回归分析数据挖掘技术

李　欣[1, 2]，刘万军[2]

（1. 海军飞行学院，辽宁葫芦岛，125000；2. 辽宁工程技术大学电子与信息工程系，辽宁葫芦岛，125015）

摘　要：统计类数据挖掘技术作为数据挖掘技术中较为成熟的一个分支，它采用统计学的方法与数据挖掘的原理对目标数据进行分析和处理。文中针对其中的回归分析技术的原理及应用进行了探讨。

关键词：数据挖掘；回归分析；线性回归

中图分类号：TP 311.12　　**文献标识码：**A

0 引言

回归分析是统计分析中应用最为广泛的一个分支，它起源于 19 世纪高斯的最小二乘法，20 世纪初形成。回归是研究自变量与因变量之间关系的分析方法，它根据已知的自变量来估计和预测因变量的总平均值。根据回归分析方法得出的数学表达式称为回归方程，它可能是直线，也可能是曲线。在统计中有许多不同类型的回归，但是它们的基本思想都是创建的模型能够匹配预测属性中的值。

1 回归分析数据挖掘技术概述

随着以数据库、数据仓库等数据仓储技术为基础的信息系统在各行各业的应用，使海量数据不断产生，导致数据库中存储的数据量急剧增大。然而，在大量的数据背后隐藏着许多重要信息，如能把这些信息从数据库中抽取出来，将会非常有用，数据挖掘技术就是伴随着这种需求从研究走向应用的[1]。数据挖掘是从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的、但又是有用的信息和知识。

回归分析是通过一个或几个变量的变化去解释另一变量的变化。包括找出自变量与因变量、设定数学模型、检验模型、估计预测等环节。变量之间的关系，有的是确定的函数关系，有的则没有，变量 y 随着变量 x 而变化，但不能由 x 的取值精确求出 y 的值，变量 y 与 x 间的这种关系称为相关关系。回归分析就是研究变量间相关关系的一种数理统计方法。回归分析以对一种变量同其他变量相互关系的过去的观察值为基础，并在某种精确度下预测未知变量的值。

2 线性回归数据挖掘技术

回归中最简单的形式是仅包含一个预测目标和一个预测属性的简单线性回归，称为一元线性回归，也称为单因素回归[2]。简单的线性方程如下：

$$y=a+bx, \tag{1}$$

变量 y 不仅受 x 的影响，还受其他随机因素的影响，因此通过相关图，可以直观地发现各个相关点并不都落在一条直线上，而是在直线上下波动，只呈现线性相关的趋势。因此我们试图在相关图的散点中引出一条模拟的回归直线，以表明两变量 x 与 y 的关系，称为估计回归线。回归方程：

$$y_1=a+bx, \tag{2}$$

式中：y_1 为 y 的估计值；a 为纵轴截距；b 为回归系数，即自变量增加一个单位时因变量的平均增加值。

2.1 回归参数的估计

模型中的参数 a 与 b 在一般情况下都是未知数，必须根据样本数据(x_i，y_i)来估计。确定参数 a 与 b 值的原则是要使得样本的回归直线同观察值的拟合状态最好，即要使得偏差最小[3]。为此，可以采用“最小二乘法”的办法来解决。

对应于每一个 x_i，根据回归直线方程可以求出一个 y_1，它就是 y_i 的一个估计值。估计值和观察值之间的偏差 y_i-y_1。有 n 个观察值就有相应的 n 个偏

收稿日期：2005-10-12
作者简介：李　欣（1980–），女，硕士生.

差。要使模型的拟合状态最好，就是说要使 n 个偏差的总和最小。但为了计算方便起见，我们以误差的平方和最小为标准来确定回归模型[4]。

这就要求

$$Q=\sum_{i=1}^{n}(y_i-y_1)^2=\sum_{i=1}^{n}(y_i-a-bx_i)^2 \qquad (3)$$

是个极小值。

根据微积分中的极值定理，要使式(3)取极值，其对 a 与 b 所求的偏导数应为 0，即

$$\frac{\partial Q}{\partial a}=-2\sum(y_i-a-bx_i)=0\ , \qquad (4)$$

$$\frac{\partial Q}{\partial b}=-2\sum(y_i-a-bx_i)x_i=0\ 。 \qquad (5)$$

经整理后可得：

$$\sum y_i=na+b\sum x_i\ ,\ \sum x_iy_i=a\sum x_i+b\sum x_i^2\ 。$$

解上式，可得：

$$b=\frac{\sum x_iy_i-\frac{1}{n}(\sum x_i)(\sum y_i)}{\sum x_i^2-\frac{1}{n}(\sum x_i)^2}\ ,\ a=\overline{Y}-b\overline{X}\ 。$$

2.2 线性回归数据挖掘举例

从某班 50 名学生中随机抽取 10 名，测得其语文考试成绩与数学考试成绩资料见表 1，试建立该 10 名学生的语文成绩对数学成绩的线性回归模型。

表 1　某班语文与数学考试成绩表

序号	1	2	3	4	5	6	7	8	9	10
语文成绩	54	66	68	76	78	82	85	87	90	94
数学成绩	61	80	62	86	84	76	85	82	88	96

解：设语文成绩为 x，数学成绩为 $\hat{y}$，则设所求线性回归模型为 $\hat{y}=bx+a$ 计算，代入公式求得：

$$b=\frac{10130}{13700}=0.7394\approx 0.74\ ,$$

$$a=\frac{800}{10}-0.74\times\frac{780}{10}=22.28\ 。$$

所求线性回归模型为 $\hat{y}$ =0.74x+22.28。

说明：将自变量 x 的值分别代入上述回归模型中，可得到相应的因变量的估计值，由回归模型知：语文成绩每增加 1 分，数学成绩平均增加 0.74 分。

3 非线性回归数据挖掘技术

上面讨论的变量间的关系是一条直线的情况。在很多的情况下，变量间的关系呈非线性，这时就应该拟合一条曲线来反映变量间的关系[5]。

3.1 非线性回归模型

一元非线性回归的种类很多，如：$y=\beta_0+\beta_1e^x$，$y=\beta_0+\beta_1\ln x$，$y=\beta_0+\beta_1x^2$ 等，y 对自变量 x 都不是线性的，但 y 对参数 β_0 和 β_1 而言是线性的，在这种情况下，我们只需把 e^x、$\ln x$、x^2 等视作变量，用简单的代换就可将上述模型化为线性模型 $y'=\beta_0+\beta_1x'$，其中：x' 可分别为 e^x、$\ln x$ 和 x^2 等。

对于可化为线性模型的回归问题，一般先将其化为线性模型，然后用最小二乘法求出参数的估计值，最后再经过适当的变换，得到所求回归曲线[6]。

3.2 非线性回归数据挖掘举例

某超市的年商品销售额和商品流通费率的对应关系见表 2。试建立回归方程，并预测当计划销售额达到 32.5 万元时，商品的流通率可降低多少？

表 2　某超市年商品销售额和商品流通率对照表

序号	1	2	3	4	5	6	7	8	9
销售额	1.5	4.5	7.5	10.5	13.5	16.5	19.5	22.5	25.5
商品流通费率/%	7.0	4.8	3.6	3.1	2.7	2.5	2.4	2.3	2.2

经过对上述数据的分析得对应关系图，见图 1。

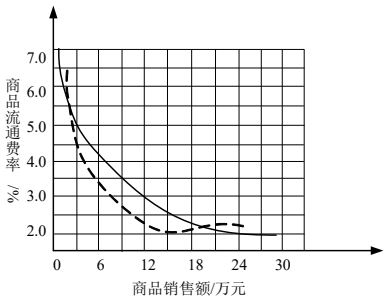


图 1　某超市年商品销售额和商品流通率对应关系图

可以看出，商品的流通率与商品的销售额间的关系呈双曲线形状，用双曲线的回归方程来描述它们之间的规律，并根据它来预测未来的情况。

$$\hat{y}=b\frac{1}{x}+a$$

对于上面的双曲线，令 Z=1/X，则上述的方程可化为：$\hat{y}=bx+a$，即得到双曲线的回归参数 a，b 的表达式。根据上面的数据，进行变换后带入方程解得：a=2.2258，b=7.6196，则双曲线的回归方程为：$\hat{y}=2.2258+7.6196\frac{1}{x}$。

因此，利用这个双曲线的回归方程就可预测 y 的值，如该超市的计划销售额达到 32.5 万元，商品

流通率可降低到：

$$\hat{y} = 2.2258 + 7.6196 \times \frac{1}{32.5} = 2.46\% \text{。}$$

4 结束语

统计作为数据挖掘的一种关键技术，应用非常广泛，具有相关关系的变量间虽然不具有确定的函数关系，但是通过大量的观测数据，可以发现它们之间存在一定的统计规律，数理统计中研究这些统计规律或者说研究变量之间相关关系的方法就是所谓的回归分析。它能帮助我们有效地从一个可以控制或可以精确观察的变量取得的值去估计另一随机变量所取的值。

参考文献：

[1] 陈文伟. 数据挖掘技术[M]. 北京：北京工业大学出版社, 2002:71-78

[2] 周根贵. 数据仓库与数据挖掘[M]. 杭州：浙江大学出版社, 2001:156-160

[3] Margaret H Dunham. 数据挖掘教程[M]. 郭崇慧等译. 北京：清华大学出版社, 2005:67-73

[4] 张云涛，龚玲. 数据挖掘技术与原理[M]. 北京：电子工业出版社, 2004:69-76

[5] Jiawei Han, Micheline Kamber. Internet/Web 数据挖掘概念与技术[M]. 范明，等译. 北京：机械工业出版社, 2001:132-141

[6] 康小东. 基于数据仓库的数据挖掘技术[M]. 北京：机械工业出版社, 2004:179-197

Technology of regression data mining

LI Xin[1, 2]，LIU Wanjun[2]

（1．Naval Flying Academy，Huludao，Liaoning，125000；

2．Department of Electrical and Information Engineering，

Liaoning Engineering and Technology University，Huludao，Liaoning，125015）

Abstract: The technology of statistical data mining has become the most matures technology of data mining. It mingles the method of statistics with the principle of data mining and deals with the data with it. This paper discusses aiming at the principle and application of regression technology.

Key words: data mining; regression; liner regression