Software Engineering and Applications 软件工程与应用, 2021, 10(1), 10-16
Published Online February 2021 in Hans. http://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2021.101002

Hans汉斯

一种基于遗传算法的回归模型寻优方法

罗晓霞，王　佳，罗香玉

西安科技大学计算机科学与技术学院，陕西　西安
Email: 1443751599@qq.com

收稿日期：2020年12月16日；录用日期：2021年1月4日；发布日期：2021年2月24日

摘　要

回归分析是数据分析和建模的重要工具，主要用于数据的预测和拟合。回归分析通常需要人工干预给定参考模型，再进行参数回归。然而，在多数情况下，用户难以给出参考模型，或者给出模型具有较大的误差。本文提出一种基于遗传算法得出回归模型的方法，主要利用遗传进化的思想，首先随机产生初始模型的种群；然后不断迭代的进行选择、交叉、变异操作，在解空间中动态地进行全局寻优，找出一个较优的模型；为了确定模型的参数，又利用梯度下降法对该模型进行参数估算。最后，将本文得出的模型与最小二乘法回归分析得出的模型进行对比，结果表明，在进行预测时，前者的误差比后者有显著减小，由14.24%减少到9.59%。

关键词

回归分析，遗传算法，最小二乘法，寻优，预测

A Regression Model Optimization Method Based on Genetic Algorithm

Xiaoxia Luo[1], Jia Wang[2], Xiangyu Luo[3]

College of Computer Science and Technology, Xi'an University of Science and Technology, Xi'an Shaanxi
Email: 1443751599@qq.com

Received: Dec. 16th, 2020; accepted: Jan. 4th, 2021; published: Feb. 24th, 2021

Abstract

Regression analysis is an important tool for data analysis and modeling, mainly used for data prediction and fitting. Regression analysis usually requires manual intervention of a given reference model followed by parametric regression. However, in most cases, it was difficult for the user to

文章引用: 罗晓霞, 王佳, 罗香玉. 一种基于遗传算法的回归模型寻优方法[J]. 软件工程与应用, 2021, 10(1): 10-16.
DOI: 10.12677/sea.2021.101002

given a reference model or given the model a large error. It proposed a method based on genetic algorithm to obtain regression model. It mainly used the idea of genetic evolution to first randomly generated an initial model populations; then iteratively selected, crossed, and mutated operations, perform global optimization dynamically in the solution space to find a better model; in order to determine the parameters of the model, the gradient descent method is used to estimate the parameters of the model. Finally, the model obtained in this paper is compared with the model obtained by least squares regression analysis. The results show that the error of the former is significantly reduced from the previous one, from 14.24% to 9.59%.

Keywords

Regression Analysis, Genetic Algorithm, Least Squares Method, Optimization, Prediction

1. 引言

回归分析是一种确定多个变量之间依赖关系的统计分析方法，已经广泛应用于各大学科领域，例如经济学[1]、医学[2]、气象学[3]、人文学[4]等一系列的领域。随着统计学的不断发展，回归分析模型更广泛的应用于数据分析。施龙青等人[5]基于多元回归方法进行防隔水断层的预测，这种多元的回归方法，基于先前的经验，人工给定参考模型，虽然预测结果也是相对准确的，但需要人工干预。卢骏等人[6]对大坝变形问题进行了变系数回归建模，该方法需要事先给出一种模型，然后进行参数估计。Taylor 首先提出了回归神经网络，不仅能运用到金融方面，避免了那些不必要的损失之外，而且还能用神经网络的结构预测没有开发出来的、潜在的曲线模型，对其进行较为准确的预测。虽然回归神经网络能对数据进行相对准确的预测，但是它的操作是一个黑箱操作，无法看清箱内的计算过程和产生的模型，在多数情况下，是很难理解这一过程的。

通过以上分析，现有回归分析方法一般要求用户给出参考模型，通过数据样本分析，求出最佳的模型参数。然而，在多数情况下，用户难以给出参考模型，或者所给出模型具有较大的误差。因此，针对该问题，本文提出一种基于遗传算法[7]的模型寻优方法。利用了遗传进化的思想，首先随机产生大量初始模型种群，然后按照一定的遗传进化法则对模型进行淘汰，找出一个接近于最优解的模型，最后对该模型进行参数估算。

2. 方法设计

首先给出问题定义，然后基于遗传算法进行模型设计。根据个体编码随机产生 N 个初始模型的种群，然后进行交叉、变异操作，按照适应度函数和个体的选择方式，从种群中选出下一代种群，继续进行寻优过程，不断迭代直到找出较优解。

2.1. 问题定义

在参考模型未知时，对数据进行回归分析，将该问题定义如下：

已知：1) n 个自变量的个数和取值。

2) 可供选择的基本单目运算符：“x”，“x^2”，“$\sqrt{x}$”，双目运算符号：“+”、“−”。

3) 双目运算符的个数 N (决定回归方程的长短，N 越大，回归方程越长)。

求解：因变量 y 关于 x 的回归方程。

2.2. 个体编码

2.2.1. 编码长度

对回归模型进行编码，模型的长度是由其中双目运算符的个数确定的，不同的双目运算符个数，产生的编码长度是不同的。把双目运算符的个数记为 n，那么编码长度 N (N 的初始值为 2)和双目运算符个数 n 之间的关系可以用公式(1)来表示：

$$N=\begin{cases}2 & n=0\\ 2*N+2 & n>0\end{cases} \tag{1}$$

2.2.2. 编码规则

将参与回归模型编码的算子分为三类，分别是单目运算符、双目运算符和操作数。然后对这 3 种类型的算子进行编码设计，如表 1 所示。

Table 1. Code
表 1. 编码

单目运算符	基本算子	x	x^2	$\sqrt{x}$
	编码	1	2	3
双目运算符	基本算子	null	+	-
	编码	0	1	2
操作数	基本算子	x_1	x_2	x_3
	编码	1	2	3

这 3 种算子按照单目运算、操作数 1、双目运算、操作数 2 的顺序进行组合，随机产生个体编码。不同的编码长度对应的编码规则有所不同，编码规则采用嵌套的格式，其中操作数可以由单目运算和操作数的组合产生。例如，当 n = 0 时，编码长度为 2 位，只存在单目运算和操作数，是最简单的原子操作，对某个操作数进行单目运算。例如：编码“21”表示“x_1^2”；“32”表示“$\sqrt{x_2}$”；“23”表示“x_3^2”。当 n = 1 时，编码长度为 6 位。例如：编码“121132”表示“$x_1^2+\sqrt{x_2}$”；“212123”表示“$\left(x_2+x_3^2\right)^2$”；“321122”表示“$\sqrt{\left(x_1^2+x_2^2\right)}$”。

2.3. 交叉和变异因子

假设染色体的长度为 L，那么计算机随机产生一个(1, L)范围内的整数 r，然后把要交叉[8]的两个母代个体从 r 这个位置截为两段，分别交换母代个体的后半段，就产生了新子代个体。

种群中每个个体都要按照变异概率判断是否变异[9]，一般情况下变异概率是小于 0.5 的，甚至小于 0.1，所以最终的种群中只有一少部分的个体发生了变异，本文采用单点变异的方法，变异的位置点是随机产生的。

2.4. 个体的选择方式

选择操作在遗传算法中起到关键作用，选择出的个体是否优胜会决定后代个体的质量，在此，适应度函数选用均方误差(MSE)来评判。使用轮盘赌算法[10]对种群中的个体进行选择。假设，先把个体适应

度函数的值记为 $F(i)$，种群中所有个体适应度的值为 $TotalF$，每个个体被选中的概率记为 $p(i)$，那么

$$p(i)=\frac{F(i)}{TotalF} \tag{2}$$

由公式(2)得出，个体 i 适应度值越大，$p(i)$的值越大。把后代种群数量记为 S，就意味着，一共要执行 S 次轮盘赌算法，从原始种群中选择出 S 个个体，放入子代种群中进行交叉变异等操作。

2.5. 建立回归模型的基本步骤

基于遗传算法的回归模型寻优方法，主要利用遗传算法适者生存、优胜劣汰的思想寻找较优的模型，然后利用梯度下降法进行参数回归。其基本步骤如下：

输入：含有 N 个自变量值的文本文件；

输出：回归模型。

Step1：根据遗传算法参数初始化的历史经验，对种群规模(M)、交叉发生的概率(Pc)、变异发生的概率(Pm)、终止进化的代数(G)进行初始化。根据 1.2 节编码规则，随机产生第一代初始种群 Pop；

Step2：do

{

计算种群 Pop 中每一个体的适应度 $F(i)$。

初始化空种群 $newPop$

do

{

根据 1.4 节中的个体选择方式，从种群 Pop 中选出 2 个个体

if($random(0,1)<Pc$)

{

对 2 个个体按交叉概率 Pc 执行交叉操作

}

if($random(0,1)<Pm$)

{

对 2 个个体按变异概率 Pm 执行变异操作

}

将 2 个新个体加入种群 $newPop$ 中

} until(M 个子代被创建)

用 $newPop$ 取代 Pop

}until(繁殖代数超过 G)

Step3：根据 Step2 产生的模型，用梯度下降法进行参数回归。初始化迭代步数为 S，模型参数为 $\theta_1, \theta_2, \cdots$；

Step4：迭代更新这些参数使目标函数 $J(\theta)$不断变小，直到迭代次数到达 S 停止迭代，$J(\theta)$的计算如公式(3)所示：

$$J(\theta)=\frac{1}{2}\sum_{i=1}^{m}\left(h\left(x^i\right)-y^i\right)^2 \tag{3}$$

3. 实验与结果分析

为了验证本文所得模型预测的准确性，进行回归分析实验。本次实验对一组化学动力学反应的数据

进行回归分析，该数据是在数学建模比赛中测得的真实数据，主要分析反应速度和反应物含量的关系。将相同的实验数据分别用于本文方法和最小二乘法回归分析方法，比较两种方法结果的误差大小。本次实验采用的操作系统是 Windows2007，编译工具为 myeclipse2014。

3.1. 实验过程

3.1.1. 基于遗传算法的回归模型寻优方法

首先，将化学动力学反应数据存入一个文本文件，各个变量之间用空格隔开，实验数据如表 2 所示。初始化算法参数，设置种群大小为 50，交叉的概率为 60%，变异的概率为 0.3，迭代次数为 50。

Table 2. Relationship between reaction rate and reactant content
表 2. 反应速度和反应物含量的关系

氧气 x_1	戊烷 x_2	异构戊烷 x_3	反应速度 y
470	300	10	8.55
285	80	10	3.79
470	300	120	4.82
470	80	120	0.02
470	80	10	2.75
100	190	10	14.39
100	80	65	2.54
470	190	65	4.35
100	300	54	13
100	300	120	8.5
100	80	120	0.05
285	300	10	11.32
285	190	120	3.13

实验得到编码为“113131”的模型，根据 1.2 节个体编码可以得出，该编码对应的回归模型为：$y = x_3 + \sqrt{x_1}$，得到参数估算结果分别为：0.76 和 0.3，综上所述，本文方法所得的模型为：$y = 0.76x_3 + 0.3\sqrt{x_3}$ 。

3.1.2. 最小二乘法回归分析方法

“统计产品与服务解决方案(spss) [11]”软件可以对数据进行回归分析，使用该软件对化学动力学反应数据进行最小二乘法回归分析得到的结果如图 1 所示。t 值是单样本检验，展现了该自变量对因变量是否有显著性影响，最后一列的 t 值所对应的 Sig 值，如果小于 0.05，代表该自变量对结果的影响程度越高，表中系数一列表达了该自变量在回归方程中的系数，如果系数为正，该自变量与因变量则为正比例的关系；如果系数为负数，那么自变量与因变量就为反比例的关系。

由图 1 可得，反应物氧气 x_1，对应系数为：−0.386，反应物戊烷 x_2，对应系数为 0.701，反应物异构戊烷 x_3，对应系数为−0.509。综上所述，最小二乘法回归分析所得模型为：$y = -0.386 * x_1 + 0.701 * x_2 - 0.509 * x_3$ 。

系数[a]

模型		非标准化系数		标准系数	t	Sig.
		B	标准误差	试用版		
1	(常量)	-0.459	2.139		-0.214	0.834
	戊烷	0.034	0.010	0.711	3.354	0.006
2	(常量)	2.528	1.934		1.307	0.220
	戊烷	0.03	0.008	0.702	4.303	0.002
	异构戊烷	-0.045	0.015	-0.478	-2.934	0.015
3	(常量)	5.818	1.672		3.480	0.007
	戊烷	0.033	0.005	0.701	6.131	0.000
	异构戊烷	-0.048	0.011	-0.509	-4.439	0.002
	氧气	-0.011	0.003	-0.386	-3.366	0.008

a.因变量:反应速度

Figure 1. Coefficient
图 1. 系数

3.2. 两种模型预测应用

分别运用以上两个模型对化学动力学反应中的反应速度进行预测，预测结果如表 3 所示。

Table 3. Comparison of prediction results of two models
表 3. 两种模型预测结果对比

原始数据	本文方法预测结果	最小二乘法预测结果
8.55	14.11	23.79
3.79	12.66	−59.02
4.82	97.71	−32.2
0.02	97.71	−186.42
2.75	14.11	−130.43
14.39	10.6	89.5
2.54	52.4	−15.605
4.35	55.91	−81.315
13	44.04	144.214
8.5	94.2	110.62
0.05	94.2	−43.6
11.32	12.66	95.2
3.13	96.26	−37.9

将以上预测数据进行对比，结果如图 2 所示。使用本文方法所得模型进行预测，结果波动不大，与原始数据相差较小，而最小二乘法回归模型预测出现负值，波动较大，与原始数据相差较远。

由公式(4)计算两种模型预测结果的均方根误差，最小二乘法回归分析预测结果的误差为 14.24，本文方法得出模型预测结果的误差为 9.59。

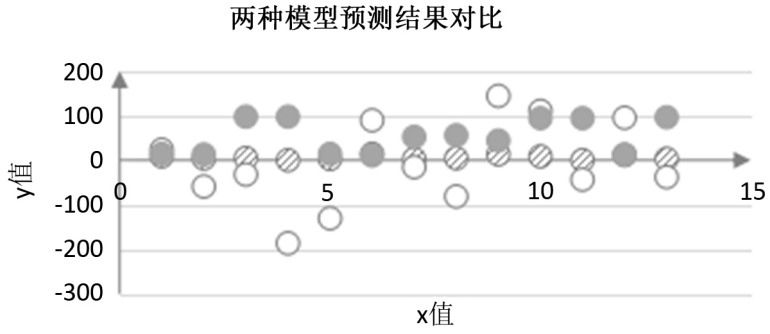

Figure 2. Comparison of prediction results of two models
图 2. 两种模型预测结果对比图

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} w_i \left(y_i - \hat{y_i}\right)^2} \tag{4}$$

两个误差结果比较可得：$9.59 \ll 14.24$，显然，本文方法得到的模型误差更小。但在设计个体编码的过程中，由于单目运算和双目运算种类的限制，模型的表达范围缩小，使得模型的准确度降低，误差仍然较大。

4. 结语

本文提出了一种基于遗传算法的回归模型寻优方法，该方法能够在参考模型未知的情况下计算出回归模型，并进行参数估算。实验表明，基于遗传算法的回归分析模型，在进行预测时，与最小二乘法回归分析模型比较，误差更小。本文方法仅针对双目运算符中的加减运算进行了研究，下一步工作是增加单目以及双目运算符的种类，使模型变得更加通用。

基金项目

国家自然科学基金青年项目(61702408)；国家自然科学基金重点项目(51634007)；陕西省自然科学基础研究计划项目(2019JM-020)。

参考文献

[1] 张钰珩. 浅析回归分析在经济金融领域的运用[J]. 商场现代化, 2018(1): 147-148.

[2] 王曼. 医学论文中常用回归分析方法的审核要点及对策[J]. 编辑学报, 2018, 30(5): 475-477.

[3] 孔德兵, 尚可政, 王式功, 等. 基于逐步回归分析的西北地区东部雷暴概率预报方法研究[J]. 干旱气象, 2016, 34(1): 181-187.

[4] 孙克, 徐中民. 基于地理加权回归的中国灰水足迹人文驱动因素分析[J]. 地理研究, 2016, 35(1): 37-48.

[5] 施龙青. 基于多元回归分析法预测断层防隔水煤柱宽度[J]. 煤炭科学技术, 2013, 41(6): 108-110.

[6] 卢骏, 戴吾蛟, 章浙涛. 大坝变形变系数回归建模[J]. 武汉大学学报(信息科学版), 2015, 40(1): 139-142.

[7] Nijhout, F. (1997) An Introduction to Genetic Algorith-MS. *Complexity*, **2**, 39-40. https://doi.org/10.1002/(SICI)1099-0526(199705/06)2:5<39::AID-CPLX8>3.0.CO;2-L

[8] 蔡良伟, 李霞. 遗传算法交叉操作的改进[J]. 系统工程与电子技术, 2006, 28(6): 925-928.

[9] 周祥, 何小荣, 陈丙珍. 基于最优变异因子的遗传算法在 ANN 训练中的应用[J]. 清华大学学报(自然科学版), 2002(5): 619-621.

[10] 马洁莹. 基于轮盘赌策略的混沌萤火虫算法研究[D]: [硕士学位论文]. 西安: 西安电子科技大学, 2018.

[11] 黄文霞, 李民. 基于 SPSS 数据分析的影响旅游地区发展的主要因素分析[J]. 软件, 2019, 40(1): 152-157.