对象受多个因素 $x_1, x_2, \cdots, x_m$(自变量)影响,各影响因素与预测目标(因变量)的关系是线性的,则其多元回归线性模型为:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im} + \varepsilon_i, \quad i = 1, 2, \cdots, n. \tag{1}$$

$y_i, x_1, x_2, \cdots, x_m$ 为预测目标和影响因素的第组观测值,ε_i 是第 i 组观测值对 y_i 的随机误差,通常取观测值 x_{i1} 恒等于1;其矩阵形式为:

$$\boldsymbol{Y} = \boldsymbol{XB} + \varepsilon, \tag{2}$$

用最小二乘法估计回归系数向量 $\boldsymbol{B}$,可得其估计值 $\hat{\boldsymbol{B}}$ 为:

$$\hat{\boldsymbol{B}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}, \tag{3}$$

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & x_{12} & \cdots & x_{1m} \\ 1 & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n2} & \cdots & x_{nm} \end{bmatrix},$$

$$\boldsymbol{B} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

1.2 模型显著性检验

为了判断多元线性回归模型所反映的各变量之间

* 收稿日期:2005 - 05 - 25
基金项目:重庆市制造业信息化重大专项资助项目(2001 - 03)
作者简介:蔡章利(1973 -),男,四川大竹人,重庆大学硕士研究生,研究方向:智能决策理论与应用.

的关系形式是否符合客观实际,引入的因素是否有效,用户将模型用于实际预测工作前,需对模型进行显著性检验. 常用方法有 R 检验法、F 检验法、t 检验法和 DW 检验法 4 种[3-4]. 下面仅给出复相关系数 R、F 统计量、t 统计量、DW 统计量的计算公式.

$$R = \sqrt{1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y}_i)^2}}, \tag{4}$$

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - m}{m - 1}, \tag{5}$$

$$t_j = \frac{\hat{\beta}_j}{S_{\hat{\beta}_j}}, j = 1,2,\cdots,m, \tag{6}$$

$$DW = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}. \tag{7}$$

1.3 预测区间估算

回归模型通过上述显著性检验后,可用于实际预测工作. 在多元线性回归模型中,将给定的一组自变量值 $x_{01},x_{02},\cdots,x_{0m}$ 代入回归模型,便可求得对应的回归预测值 $\hat{y}_0,\hat{y}_0$ 又称为点估计值. 由于在实际工作中,预测目标的实际值不一定刚好等于预测值,受各种环境因素的影响,两者总会产生或大或小的偏差. 如果人们仅根据某一点的预测计算就得出结论,则总是存在谬误. 因此,不仅要预测出 y 的点估计值,而且还要给出 y 的预测区间. 当预测值 $\hat{y}_0$ 的显著性水平为 α 时,多元线性回归模型的预测区间可用式(8)来估算得到.

$$\left.\begin{aligned} \hat{y}_0 \mp t_{\alpha/2}(n - m)S, \quad n < 30 \\ \hat{y}_0 \mp Z_{\alpha/2} \cdot S, \quad n \geqslant 30 \end{aligned}\right\} \tag{8}$$

2 模型及算法设计

2.1 模型设计

在 MDSS 系统中,定义如下结构的数据仓库来保存销售数据,并以它为数据挖掘对象[5-7]. 定义时间 T 和产品 P 为自变量,销售量 Y 为因变量,建立如下销售预测模型:

$$Y = \beta_0 + \beta_1 T + \beta_2 P + \varepsilon, \tag{9}$$

β_0,β_1,β_2 为回归系数,可用后面设计的算法求得. 当用户输入时间和产品代码时,代入上式即可得到销售量.

2.2 算法设计

为便于算法描述,此处预定义如下存储结构和常量:

```
TypeDef Struct TwoLineRegressionModel {
    float    BI[3];    //回归系数
    int      num;      //样本数
    float    Level;    //显著性水平
    float    Error;    //随机误差
};//二元线性回归模型的存储结构定义
TypeDef Struct SampleData {
    float    y;    //因变量
    float    x1;   /自变量
    float    x2;   //自变量
};//样本数据存储结构定义
#define  TRUE   1
#define  FALSE  0
```

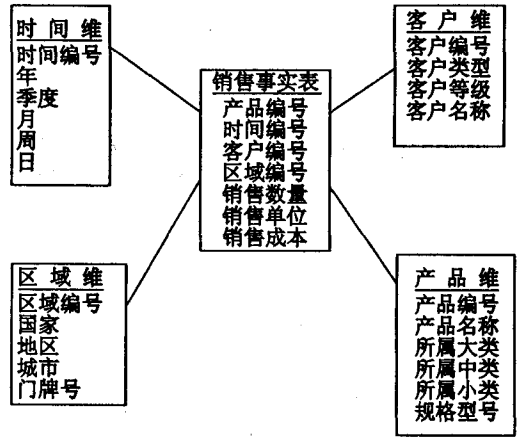

图 1 用于数据挖掘的数据仓库结构

函数 CreateTLRModel( ) 是笔者设计的一个用于创建二元线性回归模型的算法,其实现描述如下:

算法:CreateTLRModel( ) 由给定的样本数据创建一个二元线性回归模型.

输入:样本数据 SData、显著性水平 Level.

输出:二元线性回归模型 TLRM.

方法:

1) 定义表示自变量矩阵 X、因变量 Y 的数组 $X[n][3]$ 和 $Y[n]$;

2) 从数据仓库的销售事实表中根据系统指定的时间单位和时间范围取样本数据到数组 $X[n][3]$ 和 $Y[n]$;并将得到的样本数 n 和系统指定的显著性水平 α 存放到 TLRM 中;

3) 按式(3)求回归系数 β_0,β_1,β_2 的估计值,结果保存到 TLRM 中;

4) 调用函数 RCheck( ) 进行 R 检验,若检验通过,程序继续执行,否则终止程序,输出错误信息;

5) 调用函数 FCheck( ) 进行 F 检验,若检验通过,程序继续执行,否则终止程序,输出错误信息;

6) 调用函数 TCheck( ) 进行 t 检验,若检验通过,程序继续执行,否则终止程序,输出错误信息;

7) 调用函数 DWCheck( ) 进行 DW 检验,若检验通过,程序继续执行,否则终止程序,输出错误信息;

8) 调用函数 ECaculate( ) 计算随机误差 ε,并将其保存到 TLRM 中;

9) 输出并保存 TLRM,存储结构 TLRM 中保存的值即为所求的销售预测模型参数;

函数 CreateTLRModel( ) 是创建二元线性回归模

型的主算法,运行时,需要调用其它函数如 RCheck()、FCheck()、TCheck()、DWCheck()、ECaculate()等共同完成,其运算主要集中在自变量矩阵 X 和因变量矩阵 Y 的相关运算上,即求回归系数向量 B 的估计值 $\hat{B}=(X'X)^{-1}X'Y$;时间复杂度为 $O(n)$,与样本数 n 成正比.

3 应用实例

为检验上述算法的运行效果,笔者将其应用于MDSS系统的销售预测.

3.1 问题分析

对MDSS系统中数据仓库存储的销售数据进行挖掘处理,以时间和产品为自变量,销售量为因变量,建立销售预测模型;当用户输入时间和产品代码时,系统能利用该模型得到销售量.

3.2 数据准备

算法运行前,需要系统预先提供样本数据及样本数,并指定模型的显著性水平.

系统从数据仓库的销售事实表中提取样本数据时,由于时间(如1999-10-21)和产品代码(如TS1700405-0A0-472A-C)通常不是数值型的,不能直接用于运算.对于产品,由于其代码是惟一的,设有一个int型的字段与产品代码一一对应,用该int型字段参与运算;对于时间,则构建了一个哈希函数来帮助完成转换,以满足在不同时间单位(年、季、月、周、日)上进行销售预测.

3.3 运行效果

用VC++语言实现前述算法后,笔者准备了3 000条模拟数据放入MDSS系统的数据仓库中,执行算法程序,对其进行挖掘处理.从得到的结果来看,算法能达到预期效果.

4 结束语

基于数据仓库、联机分析处理、数据挖掘的决策支持系统开发已成为研究热点,如何从数据仓库中提取知识辅助用户决策是开发决策支持系统必须解决的问题.针对笔者开发MDSS系统中遇到的连续型数据目标预测问题,选用回归分析方法,系统地探讨了如何设计挖掘算法,如何建立挖掘模型等问题;将其用于销售预测,对模拟数据进行处理,结果表明算法能实现预期效果.

参考文献:

[1] HAN J W, MICHELINE KAMBER. 数据挖掘——概念与技术(影印版)[M]. 北京:高等教育出版社,2001.

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

[3] 宁宣熙,刘思峰. 管理预测与决策方法[M]. 北京:科学出版社,2003.

[4] 大漠达尔N.古亚拉提著. 经济计量学精要[M]. 第2版. 张涛,王智勇,王宏伟译. 北京:机械工业出版社,2000.

[5] HAN J W, CHEE S, CHIANG J Y. Issues for On-line Analytical Mining of Datawarehouses[A]. Proc of 1998 SIGMOD 96 Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD98)[C]. Seattle, Washington, 1998.

[6] HAN JIAWEI, MICHELINE KAMBER. Data Mining: Concepts and Techniques[M]. San Francisco: Morgan Kanfmann Publishers, 2001.

[7] 黄晓霞,萧蕴诗. 数据挖掘集成技术研究[J]. 计算机应用研究,2003,21(4):37-39.

Regression Applied in Continuous Mathematical Forecast

CAI Zhang-li[1], *SHI Wei-ren*[1], *DIAO Ling*[2]

(1. College of Automation, Chongqing University, Chongqing 400030, China;
2. Chongqing Electronics Polytechnic College, Chongqing 400043, China)

Abstract: Data Mining can distill the connotative, unknown and potential value information or pattern from many data. It is a problem that must be settled during the developing of Decision Support System how to discover knowledge from Data Warehouse. In order to realize the continuous mathematical forecast that will be used in the Marketing Decision Support System for Industrial Enterprise, the authors use the method of regression to discuss how to design the algorithm for Data Mining, how to design the model for Data Mining, and so on. When the algorithm applied to the sales forecast, the result processing to simulation data indicates it can realize the expected effect.

Key words: decision support system; regression; data mining; sales forecast

(编辑　吕赛英)