

# 基于马尔科夫模型的回归研究及其应用

何成刚<sup>1,2</sup>, 丁宏强<sup>3</sup>, 陈思宝<sup>1,2</sup>, 罗 斌<sup>1</sup>, 王家鑫<sup>1</sup>

(1. 安徽大学 计算机科学与技术学院, 安徽 合肥 230031;

2. 安徽大学 计算智能与信号处理教育部重点实验室, 安徽 合肥 230039

3. 美国德州大学阿灵顿分校 计算机科学与工程系, 美国 阿灵顿 TX76019)

**摘 要:**在国内外回归分析方法的研究中,神经网络、支持向量机等传统方法被广泛使用,但是由于其计算量太大而且对计算模型和数据的准确性要求很高,在实际的应用中局限性强。为了解决这些难题,对 Markov 理论和相关模型进行了深入的研究。首先将多元回归和 Markov 模型进行结合,提出了基于多元回归的 Markov 模型,解决了转移矩阵难以确定的问题,并将其应用于国民收入预测中,减少了运算复杂度并且解决了实际应用中的局限性,提高了模型的鲁棒性。同时将 Markov 模型和 Regime Switching Model 进行结合,提出了基于 Markov-switch 的回归算法,使用状态转移矩阵来处理数据,实验结果表明该算法可以有效地提高预测效率和大幅度减少运算时间,并且在 UCI 数据集上进行验证和传统方法相比,标准差减少 72.72%、相关系数提高 2%、运行时间减少了 50%。

**关键词:**Markov 模型;多元回归;Markov-switch 回归算法;减少运算量;缩短运算时间

**中图分类号:**TP391.4

**文献标识码:**A

**文章编号:**1673-629X(2022)04-0008-07

**doi:**10.3969/j.issn.1673-629X.2022.04.002

## Regression Research and Application Based on Markov Model

HE Cheng-gang<sup>1,2</sup>, Chris H. Q. DING<sup>3</sup>, CHEN Si-bao<sup>1,2</sup>, LUO Bin<sup>1</sup>, WANG Jia-xin<sup>1</sup>

(1. School of Computer Science and Technology, Anhui University, Hefei 230031, China;

2. Key Lab of Intelligent Computing and Signal Processing of Ministry of Education, Hefei 230039, China;

3. Department of Computer Science and Engineering, University of Texas at Arlington, Arlington TX76019, USA)

**Abstract:**In the research of regression method at home and abroad, traditional methods such as neural network and support vector machine are widely used. However, due to its large amount of calculation and high requirements on the accuracy of the calculation model and data, it has strong limitations in application. Combined multiple regression with Markov model, we propose a Markov model based on multiple regression, which solves the problem that the transfer matrix is difficult to determine, and apply it to national income prediction, which reduces the computational complexity and solves the limitation in application, and improves the robustness of the model. At the same time, the Markov model and Regime Switching Model are combined, and a regression algorithm based on Markov-switch is proposed. The experiment shows that the proposed algorithm can effectively improve the prediction efficiency and greatly shorten the calculation time. It is verified on the UCI data set and compared with traditional methods, the standard deviation is reduced by 72.72%, the correlation coefficient is increased by 2%, and the running time is reduced by 50%.

**Key words:**Markov model; multiply regression; Markov-switch regression model; reducing calculation; shortening calculation time

## 0 引言

回归分析的研究,一直是机器学习研究领域的热点,它能根据历史数据的特点,拟合出回归模型进行准确的预测,广泛应用在实际问题处理中。回归分析<sup>[1]</sup>的方法,就是对大量的数据进行相关的统计处理,通过

寻求恰当的模型来探索出这些变量的内在关系,构造具体的回归模型,然后再根据相关的数据指标来对回归的效果进行评价。通过分析评价的结果得到较好的回归模型,从而运用回归模型来进一步有效的进行预测的研究与应用。神经网络理论、支持向量机

收稿日期:2021-05-29

修回日期:2021-09-29

**基金项目:**国家自然科学基金资助项目(61976004,61572030,61671018);国家自然科学基金面上项目(61673020);国际(地区)合作与交流重点项目(61860206004);安徽大学科学研究建设经费(Y040418282)

**作者简介:**何成刚(1984-),男,博士,高级工程师,通讯作者,研究方向为机器学习、软件工程、数学建模等;丁宏强,博士,教授,博导,研究方向为机器学习等;陈思宝,博士,教授,博导,研究方向为机器学习;罗 斌,博士,教授,博导,研究方向为模式识别等。

(support vector machine, SVM)等这些人工智能(机器学习)的方法被用于回归分析之中,形成了一些回归方法,如 SVM 回归<sup>[2]</sup>、广义回归神经网络(generalized regression neural network, GRNN)<sup>[3]</sup>,但是这些方法计算量太大,对计算模型和数据的准确性要求很高,局限性很强。近些年来,随着马尔科夫理论的进一步发展,其在语音的识别<sup>[2]</sup>、金融序列分析<sup>[3]</sup>等方面取得了良好的效果。该文期望通过将马尔科夫理论与回归分析方法相结合,得到更好的基于马尔科夫理论的回归模型来进行回归预测工作的研究与应用。

主要贡献如下:

(1)将多元回归和 Markov 模型相结合,使用多元回归的方法解决 Markov 模型中转移矩阵难以确定的痛点,提出了基于多元回归的 Markov 回归模型,并将其应用到国民收入的预测之中,取得了很好的效果。

(2)将 Regime Switching Model<sup>[4-5]</sup>(体制转换模型)和 Markov 模型相结合,提出了马尔科夫转换(Markov-switch)回归模型,并将其应用于 UCI 数据集上和 SVM 回归算法进行比对,取得了非常好的效果。

(3)大胆提出将机器学习和经济学进行跨学科创新融合,利用跨学科知识互补的优势,使得回归模型的研究得到了进一步的发展和扩充。

## 1 机器学习方法的回归预测相关工作

对于回归问题的研究,国内外的科学工作者都进行了广泛的探讨。在传统的回归分析方法中,函数逼近理论分析思想严密,体系结构完整。但是由此发展而来的许多算法都有一些共同的缺点:计算量太大,对计算模型和数据的准确性要求很高,局限性很强。然而把人工神经网络应用于函数逼近有着很多的优点,具体体现在数据的特征不是很明确,数据模糊或含较多噪声和非线性等情况<sup>[6-7]</sup>。

### 1.1 BP 神经网络的回归预测

BP 神经网络<sup>[3]</sup>是 1986 年由 Rumelhart 和 McClelland 领导的科学家小组在《并行分布式处理》一书中提出来的,本书中对具有非线性连续变换函数的多层感知器的误差反向传播算法进行了详尽的分析,实现了 Minsky 关于多层网络的设想<sup>[3-4]</sup>。BP 神经网络的结构,反向传播(back propagation)神经网络,简称为 BP 神经网络。标准 BP 神经网络分 3 层,即输入层、隐含层和输出层,如图 1 所示。

在神经网络中误差反向传播网络和径向基函数(radial basis function, RBF)网络<sup>[8]</sup>是多层前向网络的两种典型网络,它们能够任意逼近任何非线性函数。由于它们结构简单、易于实现,已在时间序列分析、非

线性函数回归估计中得到了广泛的应用。然而,由于网络结构难以确定、存在过学习、容易陷入局部极值等问题,限制了此种网络的发展。

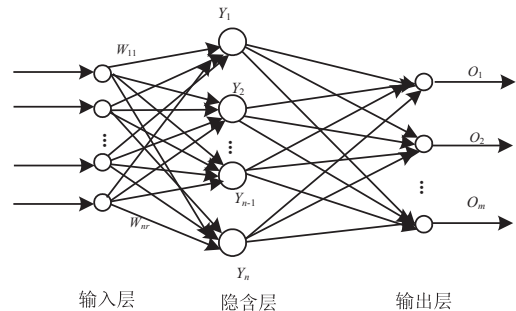


图1 三层 BP 神经网络的结构

### 1.2 GRNN 神经网络的回归预测

GRNN(广义回归神经网络)<sup>[3]</sup>最早是由 Sprech 提出的一种基于非线性回归理论的前馈式神经网络模型。由于 GRNN 网络的训练速度快和非线性映射能力很强,可以将其运用于函数逼近。广义回归神经网络的思想是:用径向基函数作为隐含层中各个节点的基,构成隐含层空间。隐含层对输入向量进行变换,将低维空间的模式变换到高维空间内,使得在低维空间的模式变换到高维空间内,使得在低维空间内的线性不可分问题在高维空间内线性可分。利用径向基神经元和线性神经元建立了广义回归神经网络(GRNN),并将 GRNN 应用于了函数的逼近,取得较好的效果。GRNN 网络的结构分析,利用径向基神经元和线性神经元可以创建广义回归神经网络。广义回归神经网络是由一个径向基网络层和一个线性网络层组成,见图 2,其中  $a_{i1}$  表示第一层输出  $a_1$  的第  $i$  个元素,  $W_{i1}$  表示第一层权值矩阵  $w_1$  的第  $i$  行元素。 $P$  表示输入向量,  $R$  表示网络输入的维数,  $S$  表示每层网络中的神经元个数,同时还表示训练样本的个数,  $b_1$  为隐含层阈值,符号  $\otimes$  表示  $\| \text{dist} \|$  输出与阈值  $b_1$  的元素之间的乘积关系。隐含层的传递函数为径向基函数,广义回归神经网络连接权值的学习修正仍然使用 BP 算法。广义回归神经网络中人为调节的参数少,只有一个阈值,网络的学习全部依赖数据样本,这个特点决定了网络得以最大限度地避免人为主观假定对预测结果的影响。

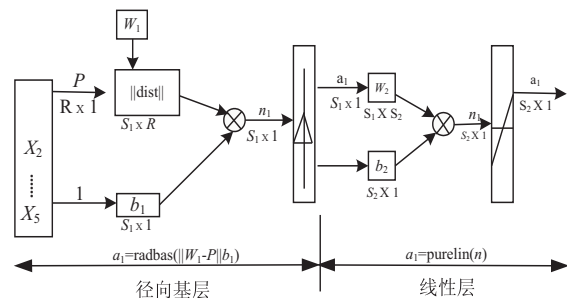


图2 广义回归神经网络的结构

但是,同样由于该网络结构难以确定、存在过学习、容易陷入局部极值等问题限制了此种网络的发展。上述三种网络都建立在渐进理论基础上,这三种网络需要无穷多的样本才能较真实地模拟样本的分布函数,而实际上所得的样本都是有限的。由此可以看出传统的神经网络在回归预测方面还是存在许多不可逆转的缺点,因此需要探索新的回归方法。

### 1.3 基于 SVM 的回归预测

Vapnik 等人<sup>[6]</sup>为了解决神经网络在处理回归问题上的瓶颈,提出了统计学习理论,专门研究小样本情况下机器学习的规律,并给出了该框架下的一种具体实现—支持向量机<sup>[6]</sup>。支持向量机,形式类似多层前向网络,在学习机器的结构复杂性和学习精度之间寻求折衷,获得最优泛化能力。它着眼于现有的有限小样本,将求取模型最优超平面问题转化为二次规划问题,求得全局最优。同时,将非线性问题通过核映射将样本映射到高维特征空间,在特征空间求取最优超平面,避免了繁杂的内积计算。该算法的主要目的是利用核函数在具体的特征空间  $\Omega$  下实现线性回归,具体的回归函数为  $f(x) = (\omega, \varphi(x)) + b$ , 其中  $\varphi: R^d \rightarrow \Omega$ ,  $(\omega, \varphi(x))$  表示特征空间中的内积运算,  $\omega \in \Omega, b \in R$ , 通过极小化下面的目标函数:

$$Q(\omega) = \frac{1}{2} \|\omega\|^2 + R_{\text{emp}}[f] = \frac{1}{N} \sum_{i=1}^N c(f(x_i), y_i) + \frac{1}{2} \|\omega\|^2$$

这里损失函数为  $\varepsilon$ -不敏感损失函数<sup>[7]</sup>,通过求解二次规划的优化问题来最小化结构风险,可以求得  $\omega = \sum_{i=1}^N (a_i^* - a_i) \varphi(x_i)$ , 其中  $a_i^*, a_i$  是通过求解二次规划问题来进行求解确定的。这里核函数  $k(x_i, y_i)$  为映射到特征空间的数据内积,即为  $k(x_i, y_i) = (\varphi(x_i), \varphi(y_i))$ , 因此回归函数可以写为  $f(x) = \sum_{i=1}^N (a_i^* - a_i) k(x_i, x_j) + b$ , 对极小化函数  $Q(x)$  的求解变为以下式子的求解。

$$\min \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^N (\xi_i + \xi_i^*)$$

$$\begin{cases} f_i(x) - y_i \leq \varepsilon + \xi_i \\ y_i - f_i \leq \varepsilon + \xi_i^*, 1 \leq i \leq N \\ \xi_i, \xi_i^* \geq 0 \end{cases}$$

其中,  $C$  为待确定的常数,  $\xi_i, \xi_i^*$  为松弛变量。通过对偶方法,可以求得最优回归线性函数  $\omega$  和支持向量。然而,在实际问题中运用支持向量机进行回归预测研究时,由于核函数参数在实际问题中的选择比较困难、复杂,且支持向量机回归算法复杂,导致出现训练速度

较慢、对大规模分类问题训练时间长等问题,支持向量机只适用于小样本数据的分类和回归问题。

## 2 基于 Markov 回归模型的研究

该文将 Markov 模型和回归相结合,提出了 Markov 多元回归模型,取得了良好的效果。同时将 Markov 和 switch 机制相结合,提出了马尔科夫转换 (Markov-switch) 回归模型,并将其应用于 UCI 公共数据集上,取得了良好的效果。

### 2.1 基于多元回归的 Markov 回归模型的研究与应用

在马尔科夫的预测问题研究<sup>[7]</sup>中,关键的地方是确定转移矩阵,但是由于具体问题的复杂性,往往使得转移矩阵难以确定。而回归分析<sup>[8-10]</sup>则是根据大量的观测数据值,建立相关的数学模型,设出相关的待估参数,然后再利用观测数据进行拟合,从而求出待估参数,得到具体的数学模型,最后对得到的模型进行评价。本部分利用多元回归理论进行数据的回归分析,求出马尔科夫模型的转移矩阵,从而建立起马尔科夫的回归模型。

#### 2.1.1 多元回归模型

一般的形如

$$y = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_{n-1} x_{n-1} + a_n x_n + \varepsilon \quad (1)$$

称为多元回归的线性模型<sup>[1]</sup>。其中,  $\varepsilon$  是误差,且要求  $\varepsilon$  独立同分布,  $a_0, a_1, \cdots, a_n$  是未知量。公式(1)中只有一次观测误差  $\varepsilon$ , 没有不同次观测值误差之间的关系,然而在实际的具体问题中,为了确定线性回归模型的未知参数,一定要有若干次的观测值序列,即:

$$\begin{cases} y_1, x_{11}, x_{12}, \cdots, x_{1m} \\ \vdots \\ y_n, x_{n1}, x_{n2}, \cdots, x_{nm} \end{cases}$$

所得到的多元回归方程为:

$$\begin{cases} y_1 = a_0 + a_1 x_{11} + \cdots + a_{n-1} x_{1(n-1)} + a_n x_{1n} + \varepsilon_1 \\ \vdots \\ y_n = a_0 + a_1 x_{n1} + \cdots + a_{n-1} x_{n(n-1)} + a_n x_{nn} + \varepsilon_n \end{cases} \quad (2)$$

在实际的统计问题中,根据 Gauss - Markov 条件<sup>[11-12]</sup>通常把误差  $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$  独立同分布的要求降低为  $E(\varepsilon_i) = 0, D(\varepsilon_i) = \sigma^2$ 。同时可以将式(2)表示为下面的情况,令:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \alpha = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ 1 & x_{21} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{bmatrix}$$

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

则式(2)可以写为:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \quad (3)$$

其中,  $E(\boldsymbol{\varepsilon}) = 0, D(\boldsymbol{\varepsilon}) = \sigma^2$ 。

在具体的问题中,通过收集或者实验可以得到一定数量的观测值序列:

$$\begin{cases} y_1, x_{11}, x_{12}, \dots, x_{1m} \\ \dots \\ y_n, x_{n1}, x_{n2}, \dots, x_{nm} \end{cases}$$

进而通过这些值来对式(2)中的参数  $a_0, a_1, \dots, a_n, \sigma^2$  进行求值估计,通常采用最小二乘法<sup>[10]</sup>,选择适当的  $\boldsymbol{\alpha}$ ,使得残差平方和:

$$H(\boldsymbol{\alpha}) = \sum_{i=1}^n (y_i - a_0 - a_1 x_{i1} - \dots - a_m x_{im})^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\alpha})'(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha})$$

达到最小值。

### 2.1.2 马尔科夫回归模型

设存在一个马尔科夫链,其状态集合为  $\{s_1, s_2, \dots, s_n\}$ ,转移矩阵为  $\mathbf{A} = (a_{ij})_{n,n}$ ,其中  $a_{ij} \geq 0, \sum_{i=1}^n a_{ij} = 1, j=1, 2, \dots, n$ 。在具体的问题中,假定国民收入有不同的分配类型,在  $t$  时刻第  $j$  种分配的占有率为  $y_t(j)$ ,  $j=1, 2, \dots, n; t=1, 2, \dots, m$ ,在  $t+1$  时刻各种分配类型的占有率为  $y_{t+1}(j)$ ,  $j=1, 2, \dots, n; t=1, 2, \dots, m$ ,易得  $\sum_{j=1}^n y_t(j) = 1; t=1, 2, \dots, m$ 。由齐次马尔科夫链的性质可得:

$$y_{t+1}(j) = \sum_{i=1}^n y_t(i) a_{ij}, \quad j=1, 2, \dots, n \quad (4)$$

将式(4)具体写出,即为:

$$\begin{bmatrix} y_t(1) & y_t(2) & \dots & y_t(n) \\ y_{t+1}(1) & y_{t+1}(2) & \dots & y_{t+1}(n) \\ \vdots & \vdots & \dots & \vdots \\ y_m(1) & y_m(2) & \dots & y_m(n) \end{bmatrix} = \begin{bmatrix} y_{t-1}(1) & y_{t-1}(2) & \dots & y_{t-1}(n) \\ y_t(1) & y_t(2) & \dots & y_t(n) \\ \vdots & \vdots & \dots & \vdots \\ y_m(1) & y_m(2) & \dots & y_m(n) \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \dots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \quad (5)$$

简写为  $\mathbf{Y} = \mathbf{X}\mathbf{A}$ ,其中  $a_{ij}, i, j=1, 2, \dots, n$  为需要估计的参数,此时采用多元回归的方法来进行参数的估计,设在  $t$  时刻误差变量为  $\varepsilon_t(j), j=1, 2, \dots, n$ , 这时有:

$$\begin{cases} y_{t+1}(1) = \sum_{i=1}^n y_t(i) a_{ij} + \varepsilon_t(1) \\ y_{t+1}(2) = \sum_{i=1}^n y_t(i) a_{ij} + \varepsilon_t(2) \\ \dots \\ y_{t+1}(n-1) = \sum_{i=1}^n y_t(i) a_{ij} + \varepsilon_t(n-1) \\ y_{t+1}(n) = \sum_{i=1}^n y_t(i) a_{ij} + \varepsilon_t(n) \end{cases}$$

### 2.1.3 马尔科夫回归模型的求解

根据多元回归的理论,可以用最小二乘法<sup>[13-15]</sup>来求解模型,令:

$$\mathbf{Y}_j = (y_1(j), y_2(j), \dots, y_n(j))'$$

$$\mathbf{A}_j = (a_{1j}, a_{2j}, \dots, a_{nj})'$$

$$\boldsymbol{\varepsilon}_j = (\varepsilon_1(j), \varepsilon_2(j), \dots, \varepsilon_n(j))'$$

$$\mathbf{X} = \begin{bmatrix} y_{t-1}(1) & y_{t-1}(2) & \dots & y_{t-1}(n) \\ y_t(1) & y_t(2) & \dots & y_t(n) \\ \vdots & \vdots & \dots & \vdots \\ y_m(1) & y_m(2) & \dots & y_m(n) \end{bmatrix}$$

从而式(5)可以写为:

$$\mathbf{Y}_j = \mathbf{X}\mathbf{A}_j + \boldsymbol{\varepsilon}_j, \quad j=1, 2, \dots, n$$

构造残差平方和可得:

$$\begin{aligned} Q(\mathbf{A}_j) &= |\mathbf{X}\mathbf{A}_j - \mathbf{Y}_j|^2 = (\mathbf{X}\mathbf{A}_j - \mathbf{Y}_j)'(\mathbf{X}\mathbf{A}_j - \mathbf{Y}_j) = \\ &= \mathbf{A}_j' \mathbf{X}' \mathbf{X} \mathbf{A}_j - \mathbf{A}_j' \mathbf{X}' \mathbf{Y}_j - \mathbf{Y}_j' \mathbf{X} \mathbf{A}_j + \mathbf{Y}_j' \mathbf{Y}_j = \\ &= (\mathbf{A}_j' \mathbf{X}' \mathbf{X} \mathbf{A}_j - \mathbf{Y}_j' \mathbf{X} \mathbf{A}_j) - (\mathbf{A}_j' \mathbf{X}' \mathbf{Y}_j - \mathbf{Y}_j' \mathbf{Y}_j) = \\ &= (\mathbf{A}_j' \mathbf{X}' \mathbf{X} - \mathbf{Y}_j' \mathbf{X}) \mathbf{A}_j - (\mathbf{A}_j' \mathbf{X}' \mathbf{X} \mathbf{X}^{-1} \mathbf{Y}_j - \mathbf{Y}_j' \mathbf{X} \mathbf{X}^{-1} \mathbf{Y}_j) = \\ &= (\mathbf{A}_j' \mathbf{X}' \mathbf{X} - \mathbf{Y}_j' \mathbf{X}) \mathbf{A}_j - (\mathbf{A}_j' \mathbf{X}' \mathbf{X} - \mathbf{Y}_j' \mathbf{X}) \mathbf{X}^{-1} \mathbf{Y}_j = \\ &= (\mathbf{A}_j' \mathbf{X}' \mathbf{X} - \mathbf{Y}_j' \mathbf{X}) (\mathbf{A}_j - \mathbf{X}^{-1} \mathbf{Y}_j) = \\ &= ((\mathbf{A}_j' \mathbf{X}' - \mathbf{Y}_j' \mathbf{X})')' (\mathbf{A}_j - \mathbf{X}^{-1} \mathbf{Y}_j) = \\ &= (\mathbf{X} \mathbf{X}' \mathbf{A}_j - \mathbf{X}' \mathbf{Y}_j)' ((\mathbf{X} \mathbf{X}')^{-1} \mathbf{A}_j - \mathbf{X}^{-1} \mathbf{Y}_j) = \\ &= (\mathbf{X} \mathbf{X}' \mathbf{A}_j - \mathbf{X}' \mathbf{Y}_j)' (\mathbf{X} \mathbf{X}')^{-1} (\mathbf{X} \mathbf{X}' \mathbf{A}_j - \mathbf{X}' \mathbf{Y}_j) \end{aligned}$$

由上式易得,当  $\mathbf{X} \mathbf{X}' \mathbf{A}_j = \mathbf{X}' \mathbf{Y}_j$  时,  $Q(\mathbf{A}_j)$  取得最小值,此时  $\mathbf{A}_j = (\mathbf{X} \mathbf{X}')^{-1} \mathbf{X}' \mathbf{Y}_j, j=1, 2, \dots, n$ ,从而得到马尔科夫的转移矩阵的第  $j$  列的估计为  $\hat{\mathbf{A}}_j = (\mathbf{X} \mathbf{X}')^{-1} \mathbf{X}' \mathbf{Y}_j, j=1, 2, \dots, n$ ,进而转移矩阵的估计为  $\mathbf{A} = (\mathbf{X} \mathbf{X}')^{-1} \mathbf{X}' \mathbf{Y}$ 。

### 2.1.4 基于多元回归的马尔科夫回归算法的设计

算法1:多元回归的马尔科夫回归算法。

输入:实验数据,并对数据进行处理。



(a) 用最小二乘法来确定 Markov 转移矩阵;

(b) 判断  $|XX'|$  是否为 0, 如果是则跳出本次程序, 否则进行 (c) 步运算;

(c) 根据 Markov 转移矩阵来确定不同参数的方程;

(d) 根据已知的数据, 由确定的方程来进行数据的拟合;

(e) 计算误差 (真实值-拟合值)。

输出: 确定显著性评价水平, 对回归方程进行假设检验 (F 检验)。

## 2.2 基于 Markov-switch 的回归研究

### 2.2.1 体制转换模型

Hamilton, Kim and Nelso 等人在 20 世纪 90 年代提出了 Regime Switching Model<sup>[16]</sup>, 并将其用于时间序列的分析之中。随着 Markov 理论的发展, 经济学家们认识到在整个经济活动的过程中, 通用的经济模型如差分自回归移动平均模型<sup>[12]</sup> (autoregressive integrated moving average model, ARIMA), 模型中的参数并不为常数, 而是存在着具体的结构变化, 必须将总体样本分解成若干个拥有不同参数的子样本, 而且由于数据来自不同的产生过程, 所以回归方程也会从一个状态向另一个状态转变。国外的许多学者对有关经济、金融时间序列的离散转换模型进行了许多的研究, 得出允许序列变量非线性、动态地发生改变。最初的体制转换模型为:

$$y_t = \mu_i + \varepsilon_t \quad (6)$$

其中,  $i = 1, 2, \dots, k$ ,  $\varepsilon_t$  服从均值为 0、方差为  $\sigma_i^2$  的正态分布。式 (6) 的含义是如果存在  $k$  个状态, 将会有  $k$  个不同的  $\mu, \sigma^2$  值。在特殊的情况下  $i = 1$  时, 此时有  $y_t = \mu_1 + \varepsilon_t$ , 这是一个最基本的线性回归模型。当  $k = 2$  时, 即有两个状态, 此时的转换模型如下:

$$y_t = \mu_1 + \varepsilon_t, \varepsilon_t \sim N(0, \sigma_1^2) \quad (7)$$

$$y_t = \mu_2 + \varepsilon_t, \varepsilon_t \sim N(0, \sigma_2^2) \quad (8)$$

这样对于不同的状态, 误差项对应于不同的状态中, 同时也反映了不同的可信度标准, 对用不同的自变量, 所得的因变量序列也将是不同的。

### 2.2.2 Markov-switch 转换回归模型

由此可以扩展, 对于基于 Markov 理论的体制转换模型<sup>[17]</sup>, 状态的转换过程是随机的, 一个确定的状态可能转换为其他的任意状态, 也可能转换为自己的状态 (即不发生状态的转换)。在整个状态的转换过程中, 每个状态转换为其他的状态或是转换为自己的状态, 可以用状态转移矩阵来进行控制。状态转移矩阵为:

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & \cdots & \vdots \\ a_{k1} & \cdots & a_{kk} \end{bmatrix}$$

其中,  $a_{ij}$  表示从  $i$  状态转换为  $j$  状态的概率。这样对于  $k$  个状态的 Markov 转换模型为:

$$y_{1t} = \alpha_{11}x_1 + \alpha_{12}x_2 + \cdots + \alpha_{1k}x_k$$

:

$$y_{kt} = \alpha_{k1}x_1 + \alpha_{k2}x_2 + \cdots + \alpha_{kk}x_k$$

### 2.2.3 马尔科夫转换回归模型的算法

输入: 数据序列, 从中提取因变量  $Y$ , 和自变量序列  $(x_1, x_2, \dots, x_n)$ , 状态转换个数  $K$ 。

(a) 计算在数据序列存在的情况下, 状态序列的条件概率, 即为:

$$P(X_t = x_t, X_{t-1} = x_{t-1}, \dots, X_1 = x_1 \mid y_{t-1}, y_{t-2}, \dots, y_1) =$$

$$P(X_t = x_t \mid X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, \dots, X_1 = x_1, y_{t-1}, y_{t-2}, \dots, y_1)$$

$$P(X_{t-1} = x_{t-1}, \dots, X_1 = x_1 \mid y_{t-1}, y_{t-2}, \dots, y_1)$$

其中, 根据 Hamilton<sup>[16]</sup> 的体制转换模型, 这里可以写为:

$$P(X_t = x_t, X_{t-1} = x_{t-1}) =$$

$$P(X_t = x_t \mid X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, \dots, X_1 = x_1, y_{t-1}, y_{t-2}, \dots, y_1)$$

(b) 计算联合条件概率密度分布 (该算法的密度函数为正态分布):

$$f(y_t, X_t = x_t, X_{t-1} = x_{t-1}, \dots, X_1 = x_1 \mid y_{t-1}, y_{t-2}, \dots, y_1) =$$

$$f(y_t \mid X_t = x_t, X_{t-1} = x_{t-1}, \dots, X_1 = x_1, y_{t-1}, y_{t-2}, \dots, y_1) \cdot P(X_t = x_t, X_{t-1} = x_{t-1}, \dots, X_1 = x_1 \mid y_{t-1}, y_{t-2}, \dots, y_1)$$

这里采用  $f \sim N(0, 1)$ 。

(c) 计算数据序列的概率密度函数。

$$f(y_t \mid y_{t-1}, y_{t-2}, \dots, y_1) =$$

$$\sum_{X_t=0}^1 \sum_{X_{t-1}=0}^1 \cdots \sum_{X_1=0}^1 f(y_t, X_t = x_t, X_{t-1} = x_{t-1}, \dots, X_1 = x_1 \mid y_{t-1}, y_{t-2}, \dots, y_1)$$

(d) 通过 Step4 的计算可以得出:

$$P(X_t = x_t, X_{t-1} = x_{t-1}, \dots, X_1 = x_1 \mid y_{t-1}, y_{t-2}, \dots, y_1) = \sum_{X_t=0}^1 P(X_t = x_t, X_{t-1} = x_{t-1}, \dots, X_1 = x_1 \mid y_{t-1}, y_{t-2}, \dots, y_1)$$

输出: 通过上面的计算输出模型的标准差、相关系数和程序运行的时间。

## 3 实验结果及分析

实验 1 (基于 Markov 多元回归模型在国民收入的应用)。

实验使用广东省主要年份国民收入使用额统计数据 (1952-1995) 来进行回归实验, 见表 1<sup>[18]</sup>。

表 1 广东省主要年份国民收入使用额统计数据(1952-1995)

年份	国民收入使用额 (亿元)	消费额	积累额	消费率	积累率
1957	51.25	42.44	8.81	0.828	0.172
1962	53.83	50.27	3.56	0.934	0.066
1965	69.26	57.43	11.83	0.829	0.171
1970	89	69.5	19.5	0.781	0.219
1975	121.45	89.75	31.7	0.739	0.261
1978	147.57	105.62	41.95	0.716	0.284
1980	209.25	162.38	46.87	0.776	0.224
1985	485.3	313.32	171.98	0.646	0.354
1990	1 024.52	688.37	336.15	0.672	0.328
1991	1 220.68	810.86	409.82	0.664	0.336
1994	3 175.14	1 911.43	1 263.71	0.602	0.398
1995	4 002.83	2 529.79	1 473.04	0.632	0.368

采用消费率和积累率作为实验的数据,由算法 1 可以得到马尔科夫的转移矩阵为:

$$A = \begin{bmatrix} 0.923 & 1 & 0.076 & 9 \\ 0.158 & 1 & 0.841 & 9 \end{bmatrix}$$

从而得到广东省主要年份的国民收入的消费率和积累率的回归方程为:

$$\begin{cases} y_{t+1}(\text{consume}) = 0.923 \ 1y_t(\text{consume}) + \\ \qquad \qquad \qquad 0.158 \ 1y_t(\text{accumulate}) \\ y_{t+1}(\text{accumulate}) = 0.076 \ 9y_t(\text{consume}) + \\ \qquad \qquad \qquad 0.841 \ 9y_t(\text{accumulate}) \end{cases}$$

用回归方程进行数据的测算实验可以得到表 2。

表 2 消费率与积累率的真实值与实验值的比对

消费率 (真实值)	积累率 (真实值)	消费率 (实验值)	积累率 (实验值)	消费率误差	积累率误差
0.861	0.139	0.816 8	0.183 2	0.044 2	0.044 2
0.828	0.172	0.791 5	0.208 5	0.036 5	-0.036 5
0.934	0.066	0.872 6	0.127 4	0.061 4	-0.061 4
0.829	0.171	0.792 3	0.207 7	0.036 7	-0.036 7
0.781	0.219	0.755 6	0.244 4	0.025 4	-0.025 4
0.739	0.261	0.723 4	0.276 6	0.015 6	-0.015 6
0.716	0.284	0.705 8	0.294 2	0.010 2	-0.010 2
0.776	0.224	0.751 7	0.248 3	0.024 3	-0.024 3
0.646	0.354	0.652 3	0.347 7	-0.006 3	0.006 3
0.672	0.328	0.672 2	0.327 8	-0.000 2	0.000 2
0.664	0.336	0.666 1	0.333 9	-0.002 1	0.002 1
0.602	0.398	0.618 6	0.381 4	-0.016 6	0.016 6
0.632	0.368	0.641 6	0.358 4	-0.009 6	0.009 6

这里对上面的方程在显著性水平为 0.05 和 0.025 的情况下,进行假设检验可得:

$$F_1(2,30) = 6.589 \ 2 > F_{0.05}(2,30) = 3.32$$

$$F_1(2,30) = 6.589 \ 2 > F_{0.025}(2,30) = 4.18$$

$$F_2(2,30) = 15.41 > F_{0.05}(2,30) = 3.32$$

$$F_2(2,30) = 15.41 > F_{0.025}(2,30) = 4.18$$

由表 2 可以看出,实验的效果很好,误差很小。因

此,应用 Markov 的回归方法可以有效预测国民收入分配的情况。

实验 2(Markov-switch 回归模型)。

本实验采用 UCI 数据集的 abalone 数据集,是由 4 177 \* 8 的一个数据集(实例是 4 177 种,属性是 8 维数),wine 数据集。采用的回归衡量标准(样本标准差,相关系数),其中样本标准差为:

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$$

相关系数为:

$$R = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}} = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sqrt{[\sum X^2 - \frac{(\sum X)^2}{n}][\sum Y^2 - \frac{(\sum Y)^2}{n}]}}$$

样本标准差是衡量一组数据分散的程度,标准差越高,说明实验数据越离散,差异越大,也就是实验越不精确。反之,标准差越低,则说明实验的数据越聚

合,差异较小,实验越精确。相关系数反映的是变量之间的相关程度,R 值越大,表明误差越小,变量之间的线性相关程度越高。相关系数越大,也说明样本点较大程度地聚集在函数的回归曲线上,相关系数越小,说明样本点分布在回归曲线上越离散。由实验结果可以看出,提出的 Markov 转换回归算法在标准差方面比 SVM 方法<sup>[19-21]</sup>小,说明实验数据聚合度好、差异小、实验精度高。Markov 转换回归算法相关系数比 SVM 大,说明误差比 SVM 算法小,变量之间的线性相关程度高。在运行时间上,Markov 转换回归算法比 SVM 算法缩短了 50%,取得了非常好的效果。

实验结果见表 3。

表 3 实验结果

abalone 数据集实验					wine 数据集实验				
		标准差 S	相关系数 R	时间/秒			标准差 S	相关系数 R	时间/秒
		Mean squared error	Squared correlation coefficient				Mean squared error	Squared correlation coefficient	
Markov 转换回归	状态 1	0.179 93	0.996 25	445.188	Markov 转换回归	状态 1	0.069 03	0.960 34	87.407
	状态 2	0.114 62	0.998 43			状态 2	0.086 45	0.962 26	
SVM		0.196 29	0.989 77	933.156	SVM		0.092 55	0.928 90	133.547

## 4 结束语

首先介绍了回归分析和预测的基本知识,其次分析了传统的神经网络如 BP 神经网络和广义回归神经网络在回归预测方面的研究,讨论了传统神经网络在回归预测方面的不足。随后分析了基于统计学理论的 SVM 回归预测算法,得出了此方法的不足之处。然后提出基于多元回归的马尔科夫模型,对其在回归预测方面的研究进行了探讨,使用多元回归的方法可以解决马尔科夫矩阵难以确定的问题,提升了马尔科夫算法的预测效率。最后对基于马尔科夫理论的转换回归模型进行了研究,提出了马尔科夫转换算法并通过实验和 SVM 回归算法进行比较得到良好的结果。

### 参考文献:

- [1] 张小蒂. 应用回归分析[M]. 杭州:浙江大学出版社,1991:11-12.
- [2] 王 睿. 关于支持向量机参数选择方法分析[J]. 重庆师范大学学报:自然科学版,2007(4):36-42.
- [3] 何成刚,张燕平,张 站,等. 机器学习中知识动态获取在函数逼近中的探究[J]. 微计算机信息,2010,26(27):134-136.
- [4] 刘嘉焜. 应用概率统计[M]. 北京:科学出版社,2004:102-115.
- [5] HAMILTON J D. Regime switching models[M]//The new palgrave dictionary of economics. [s. l.]:[s. n.],2005:5471-5475.
- [6] VAPNIK V N. 统计学习理论的本质[M]. 张学工,译. 北京:清华大学出版社,2000:88-90.
- [7] HE Chenggang,ZHANG Yanping,SUN Hui,et al. A novel regression method research based on covering algorithm[C]//2010 international symposium on intelligence information processing and trusted computing. Wuhan:IEEE,2010:41-44.
- [8] 李 勇,刘鹤飞,王 坤,等. 隐马尔科夫多元线性回归模型中未知隐状态个数的贝叶斯模型选择[J]. 西南师范大学学报:自然科学版,2020,45(7):11-17.
- [9] LEE J,JEONG J Y,JUN C H. Markov blanket-based universal feature selection for classification and regression of mixed-type data[J]. Expert Systems with Applications,2020,158:113398.
- [10] SIDDQUI A,SIDDQUI A,MAITHANI S,et al. Urban growth dynamics of an Indian metropolitan using CA Markov and logistic regression[J]. The Egyptian Journal of Remote Sensing and Space Science,2018,21(3):229-236.
- [11] MARINO M F,TZAVIDIS N,ALFÒ M. Mixed hidden Markov quantile regression models for longitudinal data with possibly incomplete sequences[J]. Statistical Methods in Medical Research,2018,27(7):2231-2246.

(下转第 38 页)