多元线性回归模型的聚类分析方法研究

Huiwen Wang, Ming Ye, Gilbert Saporta

多元线性回归模型的聚类分析方法研究

王惠文[1]，叶 明[1]，Gilbert Saporta[2]
（1.北京航空航天大学经济管理学院 北京 100083；2.Department Statistiques, Conservatoire National Des Arts et Métier，France）

摘 要：提出一种对大量的多元线性回归模型进行聚类分析的方法。首先利用增广矩阵的相关系数矩阵定义了 2 个多元回归模型之间的距离以及模型集合的质心和半径等相关概念。然后采用 Squeezer 聚类方法，以过程全自动化的方式，实现对多元线性回归模型集合进行聚类分析。通过仿真研究验证了方法的有效性，取得满意的分析结果。

关键词：多元线性回归模型集合；回归模型之间的距离；模型聚类；聚类分析
中图分类号：O212.4 文献标识码：A 文章编号：1004-731X (2009) 22-7048-03

Classification for Multiple Linear Regression Methods

WANG Hui-wen[1], *YE Ming*[1], *Gilbert Saporta*[2]
(1. Beijing University of Aeronautics and Astronautics, School of Economics & Management, Beijing 100083, China;
2. Department Statistiques, Conservatoire National Des Arts et Métier, France)

Abstract: *A cluster analysis method on massive multiple linear regression models was proposed.* Firstly, the concepts such as distance between two multiple regression models, the centroid and radius of multiple linear regression model set were defined by using the correlation coefficient matrix of augmented matrix. Then Squeezer cluster method was applied to realize the cluster analysis on multivariate linear regression models based on entire automation of the process. The simulation case confirms validity of this method and lead to satisfactory results.

Key words: multiple linear regression models set; the distance between regression models; Models clustering; cluster analysis

引 言

随着信息技术的快速发展，计算机收集和存储巨量数据的能力越来越强。于是在经济、金融和管理等研究领域中，人们开始关注如何能够快速有效地建立成千上万个多元回归模型的问题。例如，同时对几百个地区的经济指标进行回归建模，或对上千种零备件的需求进行预测分析，等等。目前在应用中，如果遇到诸如此类的回归建模问题，常用的方法还是逐一建立模型。这样做固然能比较准确地得到每个模型结果，但是建模的工作量却变得十分庞大，并且经常需要持续很长的时间。为了实现快速有效地建立大批量的回归模型，本文提出一种多元线性回归模型的聚类分析方法，该方法将成为大规模自动化建模过程中的一种十分有效的前期处理技术。

一般说来，聚类分析的工作目的是降低因规模造成的复杂程度。常用的聚类分析方法包括系统聚类法，划分方法，K-均值方法、模糊 C 均值方法、基于密度的方法、基于人工神经网格的方法[1]，等等。近年来，随着数据挖掘研究的不断深入，又涌现了大量新的聚类方法，诸如 ROCK 算法[2]，C^2P 算法[3]，DBSCAN 算法[4]，BIRTH 算法[5]，CUBE 算法[6]，CHAMELEON 算法[7]，WaveCluster 算法[8]和 CLIQUE 算法[9]等。本文所采用的 Squeezer 算法[10](2002，He Zengyou 等)是一种改善的 K-均值方法，其特点是分析人员只要设定一个阈值便可以快速的把每个参加聚类的样本点自动划归到某一个类别当中。与普通的 K-均值方法相比，该方法不需要事先人为指定分类的类别数量，从而实现了聚类过程的自动化。

为了解决多元线性回归模型集合的聚类分析问题，本文首先给出 2 个多元线性回归模型之间的距离定义；之后，再利用 Squeezer 聚类算法对众多的多元线性回归模型进行聚类分析。而这一模型聚类方法可以在保证过程完全自动化的前提下，减少需要建立模型的种类，从而大幅度降低回归建模的工作量。

论文结构如下：第二部分给出 2 个多元线性回归模型之间的距离定义。第三部分介绍具体的多元线性回归模型的分类方法。第四部分为仿真研究。第五部分为总结。

1 2 个多元线性回归模型之间的距离定义

记某个 p 元线性回归模型为 M，它的自变量为 $x_1, x_2, \cdots, x_p$，因变量为 y，并设这些变量都是标准化的（即均值为 0，方差为 1）。设自变量和因变量之间存在下面的线性回归关系：

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, 2, \cdots, n \tag{1}$$

将自变量和因变量的观测值写成矩阵形式，记

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

收稿日期：2008-07-31 修回日期：2008-09-25
基金项目：多元回归模型评价，模型分类以及模型预测理论研究及其应用（70771004）
作者简介：王惠文(1957)，女，辽宁，博士，教授，博导，研究方向为复杂数据分析；叶明(1983)，男，安徽，博士生，研究方向为复杂数据分析，数据挖掘。

则增广矩阵 (X,Y) 的相关系数矩阵为

$$V=\frac{1}{n-1}(X,Y)^T(X,Y)=\frac{1}{n-1}\begin{pmatrix}X^TX & X^TY\\ Y^TX & Y^TY\end{pmatrix} \tag{2}$$

从已有的研究结论可知，多元线性回归模型的回归系数 $(\hat{\beta}_1,\hat{\beta}_2,\cdots,\hat{\beta}_p)$，以及模型的评估参数 R^2、F—检验值以及 t—检验值等均可由式(2)中的元素运算得到[11]。由此可见，可以通过比较矩阵 V 的差异来对不同的多元线性模型进行聚类分析。于是可以给出 2 个多元线性回归模型之间距离的定义。

定义 1：记有 2 个 p 元线性回归模型为 M_1,M_2，它们的观测点的数量分别是 n_1,n_2。记 M_1,M_2 所对应的增广矩阵的相关系数矩阵分别为 V_1 与 V_2；再记 $\Delta V=V_1-V_2=(\Delta v_{ij})\in\boldsymbol{R}^{(p+1)\times(p+1)}$。定义模型 M_1与M_2 的距离为矩阵 ΔV 的契比雪夫范数：

$$D(M_1,M_2)=\|V_1-V_2\|_{V_\infty}=\|\Delta V\|_{V_\infty}=\max_{i,j}|\Delta v_{ij}| \tag{3}$$

本文之所以选择契比雪夫范数作为矩阵之间距离的定义，是为了保证矩阵 V_1 与 V_2 中的元素尽可能的接近。而通过相关系数矩阵 V_1 与 V_2 来定义模型之间距离，可以保证在数据经过标准化处理后（即所有变量的均值为 0，方差为 1），两个多元线性回归模型 M_1,M_2 的回归系数以及 R^2、F—检验值与 t—检验值等都尽可能的相象。

在模型聚类的过程中，为了衡量一个模型到某个模型集合的距离，需要定义模型集合的 “质心”概念。但是在模型集合中，由于各个模型之间的观测点数量可能会存在差异，因此很难用平均值的办法来计算模型集合的“质心”。本文采用在模型集合中与其它模型距离总和最小的模型来表示该模型集合的“质心”，下面给出模型集合“质心”的定义。

定义 2：记多元线性回归模型集合 $\Lambda=(M_1,M_2,M_3,\cdots,M_k)$，该模型集合的质心 M 由公式(4)确定：

$$M=\min_{i=1,2,\ldots,k}\sum_{j=1}^{k}D(M_i,M_j) \tag{4}$$

由此可见，计算一个模型到某个模型集合的距离就可以转化为计算它到这个模型集合“质心”的距离。

为了使聚类的结果达到一定的精度，应该把模型集合中的所有模型控制在一定的精度范围之内。下面给出模型集合“半径”的定义。

定义 3：记多元线性回归模型集合 $\Lambda=(M_1,M_2,M_3,\cdots,M_k)$，质心为 M，该模型集合的半径由公式(5)确定：

$$R=\max_{i=1,2,\ldots,k}D(M_i,M) \tag{5}$$

2 多元线性回归模型的聚类分析方法

本节介绍一种基于 Squeezer 算法的多元线性回归模型的聚类分析方法。该算法的计算过程是全自动化的，分析人员只需要根据实际工作的需要设定一个精度阈值 T，则算法会自动将距离小于阈值的对象聚为一类。在算法执行的过程中，T 也设定为模型集合质心止动阈值。即当模型集合半径 $R>T$ 时，模型集合的质心便停止变动。该算法具体描述如下。

对于多元线性回归模型的集合 $\Lambda=(M_1,M_2,M_3,\cdots,M_N)$，设定聚类的精度阈值为 T。

(1) 对于 $i=1,2,\cdots,N$，重复下列步骤(2)～(8)；

(2) 从模型集合 Λ 中任选中一个模型记为 M_i

(3) 如果 $i=1$，将 M_i 单独归为一类，并令 $\Lambda=\Lambda-M_i$，执行步骤(2)；否则直接执行步骤(4)；

(4) 计算 M_i 到各已完成聚类的模型集合质心之间的距离，从中选出与 M_i 距离最小的模型集合，记该模型集合为 Λ_j，两者之间的距离记为 D_{ij}，执行步骤(5)；

(5) 如果 $D_{ij}<T$，将 M_i 归为模型集合 Λ_j，并令 $\Lambda=\Lambda-M_i$；执行步骤(6)，否则执行步骤(8)；

(6) 如果 Λ_j 质心未固定，调整 Λ_j 的质心和半径，新的半径记为 $R(\Lambda_j)$，执行步骤(7)。否则执行步骤(2)。

(7) 如果 $R(\Lambda_j)>T$，将 Λ_j 质心固定，执行步骤(2)。否则直接执行步骤(2)。

(8) 将 M_i 单独归为一类，并令 $\Lambda=\Lambda-M_i$。执行步骤(2)；

(9) 最后重新调整各模型集合的质心

(10) 算法结束。

3 仿真研究

本节将通过仿真研究说明所提的模型聚类算法的有效性。为此，首先采用随机数发生器，按照 $x_1\sim N(5,10)$, $x_2\sim N(10,20),x_3\sim N(5,5),\zeta\sim N(0,10)$ 生成一组仿真数据：x_1,x_2,x_3，ζ；其中，每个变量都生成 50 个随机数。这样，就得到 数据($x_{i,1}$, $x_{i,2}$, $x_{i,3}$, ζ_i)，i=1,2,…,50。然后，对数据进行如下处理：

1) 对变量 x_1,x_2,x_3 及 ζ 做中心化处理。

2) 利用 Gram-Schmidt 过程[12]将变量 x_1,x_2,x_3 及 ζ 变换成一组正交变量，由此得到一个与变量 x_1,x_2,x_3 垂直的向量 ξ。

由于对数据进行了中心化处理，并且 ξ 垂直于由 x_1,x_2,x_3 生成的空间，所以 ξ 与 x_1,x_2,x_3 的相关系数均等于 0。因此，可以用 ξ 来表示与 x_1,x_2,x_3 无关的随机误差项。

下面利用 x_1,x_2,x_3 以及 ξ 来构造 3 种不同类型的基础模型，构造方法如表 1 所示。

表 **1** 基础模型构造方法

模型名称	自变量	因变量	模型表达式	因变量 y 的计算方法
模型 1	x_1,x_2,x_3	y_1	$y_1=x_1+x_2+x_3$	$y_1=x_1+x_2+x_3+\xi$
模型 2	x_1,x_2,x_3	y_2	$y_2=2x_1+2x_2+x_3$	$y_2=2x_1+2x_2+x_3+\xi$
模型 3	x_1,x_2,x_3	y_3	$y_3=x_1+x_2+x_3$	$y_3=x_1+x_2+x_3+10\xi$

在表 1 中，模型 1、模型 2 和模型 3 分别代表 3 类不同的模型。首先，模型 1、模型 3 与模型 2 相比，具有不同的模型参数，因此它们与模型 2 不属于同类的总体模型；另一方面，虽然模型 1 与模型 3 的模型参数是一致的，但是由于这 2 个模型的随机误差项相差非常大，因此也成为 2 类不同的模型。

根据上述 3 个基础模型，可以通过仿真方法衍生出若干

新的模型，生成这些新模型的基本思路是：对于同样的基础

表 2 衍生模型的生成方式

模型类别	模型名称	自变量	因变量	各实际变量的生成方式
模型组 1	基础模型 1	x_1, x_2, x_3	y_1	$y_1 = x_1 + x_2 + x_3 + \xi$
	模型 1.1	$x_{11}^1, x_{12}^1, x_{13}^1$	y_1^1	$y_1^1 = y_1 + \eta;\ x_{11}^1 = 10x_1 + \eta$
				$x_{12}^1 = 20x_2 + \eta;\ x_{13}^1 = 30x_3 + \eta$
	模型 1.2	$x_{11}^2, x_{12}^2, x_{13}^2$	y_1^2	$y_1^2 = y_1 + \eta;\ x_{11}^2 = 0.1x_1 + \eta$
				$x_{12}^2 = 0.2x_2 + \eta;\ x_{13}^2 = 0.3x_3 + \eta$
	模型 1.3	$x_{11}^3, x_{12}^3, x_{13}^3$	y_1^3	$y_1^3 = 10y_1 + \eta;\ x_{11}^3 = 0.1x_1 + \eta$
				$x_{12}^3 = 0.2x_2 + \eta;\ x_{13}^3 = 0.3x_3 + \eta$
	模型 1.4	$x_{11}^4, x_{12}^4, x_{13}^4$	y_1^4	$y_1^4 = 0.1y_1 + \eta;\ x_{11}^4 = 10x_1 + \eta$
				$x_{12}^4 = 20x_2 + \eta;\ x_{13}^4 = 30x_3 + \eta$
模型组 2	基础模型 2	x_1, x_2, x_3	y_2	$y_2 = 2x_1 + 2x_2 + x_3 + \xi$
	模型 2.1	$x_{21}^1, x_{22}^1, x_{23}^1$	y_2^1	$y_2^1 = y_2 + \eta;\ x_{21}^1 = 10 \times 2x_1 + \eta$
				$x_{22}^1 = 20 \times 2x_2 + \eta;\ x_{23}^1 = 30x_3 + \eta$
	模型 2.2	$x_{21}^2, x_{22}^2, x_{23}^2$	y_2^2	$y_2^2 = y_2 + \eta;\ x_{21}^2 = 0.1 \times 2x_1 + \eta$
				$x_{22}^2 = 0.2 \times 2x_2 + \eta;\ x_{23}^2 = 0.3x_3 + \eta$
	模型 2.3	$x_{21}^3, x_{22}^3, x_{23}^3$	y_2^3	$y_2^3 = 10y_1 + \eta;\ x_{21}^3 = 0.1 \times 2x_1 + \eta$
				$x_{22}^3 = 0.2 \times 2x_2 + \eta;\ x_{23}^3 = 0.3x_3 + \eta$
	模型 2.4	$x_{21}^4, x_{22}^4, x_{23}^4$	y_2^4	$y_2^4 = 0.1y_1 + \eta;\ x_{21}^4 = 10 \times 2x_1 + \eta$
				$x_{22}^4 = 20 \times 2x_2 + \eta;\ x_{23}^4 = 30x_3 + \eta$
模型组 3	基础模型 3	x_1, x_2, x_3	y_3	$y_3 = x_1 + x_2 + x_3 + 10\xi$
	模型 3.1	$x_{31}^1, x_{32}^1, x_{33}^1$	y_3^1	$y_3^1 = y_3 + \eta;\ x_{31}^1 = 10x_1 + \eta$
				$x_{32}^1 = 20x_2 + \eta;\ x_{33}^1 = 30x_3 + \eta$
	模型 3.2	$x_{31}^2, x_{32}^2, x_{33}^2$	y_3^2	$y_3^2 = y_3 + \eta;\ x_{31}^2 = 0.1x_1 + \eta$
				$x_{32}^2 = 0.2x_2 + \eta;\ x_{33}^2 = 0.3x_3 + \eta$
	模型 3.3	$x_{31}^3, x_{32}^3, x_{33}^3$	y_3^3	$y_3^3 = 10y_3 + \eta;\ x_{31}^3 = 0.1x_1 + \eta$
				$x_{32}^3 = 0.2x_2 + \eta;\ x_{33}^3 = 0.3x_3 + \eta$
	模型 3.4	$x_{31}^4, x_{32}^4, x_{33}^4$	y_3^4	$y_3^4 = 0.1y_3 + \eta;\ x_{31}^4 = 10x_1 + \eta$
				$x_{32}^4 = 20x_2 + \eta;\ x_{33}^4 = 30x_3 + \eta$

在模型聚类算法的执行过程中，分类的精度阈值设为 0.05。通过计算，模型的最终分类结果见表 3、表 4 和表 5。在这些表中，分别给出了 3 类模型的聚类结果，以及标准化数据的回归模型及其相关参数。

从表 3～表 5 可以看到，每类模型中的各种参数值都是比较接近的。该聚类方法合理地将仿真实验中设定的同类模型都划分到其所属的类型中，达到了预期的分析效果。

表 3 第 1 类模型的分类结果与相关参数

聚类结果	标准化数据的回归模型	R^2	F value
模型 1	$y_1 = 0.484x_1 + 0.448x_2 + 0.460x_3$	0.749	45.836
模型 1.1	$y_1^1 = 0.483x_{11}^1 + 0.449x_{12}^1 + 0.459x_{13}^1$	0.749	45.756
模型 1.2	$y_1^2 = 0.483x_{11}^2 + 0.453x_{12}^2 + 0.449x_{13}^2$	0.748	45.566
模型 1.3	$y_1^3 = 0.474x_{11}^3 + 0.441x_{12}^3 + 0.467x_{13}^3$	0.746	45.115
模型 1.4	$y_1^4 = 0.483x_{11}^4 + 0.448x_{12}^4 + 0.458x_{13}^4$	0.746	45.142

表 4 第 2 类模型的分类结果与相关参数

聚类结果	标准化数据的回归模型	R^2	F value
模型 2	$y_2 = 0.422x_1 + 0.388x_2 + 0.404x_3$	0.571	20.390
模型 2.1	$y_2^1 = 0.423x_{21}^1 + 0.387x_{22}^1 + 0.404x_{23}^1$	0.570	20.419
模型 2.2	$y_2^2 = 0.414x_{21}^2 + 0.380x_{22}^2 + 0.411x_{23}^2$	0.569	20.202
模型 2.3	$y_2^3 = 0.424x_{21}^3 + 0.390x_{22}^3 + 0.403x_{23}^3$	0.571	20.435
模型 2.4	$y_2^4 = 0.428x_{21}^4 + 0.386x_{22}^4 + 0.400x_{23}^4$	0.575	20.418

表 5 第 3 类模型的分类结果与相关参数

聚类结果	标准化数据的回归模型	R^2	F value
模型 3	$y_3 = 0.422x_1 + 0.388x_2 + 0.404x_3$	0.571	20.390
模型 3.1	$y_3^1 = 0.423x_{31}^1 + 0.387x_{32}^1 + 0.404x_{33}^1$	0.570	20.419
模型 3.2	$y_3^2 = 0.414x_{31}^2 + 0.380x_{32}^2 + 0.411x_{33}^2$	0.569	20.202
模型 3.3	$y_3^3 = 0.424x_{31}^3 + 0.390x_{32}^3 + 0.403x_{33}^3$	0.571	20.435
模型 3.4	$y_3^4 = 0.428x_{31}^4 + 0.386x_{32}^4 + 0.400x_{33}^4$	0.575	20.418

特别值得注意的是，第 3 类模型的回归系数与第 1 类模型比较相近，但是该类模型的 R^2 均明显低于第 1 类模型。这也是和基础模型中关于“第 3 类模型的随机误差更大”的假设是一致的。

下面，进一步说明一类模型的质心对同类其它模型的代表性。由于文章篇幅有限，本文仅以第 1 类模型为例。根据本文第 2 节的定义 2 可计算出，第 1 类模型的质心为模型 1.1，其标准化回归模型表达式为 $y = 0.483x_1 + 0.449x_2 + 0.459x_3$。下面将第 1 类其它模型的标准化数据代入该表达式，并计算其因变量 y 的估计标准误差 $s_e = \sqrt{\dfrac{1}{n-4}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$，具体结果如表 6 所示。

表 6 各模型回归标准误差

模型名称	原模型的估计标准误差	使用质心模型的估计标准误差
模型 1	0.516	0.517
模型 1.2	0.516	0.518
模型 1.3	0.517	0.520
模型 1.4	0.518	0.519

从表 6 可以看到，若将同一模型类别中的其它模型的数据代入其质心模型的表达式中，其估计标准误差与原模型相差不大，这说明在同一模型类中，质心模型可以较好地代表该模型类别中的其它模型。这样，今后在对同一模型类别进行建模的工作中，就不必逐一建立回归模型，而只须选取该类模型的质心模型即可以很好的代表该类，这样便能大大降低建模的工作量，提高建模的工作效率。

4 结论

本文提出了一种对多元线性回归模型进行聚类分析的方法。为此，文中首先给出 2 个多元线性回归模型之间的距离定义以及模型集合的质心定义；之后，采用 Squeezer 聚类算法对多元线性回归模型进行聚类分析。这种回归模型的聚类方法仅须事先指定精度阈值，就可以快速完成模型的自动分类，从而在建立大批量的回归模型的工作中，为减少回归

（下转第 7056 页）

相比较，验证了基于仿真的优化方法的可行性和有效性，有助于供应链管理者做出正确的库存控制策略，对如何改善供应链管理具有一定的指导意义。但对于带有价格波动或突发事件等随机因素的多级网状随机性库存系统的库存控制策略优化问题，还有待进一步研究。

参考文献：

[1] 蒋长兵, 吴承健. 现代物流理论与供应链管理实践[M]. 杭州: 浙江大学出版社, 2006: 251-262.

[2] Clark A J, Scarf H. Optimal Policies of a Multi-echelon Inventory Problem [J]. Management Science (S0025-1909), 1960, 6(4): 475-490.

[3] Deuermeyer B L, Schwarz L B. A Model for the Analysis of System Service Level in Warehouse-retailer Distribution Systems: The Identical Retailer Case [C]// Schwarz L B. Studies in Management Science: Multi-level Production /Inventory Control Systems, Amsterdam, North-Holland, The Netherlands, 1981. 163-193.

[4] Bashyam S, Fu M C. Optimization of (s,S) Inventory Systems with Random Lead Time and a Service Level Constrain [J]. Management Science (S0025-1909), 1998, 44(12): 243-256.

[5] Hopp W J, Spearman M L, Zhang R Q. Easily Implementable Inventory Control Policies [J]. Operations Research (S0030-364X), 1997, 45(3): 327-340.

[6] 任常锐, 柴跃廷. 供需链仿真技术的发展现状与趋势[J]. 计算机集成制造系统, 2004, 10(2): 121-126.

[7] Arnold J, Köchel P. Evolutionary Optimization of a Multi-location Inventory Model with Lateral Transshipments [C]// Proceedings of Ninth International Working Seminar on Production Economics. Lynkoping: University of Lynkoping, 1996: 401-412.

[8] Köchel P, Nieländer U. Simulation-based Optimisation of Multi-echelon Inventory Systems [J]. International Journal of Production Economics (S0925-5273), 2005, 93-94: 505-513.

[9] Lee Y H, Kim S H. Optimal Production-distribution Planning in Supply Chain Management Using a Hybrid Simulation-analytic Approach [C]// Proceedings of The 2000 Winter Simulation Conference. Piscataway, USA: IEEE Press, 2000: 1252-1259.

[10] Ding H W, Benyoucef L, Xie X L. A Simulation-based Optimization Method for Production-distribution Network Design [C]// Proceedings of 2004 IEEE International Conference on Systems, Man and Cybernetics. Piscataway, USA: IEEE Press, 2004: 4521-4526.

[11] 田俊峰, 杨梅. 随机需求条件下生产-库存系统优化与仿真[J]. 系统仿真学报, 2004, 16(11): 2522-2524.

[12] Lee L H, Chew E P, Teng S, Chen Y. Multi-objective Simulation-based Evolutionary Algorithm for an Aircraft Spare Parts Allocation Problem [J]. European Journal of Operational Research (S0377-2217), 2008, 189(2): 476-491.

[13] Kämpf M, Köchel P. Simulation-based Sequencing and Lot Size Optimisation for a Production-and-inventory System with Multiple Items [J]. International Journal of Production Economics (S0925-5273), 2006, 104(1): 191-200.

[14] Arakawa M, Fuyuki M, Inoue I. An Optimization-oriented Method for Simulation-based Job Shop Scheduling Incorporating Capacity Adjustment Function [J]. International Journal of Production Economics (S0925-5273), 2003, 85(3): 359-369.

[15] Adachi J, Gupta A. Simulation-based Parametric Optimization for Long-term Asset Allocation Using Behavioral Utilities [J]. Applied Mathematical Modeling (S0307-904X), 2005, 29(4): 309-320.

[16] 姜昌华, 戴树贵. 基于遗传算法的随机性 (Q,r) 库存系统仿真优化 [J]. 计算机应用, 2006, 26(1): 184-187.

[17] Shi Y, Eberhart R C. A Modified Particle Swarm Optimizer [C]// IEEE World Congress on Computational Intelligence. Piscataway, USA: IEEE Press, 1998: 69-73.

[18] Parsopoulos K E, Vrahatis M N. Recent Approaches to Global Optimization Problems through Particle Swarm Optimization [J]. National Computing (S1567-7818), 2002, 1(2-3): 235-306.

(上接第 7050 页)

模型的种类，大幅度降低建模工作量，提供了一种十分有效的处理技术。文中通过仿真分析表明，该方法可以有效地对诸多线性回归模型进行类别划分，并说明这种模型聚类分析方法在变量取不同模长或数据存在微小随机扰动的情况下，也都是有效的，因而进一步指出了本文所提方法的应用特点。

参考文献：

[1] 吴文丽, 刘玉树, 赵基海. 一种新的混合聚类算法[J]. 系统仿真学报, 2007, 19(1): 16-18.

[2] Sudipto Guha, Rajeev Rastogi, Kyusemk Shim. ROCK: A robust clustering algorithm for categorical attributes [C]// Proc. 1999 Int. Conf. Data Engineering, Sydney, Australia, Mar, 1999: 512-521.

[3] Alexandros Nanopoulos, Yannis Theodoridis, Yannis Manolopoulos. C2P: Clustering based on closest pairs [C]// Proc. 27th Int Conf. Very Large Database, Rome, Italy, September, 2001: 331-340.

[4] Ester M, Kriegel H P, Sander J Xu X. A density-based algorithm for discovering clusters in large spatial databases [C]// Proc. 1996 Int. Conf. Knowledge Discovery and Data Mining (KDD'96), Portland, Oregon, USA, Aug, 1996: 226-231.

[5] Zhang T, Ramakrishnan R, Livny M. BIRTH: An efficient data clustering method for very large database [C]// Proc. The ACM-SIGMOD Int. Conf. Management of Data, Montreal, Quebec, Canada, June, 1996. USA: ACM, 1996: 103-114.

[6] Sudipto Guha, Rajeev Rastogi, Kyuseok Shim. CURE: A clustering algorithm for large database [C]// Proc. The ACM-SIGMOD Int. Conf. Management of Data, Seattle, Washington, USA, June, 1998. USA: ACM, 1998: 73-84.

[7] Karypis G, Han E-H, Kumar V. CHAMELEON: A hierarchical clustering algorithm using dynamic modeling [J]. IEEE Computer (S0018-9162), 1999, 32(8): 68-75.

[8] Sheikholeslami G, Chatterjee S, Zhang A. Wave Cluster: A multi-resolution clustering approach for very large spatial databases [C]// Proc. 1998 Int. Conf. Very Large Databases, New York, USA, August, 1998: 428-439.

[9] Agrawal R, Gehrike J, Gunopulos D, Raghavan P. Automatic subspace clustering of high dimensional data for data mining applications [C]// Proc. The 1998 ACM-SIGMOD Int. Conf. Management of Data, Washington, USA, June, 1998. USA: ACM, 1998: 94-105.

[10] HE Zeng-you, XU Xiao-fei, DENG Sheng-chun. Squeezer. An efficient algorithm for clustering categorical data [J]. Journal of Computer Science and Technology (S1000-9000), 2002, 17(5): 611-624.

[11] 王惠文, 孟洁. 多元线性回归的预测建模方法[J]. 北京航空航天大学学报, 2007, 33(4): 501-504.

[12] S K Jain, A D Gunawarddena. Linear Algebra: An Interactive Approach [M]. United States of America: Brooks/Cole, 2004: 75-85.