文章编号：2095-2163(2021)08-0143-04 中图分类号：TP399 文献标志码：A

遗传算法优化的支持向量机回归计算老龄化人口方法

张馨予，孙宏宇，逯 洋，郭天岚
（吉林师范大学 计算机学院，吉林 四平 136000）

摘 要：老龄化问题对中国经济、就业、医疗等方面的影响越来越大，因此对老龄化人口进行准确预测具有重要的意义。本文利用遗传算法（GA），对支持向量回归模型（SVR）的初始参数进行优化，并利用优化后的 SVR 模型预测中国老龄化人口。实验证明，使用遗传算法参数寻优后的 SVR 模型具有良好的预测精度。

关键词：老龄化人口；遗传算法；支持向量回归模型

Genetic algorithm optimized support vector machine regression method for aging population calculation in China

ZHANG Xinyu, SUN Hongyu, LU Yang, GUO Tianlan
(College of Computer Science, Jilin Normal University, Siping Jilin 136000, China)

【Abstract】The aging problem has an increasing impact on China's economy, employment, medical care and other aspects, so it is of great significance to accurately predict the aging population. In this paper, genetic algorithm (GA) is used to optimize the initial parameters of support vector regression model (SVR), and the optimized SVR model is used to predict the aging population in China. Experiments show that the SVR model optimized by genetic algorithm has good prediction accuracy.

【Key words】aging population; genetic algorithm; support vector regression model

0 引 言

随着科学技术的迅速发展，人民生活水平不断提高，人的寿命正不断延长。据国家统计局发布，中国平均预期寿命已由 2005 年的 72.95 岁增长到 2015 年的 76.34 岁。与此同时，中国人口出生率由 2001 年的 13.38%下降到 2019 年的 10.48%（如图 1）。为尽可能减小老龄化社会带来的负面影响，有必要对老龄化人口进行精准预测。本文利用支持向量回归模型（SVR）进行预测，但由于 SVR 的预测能力在很大程度上受到初始参数的影响，参数不同可能会导致欠拟合或过拟合的问题。因此，本文使用具有强大全局搜索能力的遗传算法（GA），对 SVR 的初始参数进行优化，进一步提高其预测准确率。

1 SVR 原理

支持向量机回归，是由支持向量机模型[1]衍生出来的，其目的是发现非线性回归问题中存在的自变量与因变量之间的关系。通过引入非线性映射函数，将低维空间中具有非线性回归关系的数据集映射到高维空间，然后对其进行线性回归关系变换[2]。其数学表示为：

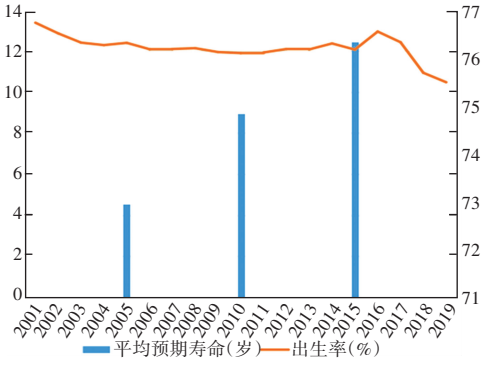

图 1 2001-2019 年平均预期寿命与出生率

Fig. 1 Life expectancy and birth rate from 2001 to 2019

基金项目：吉林省教育厅科学技术研究项目（JJKH20210457KJ）；吉林省大学生创新创业训练项目（2020JLSFDX-JSJ03）；赛尔网下一代互联网技术创新项目（NGII20180315）。

作者简介：张馨予（1996-），女，硕士研究生，主要研究方向：机器学习与无线网络；孙宏宇（1986-），女，博士，讲师，主要研究方向：无线网络与智能计算；逯 洋（1979-），女，博士，教授，主要研究方向：机器学习与数值模拟；郭天岚（1999-），女，本科生，主要研究方向：无线网络与智能计算。

通讯作者：孙宏宇 Email：hongyu@jlnu.edu.cn.

收稿日期：2021-05-25

$$\min \frac{1}{2}\sum_{i,j=1}^{l}(\overset{*}{\alpha}_i-\alpha_i)(\overset{*}{\alpha}_j-\alpha_j)K(x_i,x_j)+\varepsilon\sum_{i=1}^{l}(\overset{*}{\alpha}_i-\alpha_i)-\sum_{i=1}^{l}(\overset{*}{\alpha}_i-\alpha_i), \quad (1)$$

$$\text{s.t.}\begin{cases}\sum_{i=1}^{l}(\overset{*}{\alpha}_i-\alpha_i)=0,\\ 0\leqslant\alpha_i,\alpha_i\leqslant\dfrac{C}{l},\\ i=1,2,\cdots,l.\end{cases}$$

其中，训练样本为 x_i，$i=1,2,3,4,\cdots,l$；K 为核函数；C 为惩罚函数，C 越大表示对误差 ε 的惩罚越大。

若最优解为 $\alpha=(\bar{\alpha}_1,\overset{*}{\bar{\alpha}}_1,\cdots,\bar{\alpha}_l,\overset{*}{\bar{\alpha}}_l)^{\mathrm{T}}$，则支持向量机回归的决策函数为：

$$f(x)=\sum_{i=1}^{l}(\overset{*}{\bar{\alpha}}_i-\bar{\alpha}_i)K(x_i,x)+\bar{b}. \quad (2)$$

2 GA-SVR 模型构建

遗传算法(GA)[3]在 1975 年由 Holland 等人提出，随后 Goldberg[4]与 DeJong[5]等人将遗传算法归纳为一种模拟自然界生物遗传和进化的随机搜索智能算法，适用于复杂系统的优化问题。与其它传统的搜索算法不同，GA 算法并不是基于单一评估函数的较高次统计或梯度产生的确定性的实验解序列，而是通过模拟生物的进化过程来搜索最优解。GA 算法适用于解决大部分优化问题，随着算法的迅速发展，其影响范围也越来越大。

本文引入遗传算法，解决 SVR 的预测能力依赖于初始参数的问题，利用遗传算法的全局搜索能力，搜索支持向量机的最优参数，其中包括：惩罚函数 (C)、$gamma(g)$。工作流程如图 2 所示。

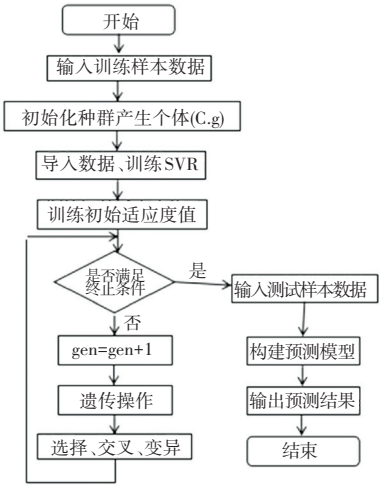

图 2 GA 优化 SVR 初始参数流程图

Fig. 2 GA optimization flow chart of SVR initial parameters

实现步骤如下：

(1)设置遗传算法相关参数。

(2)将支持向量机回归模型的惩罚函数 (C) 与 $gamma(g)$ 进行二进制编码，产生遗传算法初始种群。

(3)将随机产生的数据输入到支持向量机回归模型中，进行交叉验证得到平均准确率，作为遗传算法的目标函数。

(4)进行交叉、变异、选择等操作。

(5)判断是否满足终止条件，满足则输出结果，否则转到步骤(4)。

(6)将最终结果输入到向量机回归模型中，从而得到参数优化后的向量机回归模型，用于进行老龄化人口预测。

3 实验及结果分析

3.1 数据集

研究人口老龄化问题需要相关指标进行实证分析，本文最终选取了 7 个与人口老龄化相关的因素指标：15 ~ 64 岁人口数 - x_1(万人)、出生率 - x_2(%)、死亡率 - x_3(%)、人口自然增长率 - x_4(%)、居民消费水平 - x_5(元)、人均 GDP - x_6(元)，离退人员参加养老保险人数 - x_7(万人)，见表 1。

表 1 2001-2019 年中国人口老龄化预测影响因素表

Tab. 1 Influencing factors of population aging prediction in China from 2001 to 2019

年份	x_1	x_2	x_3	x_4	x_5	x_6	x_7
2001	89 849	13.38	6.43	6.95	3 968	8 717	3 380.6
2002	90 302	12.86	6.41	9.45	4 270	9 506	3 607.8
2003	90 976	12.41	6.4	6.01	4 555	10 666	3 860.2
2004	92 184	12.29	6.42	5.87	5 071	12 487	4 102.6
2005	94 197	12.40	6.51	5.89	5 688	14 368	4 367.5
2006	95 068	12.09	6.81	5.28	6 319	16 738	4 635.4
2007	95 833	12.10	6.93	5.17	7 454	20 494	4 953.7
2008	96 680	12.14	7.06	5.08	8 504	24 100	5 303.6
2009	97 484	11.95	7.08	4.87	9 249	26 180	5 806.9
2010	99 938	11.90	7.11	4.79	10 575	30 808	6 305.0
2011	100 283	11.93	7.14	4.79	12 677	36 302	6 826.2
2012	100 403	12.10	7.15	4.95	14 110	39 874	7 445.7
2013	100 582	12.08	7.16	4.92	15 653	43 684	8 041.0
2014	100 469	12.37	7.16	5.21	17 316	47 173	8 593.4
2015	100 361	12.07	7.11	4.96	18 976	50 237	9 141.9
2016	100 260	12.95	7.09	5.86	20 938	54 139	10 103.4
2017	99 829	12.43	7.11	5.32	23 131	60 014	11 025.7
2018	99 357	10.94	7.13	3.81	25 427	66 006	11 797.7
2019	98 910	10.48	7.14	3.34	27 702	70 581	12 310.4

数据来源：中国统计年鉴

3.2　实验及结果

由于指标数量过多，各指标之间的数据级差异会对预测模型产生影响，因此将表 1 的数据分别采用公式(3)和公式(4)的方法将数据统一到[0,1]区间内，从而保证指标的一致性。

$$x_k = \frac{x_k - x_{\min}}{x_{\max} - x_{\min}}, \tag{3}$$

$$x_k = \frac{x_{\max} - x_k}{x_{\max} - x_{\min}}. \tag{4}$$

式中：x_k 为第 k 个指标归一化后的值，$x_{\min}$ 与 $x_{\max}$ 分别为指标所在列的最小值与最大值。

归一化后的 7 个指标 $x = (x_1, x_2, \cdots, x_7)$ 作为 SVR 模型的自变量，将 65 岁以上人口(万人)作为因变量，从研究总体中选择 2001～2003 年、2005～2006 年、2008～2012 年和 2014～2018 年共 15 组数据作为训练集，将 2004 年、2007 年、2013 年和 2019 年的数据作为测试集。采用遗传算法，求得 SVR 模型的最优参数组合 C 和 g，由于各指标与输入输出量之间是非线性关系，而径向基(*RBF*)核函数适用于解决非线性关系的问题。因此，本文利用 *RBF* 作为核函数，实现非线性映射，最后将最优参数组合输入到 SVR 模型中。参数寻优适应度曲线如图 3 所示。

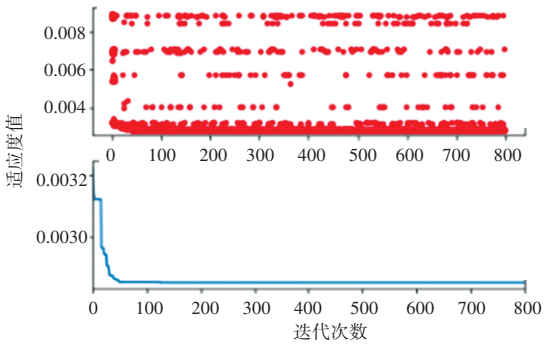

图 3　参数寻优适应度曲线

Fig. 3　Fitness curve of parameter optimization

经过迭代后，SVR 最优参数值为：C = 279.569 307 539，g = 0.000 354 796 650 669，此时训练集的适应度值(均方误差)为 0.006 769 237 246 58。适应度曲线反映出每一代群体的最佳适应度和平均适应度的进化过程。从图 3 中可以看出，随着迭代次数的增加，在后期基本达到了稳定的适应度值，收敛性能较好。

使用初始参数优化后的 SVR 对训练集数据进行拟合，将拟合后的结果与真实值进行比较，与传统 SVR 算法进行比较，实验结果见表 2。

表 2　模型评价

Tab. 2　Model evaluation

算法	*MAE*	*MSE*	*EVS*	*R2*
SVR	0.198 300 740 909	0.050 774 866 967 8	0.957 390 686 186	0.949 225 133 032
本文算法	0.071 555 222 018 6	0.006 769 237 246 58	0.993 385 249 018	0.993 230 762 753

注：*MAE* 为计算的平均绝对误差；*MSE* 为均方误差；*EVS* 为可解释方差值；*R2* 为 *R* 方值。

从表 2 可以看出，本文提出的算法与传统 SVR 相比，平均绝对误差和均方误差更接近于数值 0，可解释方差值和 R 方值更接近于数值 1。将测试集数据输入到训练好的本文构建的模型中，与传统 SVR 算法作对比测试结果见表 3。

从表 3 可以看出，本文算法的预测值与真实值差值较传统 SVR 预测差值相比较小。

实验结果表明：通过遗传算法优化初始参数的 SVR 更加精确；本文提出的使用遗传算法优化支持向量机回归初始参数具有可实施性。

表 3　本文算法与传统 SVR 预测结果比较

Tab. 3　Comparison of prediction results between the proposed algorithm and traditional SVR

预测年份/年	真实值	本文算法预测值	本文算法预测差值	传统 SVR 预测值	传统 SVR 预测差值
2004	9 857	9 756.909 278 49	100.090 721 5	9 132.083 576 63	727.916 423 4
2007	10 636	10 744.492 188 69	−108.492 188 7	10 766.633 441 15	−130.633 441 1
2013	13 161	13 560.196 987 81	−399.196 987 8	13 682.851 388 58	−521.851 388 6
2019	17 603	17 345.672 082 14	257.327 917 9	16 582.930 935 56	1 020.069 064

4　结束语

针对支持向量回归模型的分类精度依赖于初始参数选择的问题，本文设计了一种利用遗传算法优化支持向量回归模型初始参数的方法，并应用于中国老龄化人口预测问题上。实验结果表明，本文提出的优化方法的预测值优于传统SVR算法，为老龄

(下转第 150 页)