文章编号:1004-1478(2011)06-0080-03

双向数据挖掘的反馈预测分析

高利娟, 刘云, 赵玲

(昆明理工大学 信息工程与自动化学院, 云南 昆明 650500)

摘要:针对常规的数据挖掘预测模型只进行单一预测且未对预测的准确度进行分析等问题,提出了一种将关联模型和基于最小二乘法的回归分析模型相结合的反馈预测模型,并使用最小均方算法对预测进行误差分析.仿真结果表明该模型对中长期预测具有较高的精确度.

关键词:数据挖掘;关联规则;最小二乘法;反馈预测;最小均方算法

中图分类号:TP311.132.4 **文献标志码**:A

Feedback prediction analysis of two-way data mining

GAO Li-juan, LIU Yun, ZHAO Ling

(*Faculty of Infor. Eng. and Auto., Kunming Univ. of Sci. and Tech., Kunming* 650500, *China*)

Abstract: Aiming at the problem that data mining prediction model can only make simplex forecast and not respond to the accuracy of forecast analysis, in order to forecast future trends, a feedback predictive model was put forward based on combined correlation model and LMS regression model, and minimum mean square algorithm was used to verify accuracy of the model. The simulation results showed that this model has high accuracy in prediction for mid and long-term.

Key words: data mining; association rules; LS; feedback prediction; LMS algorithm

0 引言

数据挖掘(data mining)是发现数据中有用模式的过程,其会话的目的是确定数据的趋势和模式.数据挖掘模型按照功能分为预测模型和描述模型,描述性数据挖掘的任务是刻画数据的一般特性,预测性数据挖掘的任务是在当前的数据上进行判断,以便预测.

目前所用的预测技术只是单一地采用预测模型,如自回归模型、自回归滑动平均模型等,或者采用改造后的统计学方法,如基于n阶移动平均值、最小二乘法(LS)、徒手法等的回归预测技术[1].另外,一些研究较早的数据挖掘分支,如分类、关联规则等,其技术也被应用到趋势预测中.然而,常规趋势预测方法存在许多缺陷:只单纯地采用预测模型或者关联规则模型[2],而没有考虑到众多看起来不相关数据信息之间的关联性;直接采用预测模型对未来趋势进行预测,没有对预测的准确度进行核对.因此这种预测技术可能会造成错误的预测.

针对以上问题,本文拟利用关联模型和预测模型中的基于最小二乘法的回归模型相结合模型进行趋势预测,并使用最小均方(LMS)算法对预测进行误差分析,以提高预测准确度,进而提高中长期预测的精确度.

收稿日期:2011-05-31

基金项目:国家自然科学基金项目(10502050)

作者简介:高利娟(1986—),女,河南省濮阳县人,昆明理工大学硕士研究生,主要研究方向为无线通信、数据通信、多媒体通信.

1 数据挖掘预测

1.1 反馈预测模型

利用最小二乘法回归模型进行数据挖掘预测分析,在预测准确度的验证上存在一定的缺陷. 为了弥补这种缺陷,本文利用最小均方(LMS)算法对模型的准确性进行分析. 反馈预测模型如图1所示.

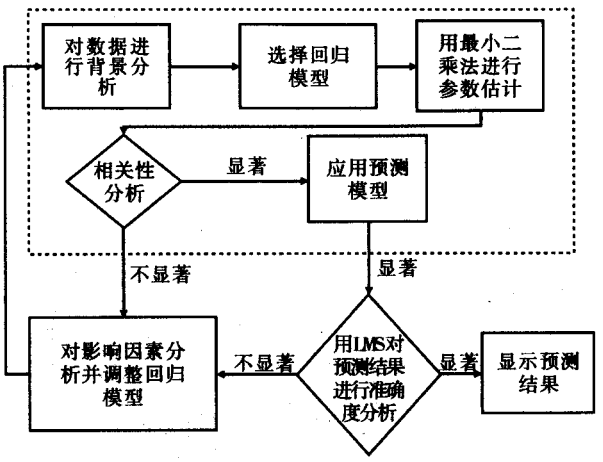

图1　反馈预测模型

1.2 预测分析

预测分析是数据挖掘的一个重要任务. 预测的目的是从历史数据中自动推导出对给定数据的推广描述,以便对未来数据进行预测. 本文主要通过基于最小二乘法的回归分析模型来进行数据挖掘的趋势预测. 在进行预测分析时,首先应对数据进行相关性分析.

1.2.1 相关性分析　关联规则挖掘反映一个事件和其他事件之间依赖或关联的知识[3]. 如果2项或多项变量之间存在关联,那么其中1项的变量值就可以依据其他变量值进行预测,在此挖掘中应对多种变量之间进行关联分析[4].

相关分析主要是分析变量之间关系的密切程度,其目的是确定变量之间的关系类型及其关联的程度,并探索其内在的数据规律性,它是回归模型分析的基础[5].

文中采用一元线性关联模型 $y = a + bx$. 为了了解变量 x 与 y 之间的线性相关程度,可根据公式①计算2个变量之间的相关系数 r.

$$r = n\left(\sum x_i y_i - \sum x_i \sum y_i\right) \times \sqrt{\left(n\sum x_i^2 - \left(n\sum x_i\right)^2\right)} \times \sqrt{\left(n\sum y_i^2 - \left(n\sum y_i\right)^2\right)} \qquad ①$$

如果2个变量之间的相关系数为1或 -1,那么就可以由变量 x 得到变量 y 的值,相关系数的绝对值越高,从1个变量去预测另1个变量的精确度就越高,这是因为相关系数越高,就意味着这2个变量的共变部分越多.

1.2.2 回归模型　回归模型是在对历史数据进行分析的基础上建立的[6]. 其目的是通过已知的变量值来预测其他变量值,进而找到一个联系输入变量和输出变量的最优模型. 回归分析试图从实际数据中寻找某种规律,确立和分析某种因变量 y 和自变量 x_i 之间的函数关系,即找出适当的函数使得 $y = f(x_1, x_2, \cdots, x_i) + e$,其中估计误差 e 在某种程度下最小. 若找到相应的回归关系,就可以进行预测和控制. 本文主要分析一元线性回归模型.

一元线性回归模型通过观测得到 n 个数据 (x_i, y_i),在此基础上,获得因变量 y 对自变量 x 的回归关系 $y = a + bx + e$,其中 $e \sim N(0, \sigma^2)$,a,b 是回归系数.

一般情况下 n 对 (x_i, y_i) 不完全相等,所以利用最小二乘法来计算 a,b.

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

$$\hat{b} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) / \sum_{i=1}^{n}(x_i - \bar{x})^2$$

其中,$\bar{x}$,$\bar{y}$ 为 x_i,y_i 的平均值. 然后利用获得的 a,b 估计 y_i.

$$d = \sum_{i=1}^{n}\{y_i - (a + bx_i)\}^2$$

当 d 达到最小值时,表示该直线最靠近观测的 n 个数据点.

实际问题中,因变量常常受到不止一个自变量的影响,因此有必要研究多个自变量的回归分析.

当有 m 个自变量时,线性回归模型将变为 $y = a + \sum_{i=1}^{m} b_i x_i + e$,同样采用最小二乘法求出回归系数 a 和 b_i.

2 预测模型的准确性分析

本文采用经典的误差估计算法——LMS算法[7]对数据挖掘趋势预测中的准确度进行分析.

LMS自适应算法步骤如下.

输出:$y = x_1 \hat{w}^H$

估计误差:$e_1 = d_1 - y$

权自适应更新:$\hat{w}_1 = \hat{w} + \mu x e_1$

其中,x_1 为输入,w 为自适应的调整权矢量,e_1

为估计误差,d_1 为期望输出,y 为输出,μ 为迭代步长参数. e_1 越趋近于 0 说明趋势预测越准确,错误率越小.

本文将回归模型与 LMS 算法相结合建立准确度分析反馈预测模型. 通过最小二乘法求出回归系数 a 和 b_i,再根据回归分析模型求出因变量 y_i,最后将 y_i 作为 LMS 算法的输入量 x_1.

3 仿真分析

利用 Matlab 仿真软件对双向数据挖掘的反馈预测模型的预测性和准确度进行仿真分析.

3.1 预测性仿真分析

利用最小二乘法对回归系数进行仿真,结果如图 2 所示.

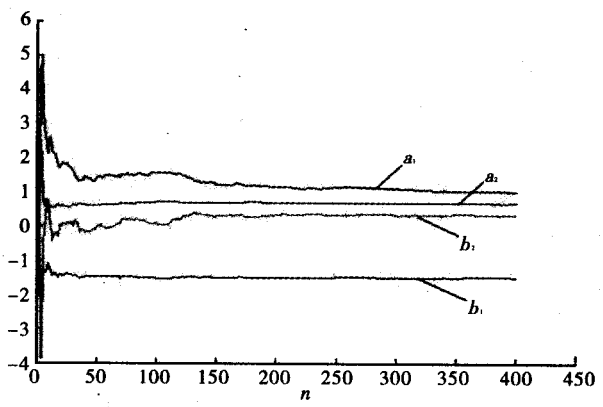

图 2 回归系数仿真结果

图 2 所示的仿真程序中,设定了 2 对回归系数进行仿真分析. 从中可以看出 2 个回归系数之间是息息相关的,即证明用最小二乘法对回归模型中的回归系数进行分析是准确的.

3.2 准确度仿真分析

准确度分析并不只有 LMS 算法这一种检验方法,也可以用校准法进行准确度分析. 但比较这 2 种分析方法,LMS 算法相应的计算量会减少很多,并且 LMS 算法比较经典,可信度更高. 在对准确度进行分析时,选用的是相关性数据和非相关性数据,对相关性数据用 LMS 算法进行仿真分析,其误差估计如图 3 所示.

从图 3 可以看出,选用 LMS 算法对相关性数据进行准确度分析时,e_1 无限接近 0,同时对于非相关性数据分析时,其误差曲线一直为 0. 这说明采用的准确度分析模型是可行的.

为了预测得更准确,可以在用LMS算法进行准确度分析的基础上,再应用标准偏差进行核实. 标

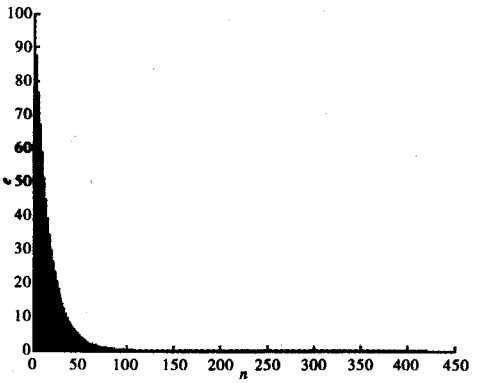

图 3 误差估计

准偏差为

$$s_{xy} = \frac{1}{n}\sum (y_i - \hat{Y}_i)^2$$

其中,s_{xy} 是标准偏差,y_i 是实际值,$\hat{Y}_i$ 是预测值,n 是数据的点数.

然而,自适应算法对系统硬件有一定的要求,选用不同的自适应算法时,对系统配置的要求又不同. 因此,在做相应的预测挖掘时,一定要根据不同的自适应算法和接收数据的来源选择相应的系统配置.

4 结语

本文将关联模型和最小二乘法回归模型结合起来建立了一种反馈预测模型,并且通过 LMS 算法验证了该反馈预测模型的准确度,证实了反馈模型在对数据进行中长期预测时预测结果具有较高的精确度. 此模型可以用于预测火灾、天气等,具有推广价值.

参考文献:

[1] 孟晓东,袁道华,施惠丰. 基于回归模型的数据挖掘研究[J]. 计算机与现代化,2010,4(1).

[2] Sandborn P A, Mauro F, Knox R. A data mining based approach to electronic part obsolescence forecasting[J]. IEEE Trans on Components and Packaging Tech, 2007, 9(3):1521.

[3] 程苗. 关联分析在数据挖掘中的应用[J]. 激光杂志,2007,28(3):65.

[4] 潘庆先,于萍,娄兰芳. 关联规则算法的研究及其在教学评价中的应用[J]. 烟台大学学报:自然科学与工程版,2010,23(2):127.

[5] 徐庆华,刘勇,马履一,等. 长白落叶松苗高生长与气象因子相关关系分析[J]. 林业科技,2010(1):1.

[6] 杨种学. 基于回归技术商品销售趋势预测模型的实现[J]. 保山师专学报,2009,28(5):64.

[7] 张旭东. 离散随机信号处理[M]. 北京:清华大学出版社,2006:151-256.