文章编号：2095-2163(2021)09-0012-05 中图分类号：F224 文献标志码：A

基于 Lasso 回归和 SVR 模型的消费者信心指数的预测

顾艳文，刘媛华
（上海理工大学 管理学院，上海 200093）

摘 要： 消费者信心指数是反映消费者消费趋向的重要指标，为了掌握消费者信心指数的发展趋势，本文对消费者信心指数进行预测。首先，构建与消费者信心指数相关的关键词作为其影响指标；其次，收集 2011～2019 年的百度指数数据，并采用 Lasso 回归等方法对变量进行筛选；最后，建立 SVR 模型进行预测，并比较在使用不同核函数时，SVR 模型的预测效果。结果表明，使用高斯核函数时，SVR 模型的预测效果最好，能够较好地预测消费者信心指数，从而为有关部门政策的制定提供参考。

关键词： 消费者信心指数；百度指数；Lasso 回归；核函数；SVR 模型

Prediction of consumer confidence index based on Lasso regression and SVR model

GU Yanwen, LIU Yuanhua
(Business School of University of Shanghai for Science and Technology, Shanghai 200093, China)

【Abstract】 Consumer confidence index is an important indicator reflecting consumer consumption trends. In order to grasp the development trend of the consumer confidence index, the paper predicts the consumer confidence index. First, keywords related to the consumer confidence index are constructed as its impact indicators. Secondly, Baidu index data from 2011 to 2019 is collected and methods such as Lasso regression are used to filter the variables. Finally, the an SVR model is used to predict the index and the comparison is conducted about the prediction effect of the SVR model when using different kernel functions. The results show that the SVR model has the best predictive effect when using the Gaussian kernel function, and can better predict the consumer confidence index, thereby providing a reference for the formulation of policies by relevant departments.

【Key words】 consumer confidence index; Baidu index; Lasso regression; kernel function; SVR

0 引 言

中国的消费结构不断升级，消费亮点纷纷涌现，使得消费逐渐成为中国经济增长的主引擎。为了应对国内外动荡的经济形势，构建以国内大循环为主体的新发展格局，需要进一步加强消费对经济的拉动作用[1]。消费与消费者信心息息相关，增强消费的重要举措就是增强消费者信心。消费者信心指数是用来衡量消费者信心的指标，其反映了消费者对当前经济发展状况和未来经济发展预期的内心想法，科学有效的把握消费者信心指数的发展趋势，有助于了解消费者内心的真实感受，对有关部门制定宏观政策，促进经济健康发展具有重要意义。

消费者信心指数的获取通常是通过调查问卷的形式，但传统调查问卷的方法存在工作量大，时效性差，覆盖不全面等问题，所以国内外学者纷纷针对消费者信心指数进行预测研究。一些学者采用传统计量经济学模型，如杨娜、王静雅利用 ARIMA 模型预测消费者信心指数[2]；董现垒、Bollen Johan、胡蓓蓓利用谷歌趋势建立计量经济学模型，对消费者信心指数进行预测[3]；刘伟江、李映桥以网络搜索数据为基础，利用主成分分析法合成搜索指数，建立回归模型，预测台湾地区的消费者信心指数[4]。由于传统计量经济学模型通常适用于线性关系的情况，而消费者信心指数与变量之间的关系复杂多样，因此一些学者提出采用机器学习模型或者深度学习模型对其进行预测，如邹鸿飞、王建州建立了 CEEMD-DEGWO-BPNN 模型预测消费者信心指数[5]；唐晓彬、董曼茹、张瑞引入百度指数数据，建立长短时间记忆神经网络模型进行消费者信心指数的预测[6]。

Hanjo Odendaal、Monique Reid、Johann F. Kirsten 认为在线情感指数对消费者信心指数具有预测作用，可为消费者信心指数的预测提供思路[7]。在以往的研究中，预测消费者信心指数所使用的影响因素也常为非结构化数据，然而非结构化数据的数据量较大，不能全部放入预测模型中建模，需要对变量进行筛选。本文采用对数据类型没有太多限制，且可以弥补最小二乘法和逐步回归法局部最优估计不

作者简介：顾艳文(1997-)，女，硕士研究生，主要研究方向：统计学；刘媛华(1974-)，女，博士，副教授，主要研究方向：系统工程、经济与管理统计、复杂系统理论方法及应用。

收稿日期：2021-07-15

足的 Lasso 回归对变量进行处理，同时采用既可以解决线性关系问题又可以解决非线性关系问题的机器学习模型——支持向量机回归，对消费者信心指数进行预测。

1　模型理论概述

1.1　Lasso 回归

当数据特征较多时，为了防止模型的过拟合，常常需要对数据进行筛选降维。1996 年国外学者 Robert Tibshirani 提出了 Lasso 回归。Lasso 回归是一种缩减性估计，在回归过程中，可以将一些不重要的回归系数直接缩减为 0，以此实现变量筛选的功能。Lasso 回归可以降低模型训练时的计算量，因此在高维数据中得到广泛应用。Lasso 回归的目标函数为式(1)：

$$J(\beta)=\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| \tag{1}$$

其中，λ 是惩罚项系数，控制着模型的复杂程度，λ 越大对特征较多的模型惩罚力度越大，通过调整 λ，最终可以获得特征较少的模型，以达到降维的目的。

1.2　SVR 模型

SVR 模型又称支持向量回归模型，其采用支持向量的思想，可将低维数据非线性映射到高维空间，从而在高维空间中对数据进行回归分析。支持向量回归模型的优点在于模型对数据的分布没有限制，可以有效解决小样本、非线性、高维度问题。SVR 模型的目标函数为式(2)：

$$\begin{cases}\min_{w,b,\xi_i,\hat{\xi}_i}\dfrac{1}{2}\|w\|^2+C\sum_{i=1}^{n}(\xi_i+\hat{\xi}_i)\\ y_i-f(x_i)\leqslant\varepsilon+\xi_i\\ f(x_i)-y_i\leqslant\varepsilon+\hat{\xi}_i\\ \xi_i\geqslant 0,\hat{\xi}_i\geqslant 0\end{cases} \tag{2}$$

其中，$f(x_i)=w^{'}x_i+b$，ξ_i 和 $\hat{\xi}_i$ 是松弛因子。

支持向量回归允许预测值和实际值之间存在一个合理的误差，即 $|y_i-f(x_i)|\leqslant\varepsilon$。根据拉格朗日函数的对偶性和极小值求解的方法，可以得到 $f(x_i)$ 中参数 w 与 b 的值，式(3)：

$$\begin{cases}w=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)x_i\\ b=y_i+\varepsilon-\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)x_i x_j\end{cases} \tag{3}$$

其中，α_i 和 α_i^* 是拉格朗日乘子。

为了使模型能够解决非线性回归问题，引入核函数 $K(x_i,x_j)$ 替换高维空间的内积，此时函数 $f(x_i)$ 可以表示为式(4)：

$$f(x_i)=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)K(x_i,x)+y_i+\varepsilon-\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)K(x_i,x_j) \tag{4}$$

SVR 模型对核函数的选择比较敏感，不同的核函数会使模型产生不同的结果。常用的核函数有多项式核函数(*ploy* 核函数)、高斯核函数(*rbf* 核函数)、*Sigmoid* 核函数等，通过网格搜索的方法可以确定核函数的参数，从而使模型达到最好的效果。

1.3　消费者信心指数预测模型

由于 Lasso 回归的降维能力和 SVR 模型的优点，本文结合两个模型对消费者信心指数进行预测。首先，对数据进行预处理，提高数据质量；其次，对变量进行领先期数的确定，使选取的变量具有预测能力；然后利用相关系数选取与消费者信心指数相关的变量，再将新得到的数据集输入 Lasso 回归模型中降维，从而得到最终的预测变量；最后，把变量放入 SVR 模型中进行消费者信心指数的预测，并比较使用不同核函数模型的预测效果，从而确定最终的预测模型。消费者信心指数预测模型的构建思路如图 1 所示。

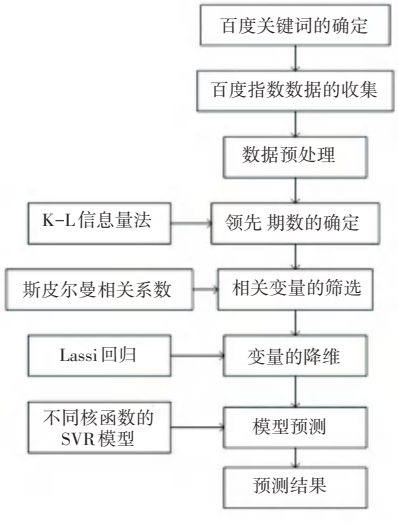


图 1　消费者信心指数预测模型构建思路

Fig. 1　Thoughts on constructing consumer confidence index forecast model

2　消费者信心指数预测分析

2.1　数据来源

近年来，互联网快速发展，出现大量的非结构化

数据,这些非结构化数据往往与经济现象之间存在某种联系,或多或少反映着真实的经济生活。因此,本文采用非结构化数据中的百度指数数据作为预测消费者信心指数的数据支撑,并通过文献参考和需求图谱的关键词推荐,选取了 133 个百度关键词,部分关键词见表 1。百度指数数据分为移动端和 PC 端,而移动端的百度指数数据从 2011 年开始收录,故本文的数据从 2011 年开始收集,通过爬虫技术获取 2011~2019 年的 PC 端和移动端的百度指数。本文的研究对象为消费者信心指数,为保持数据的一致,选取了 2011~2019 年的月度数据作为本文的样本,其数据来源于中经网统计数据库。

表 1 部分关键词

Tab. 1 Some keywords

类型	部分关键词
收入状况	个人收入 个人所得税 个税改革 收入分配 收入差距等(共 10 个)
生活质量	幸福感 满足感 福利 电影 旅游 健康 休闲 网购等(共 21 个)
经济形势	GDP GNP PPI 进出口 国际贸易 人民币汇率 美元指数等(共 39 个)
消费支出	物价 猪肉价格 鸡蛋价格 粮食价格 汽油价格等(共 33 个)
就业情况	就业率 找工作 招聘 兼职 实习 58 同城等(共 14 个)
房屋与汽车	安居客 二手房 链家 新房 房地产 汽车配置 车展(共 8 个)
储蓄与投资	股票 证券 基金 黄金 保险 理财 债券 期货 (共 8 个)

2.2 数据预处理

百度指数数据是非结构化数据,可能会受到各种各样的干扰,存在噪声较大的问题,需要对其进行预处理。

第一步:异常值处理。百度指数数据会受到特殊事件的影响,导致出现异常值,而异常值会影响模型的预测效果,故需要对异常值进行处理。本文采用箱线图法对异常值进行判断,将筛选出的异常值用前后两期的均值进行替换。

第二步:去除长期趋势。随着近些年来互联网的高速发展,搜索引擎的使用频率也会随着时间的增加而增加,为了消除由于互联网发展导致搜索量的增加,需要寻找与本文研究对象相关性不大,且能代表互联网发展趋势的关键词[8]。因此计算选取的 133 个关键词与其百度指数的比值,以消除互联网长期发展趋势。通过参考相关文献,本文选取的关键词为百度。

第三步:合并数据。由于消费者信心指数为月度数据,故将百度指数的日度数据转为月度数据。

2.3 预测模型的建立

2.3.1 基于 Lasso 回归模型的变量的降维

本文选取的 133 个百度关键词并非都适合放入模型中作为变量进行预测,需要对其进行筛选。首先,通过 K-L 信息量法确定每个关键词的最佳阶数,将关键词领先阶数设为 1~12 阶,计算每个关键词领先 1~12 阶的 K-L 信息量,并从中选取 K-L 信息量最小值所对应的阶数作为该关键词的最佳阶数,根据最佳阶数将原始数据错位补齐;其次,计算错位补齐后的每个关键词和消费者信心指数之间的斯皮尔曼相关系数,并将阈值设为 0.5,以此获得 43 个与消费者信心指数相关的关键词;最后,为了进一步减少模型的输入变量,提高模型的预测效果,建立 Lasso 回归模型对 43 个百度关键词进行筛选。

Lasso 回归模型中的 λ 值是未知的,可以通过可视化方法大致确定 λ 的取值范围,然后通过交叉验证法确定最终的 λ 值。

λ 和回归系数之间的关系如图 2 所示,每条折线图代表了每个变量。从图 2 可知,当 λ 的值大概在 0.02~0.76 之间时,绝大多数变量的回归系数趋于稳定。为确定准确的 λ 值,利用 sklearn 模块中的 LassoCV 类进行交叉验证,对每一个 λ 值,进行 10 重交叉验证,从而确定 λ 的值为 0.141。以最佳 λ 值重新建立 Lasso 回归模型,最终筛选出 6 个百度关键词,分别为股票、赶集网、58 同城、民宿、大众点评和个人所得税。表 2 是最终百度关键词的滞后阶数及斯皮尔曼相关系数。

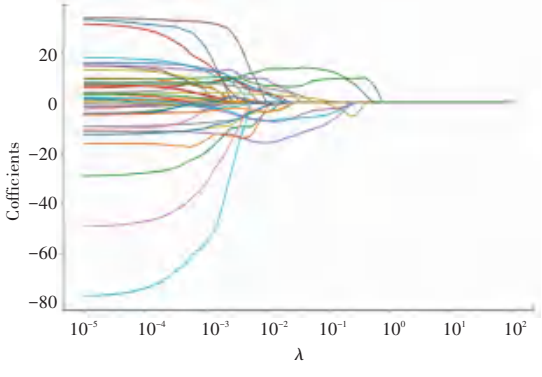


图 2 λ 与回归系数的关系

Fig. 2 Relation between λ and regression coefficient

2.3.2 SVR 预测模型

经过上述处理和变量筛选后,还剩余 96 期数

据。将数据集按照 7:3 的比例划分训练集和测试集，并对其进行归一化处理，以消除不同数量级造成的影响。由于 SVR 模型的预测效果受核函数的影响较大，所以本文选取常用的多项式核函数高斯核函数，*Sigmoid* 核函数进行建模，并采用网格搜索的方法对核函数参数、惩罚系数、损失函数参数进行寻优。SVR 模型使用不同核函数的最终参数值见表 3。

表 2　最终百度关键词

Tab. 2　Final Baidu keywords

百度关键词	滞后阶数	斯皮尔曼相关系数
股票	−4	−0.608
赶集网	−12	−0.824
58 同城	−12	−0.534
民宿	−1	0.763
大众点评	−12	−0.548
个人所得税	−1	0.644

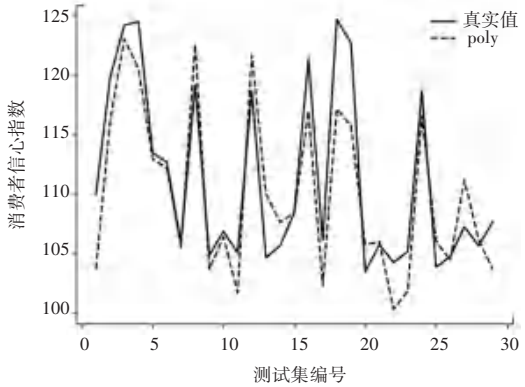
表 3　模型参数值

Tab. 3　Model parameter values

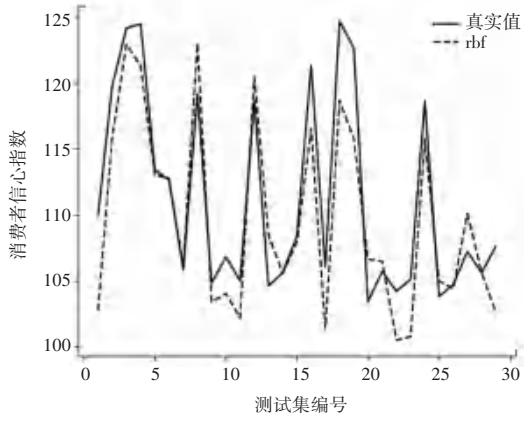
核函数	惩罚系数	损失参数	核半径	偏执系数	多项式阶数
多项式核函数	90	0.001	0.1	0.03	1
高斯核函数	20	0.01	0.466	–	–
Sigmoid 核函数	11	0.009	0.42	0.09	–

根据网格搜索法得到的参数值，分别建立 SVR 模型，并对测试集进行预测，不同核函数预测结果如图 3 所示。

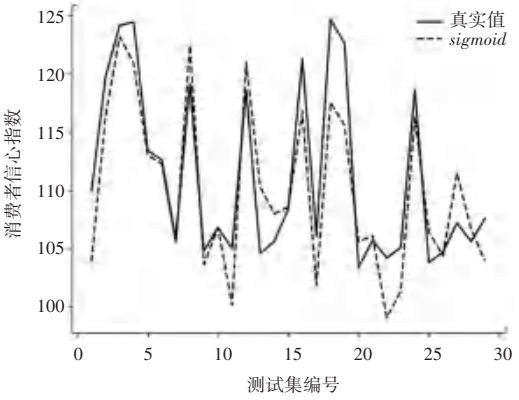
由图 3 可知，无论使用多项式核函数，高斯核函数还是 *Sigmoid* 核函数都可以对消费者信心指数进行大致的刻画，说明 SVR 模型对消费者信心指数具有一定的预测能力。但不同的核函数之间还存在一定的差异，为了选择更好的模型，对 3 种核函数的预测结果进行定量分析，采用均方根误差和平均绝对误差对其进行评价，评价结果见表 4。



（a）多项式核函数



（b）高斯核函数



（c）*sigmoid* 核函数

图 3　不同核函数真实值和预测值对比

Fig. 3　Comparison of real and predicted values of different kernel functions

表 4　不同核函数预测结果

Tab. 4　Forecast results of different kernel functions

评价指标	多项式核函数	高斯核函数	*Sigmoid* 核函数
RMSE	3.460	3.441	3.573
MAE	2.776	2.756	2.880

由表 4 可知，多项式核函数和 *Sigmoid* 核函数的预测效果不如高斯核函数，当模型使用高斯核函数时，模型的均方根误差和平均绝对误差最小，分别为 3.441 和 2.756；其次是多项式核函数，均方根误差为 3.460，平均绝对误差为 2.776；预测结果最差的是 *sigmoid* 核函数，均方根误差为 3.573，平均绝对误差为 2.88。

3　结束语

本文以非结构化数据中的百度关键词作为消费者信心指数的影响因素，将 Lasso 回归和 SVR 模型相结合，对消费者信心指数进行预测。同时，通过对比不同的核函数，认为在使用高斯核函数时，可以使

（下转第 21 页）