

基于网络搜索数据的 GDP 组合预测研究

王书平, 卢子晗, 冀承秀

(北方工业大学经济管理学院, 北京 100144)

摘要: 网络搜索数据 (Web Search Data, WSD) 是研究宏观经济现象的重要微观信息依据。从需求、供给与政策环境等方面选取和筛选关键词来合成网络搜索指数, 采用金枪鱼群 (Tuna Swarm Optimization, TSO) 算法优化的最小二乘支持向量回归 (Least Squares Support Vector Regression, LSSVR) 模型, 对 GDP 进行预测。结果表明, 网络搜索指数与 GDP 具有强相关性, 合成的网络搜索指数能较好地反映 GDP 的波动走势; 网络搜索数据的加入使得预测结果呈现出强时效性, 预测效果及预测精度都取决于对最优模型的选择, 引入参数智能优化算法可提高模型的预测性能。提出的 TSO-LSSVR&WSD 模型充分利用网络搜索数据及组合预测优势, 提高了 GDP 的预测精度和时效性, 可应用于宏观经济指标预测中。

关键词: GDP 预测; 组合预测; 网络搜索数据; 金枪鱼群算法; LSSVR 模型

中图分类号: F726; O29 文献标志码: A 文章编号: 1674-8646(2024)08-0044-05

Research on GDP Combined Forecasting Based on Web Search Data

Wang Shuping, Lu Zihan, Ji Chengxiu

(School of Economics and Management, North China University of Technology, Beijing 100144, China)

Abstract: Web Search Data (WSD) is an important micro-information basis for studying macroeconomic phenomena. In this paper, keywords are selected and screened from the aspects of demand, supply and policy environment to synthesize the web search index, and the Least Squares Support Vector Regression (LSSVR) model optimized by the Tuna Swarm Optimization (TSO) algorithm is used to predict GDP. The results show that the web search index has a strong correlation with GDP, and the synthetic web search index can better reflect the fluctuation trend of GDP. The addition of web search data makes the prediction results show strong timeliness, the prediction effect and prediction accuracy depend on the selection of the optimal model, and the introduction of parameter intelligent optimization algorithm can improve the prediction performance of the model. The proposed TSO-LSSVR & WSD model makes full use of the advantages of web search data and combined forecasting to improve the accuracy and timeliness of GDP forecasting, and can be applied to the field of macroeconomic index forecasting.

Key words: GDP forecasts; Combined forecasting; Web search data; Tuna swarm optimization algorithm; LSSVR model

0 引言

网络搜索数据是近年来新兴的一种能够准确记录网民在互联网上搜索活动的指数化数据源。截至 2022 年 12 月, 我国网民为 10.67 亿人, 互联网普及率达 75.6%, 可以看出, 互联网用户是我国目前经济活动中的主要参与者, 也是宏观经济的微观基础, 网络搜索量则是他们把信息当作决策基础的一种数据表现。基于互联网数据实时更新的特点, 将这些网络搜索数据利用起来可及时有效地了解经济运行状况, 分析我

国所处的经济周期状态。其作为一种新型的数据资源, 还可为国家发展和其他各领域研究提供思路。

GDP(即国内生产总值)是体现一国宏观经济运行情况的重要指标。ARIMA 模型^[1]、灰色预测模型^[2]等线性预测方法是国内外学者预测 GDP 时最为常用的方法。随着我国市场经济的不断发展, 机器学习提供了一种新的非线性建模思路, 其中反向传播神经网络中的人工神经网络^[3]、极限学习机^[4]和长短期记忆网络模型^[5]的应用最为广泛。而传统模型也具有一定作用, 以此为基础, 将不同模型结合在一起实现各自的优势互补可在一定程度上扩大模型的适用范围, 提升预测精度。谢成兴等^[6]基于喀什地区 2000—2018 年的 GDP 建立了 ARIMA-DGM-BP 组合预测模型, 结果显示, 三因素组合预测模型最优。目前, 预测 GDP 大多是以传统的宏观数据为主, 统计时效性不强, 会造成数据发布的延迟。

近年来, 学者们基于网络搜索数据, 聚焦于消费者

收稿日期: 2023-12-20

基金项目: 北方工业大学毓优团队培养计划项目(107051360021XN083/045); 北方工业大学北京城市治理研究基地开放课题(110051360023XN277-02)

通讯作者: 卢子晗(1999-), 女, 硕士研究生。研究方向: 数量经济分析、预测科学。E-mail: luzihantt158@163.com。

信心指数^[7]、CPI^[8]及 GDP^[9]等宏观经济指标的预测,通过与多个基准模型进行对比发现,加入网络搜索数据的模型可以进一步提高解释能力和预测精度。在关键词的选取上,主要涉及需求端、供给端、个人收入、投资策略等方面及其他影响宏观经济的因素。在模型建构上,大多在网络搜索数据的基础上运用时间序列模型、混频数据模型、经算法优化的机器学习等模型预测宏观经济指标。

本研究基于网络搜索数据,采用经 TSO 算法优化的 LSSVR 模型预测 GDP,利用组合预测模型的高精确度和网络搜索数据的时效性来预测 GDP,可及时捕捉到宏观经济的走势变化,为宏观调控的时机选择及力度把握提供决策依据。

1 理论分析

对 GDP 的构成因素分析可从需求侧和供给侧两方面展开。需求侧从消费、投资和净出口入手,三者分别体现内部需求、财政投入和外部需求。供给侧以经济增长理论和柯布 - 道格拉斯生产函数为基础,从劳动力、资本、技术三个方面反映国民经济增长动力。由此可见,需求侧体现 GDP 的静态构成,供给侧体现 GDP 的动态增长,需将两方面结合进行分析。

基于网络搜索数据监测预测并分析社会经济活动的研究在国内外已取得了较多成果,该方法的主要原理是捕捉消费者对微观经济活动的关注行为,通过对蕴含用户意图的关键词搜索数据汇总分析,形成对宏观指标的监测与预测。

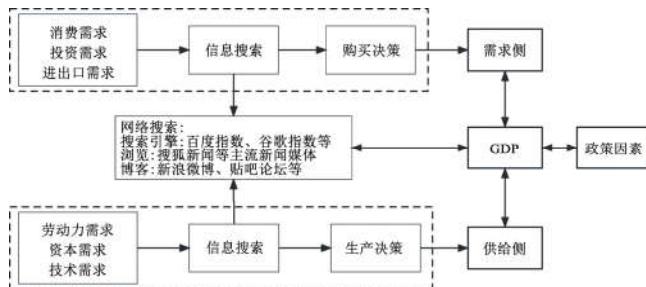


图 1 网络搜索数据与 GDP 关联

Fig. 1 Correlation of Web search data and GDP

宏观经济形势影响着搜索引擎用户对经济活动热点的关注,消费者或生产者在开展经济活动前,需结合宏观经济形势和行业经济特征为实施购买或投资决策提供判断依据。以投资为例,城镇住房价格水平的波动或住房政策调整将吸引投资者在网络上进行搜索活动,了解相关政策和行业形势变化。搜索引擎用户自身对微观经济活动的关注也将产生对宏观经济发展的影响。消费者或生产者通过搜索活动作出的购买或生产决策将影响相关行业的发展动力,从而影响未来走势。

我国经济进入高质量发展阶段,宏观经济增速变

慢,宏观经济结构不断改善。供给侧结构性改革、经济发展双循环等的提出将作为选取政策环境维度关键词时的部分参考依据。

2 研究方法

2.1 主成分分析法

主成分分析法是将众多变量中相关程度较高的变量删除,保留尽可能少的变量,从剩下的众多变量中提取新变量的方法。其基本表达式为。

设 Z_i 表示第 i 个主成分, $i = 1, 2, \dots, p$

$$\begin{cases} Z_1 = c_{11}X_1 + c_{12}X_2 + \dots + c_{1p}X_p \\ Z_2 = c_{21}X_1 + c_{22}X_2 + \dots + c_{2p}X_p \\ \dots \\ Z_p = c_{p1}X_1 + c_{p2}X_2 + \dots + c_{pp}X_p \end{cases} \quad (1)$$

$$\frac{\lambda_i}{\sum_{j=1}^p \lambda_j}, (i = 1, 2, \dots, l) \quad (2)$$

$$\frac{\sum_{j=1}^{nj} \lambda_j}{\sum_{j=1}^p \lambda_j} \quad (3)$$

对于每个 i ,都有 $c_{11}^2 + c_{12}^2 + \dots + c_{ip}^2 = 1$,且($c_{11}, c_{12}, \dots, c_{1p}$),使得 $Var(Z_i)$ 的值达到最大;($c_{21}, c_{22}, \dots, c_{2p}$)不仅垂直于($c_{11}, c_{12}, \dots, c_{1p}$),且使得 $Var(Z_2)$ 的值达到最大,以此类推得到全部 p 个主成分。各主成分的贡献率可由该主成分的特征根和所有主成分的特征根总和的比值确定,表达式如式(2),从第 1 个到第 n 个主成分的累计贡献率可表示为式(3)。

2.2 TSO 算法

TSO 算法是一种基于群的元启发式全局最优算法,由 Xie 等^[10]提出。该算法是受金枪鱼群体合作捕食的启发,通过对金枪鱼的螺旋式捕食和抛物线式捕食两种行为的模拟来进行最优化问题求解,算法结构简单,需要调整的参数少,易于实现。

金枪鱼觅食的第一种策略是螺旋觅食。在觅食过程中,金枪鱼可以随时对自己的方向进行调整,从而形成一个紧密的螺旋结构,对周围的猎物进行追逐,最优个体在追逐猎物的过程中可以与鱼群进行交互,实现信息共享。螺旋觅食时的位置更新如式(4)所示。

$$\begin{cases} X_i^{t+1} = \alpha_1 \cdot (X_{rand}^t + \beta \cdot |X_{rand}^t - X_i^t|) + \alpha_2 \cdot X_i^t, & i=1, \text{if } rand < \frac{t}{t_{max}} \\ \alpha_1 \cdot (X_{rand}^t + \beta \cdot |X_{rand}^t - X_i^t|) + \alpha_2 \cdot X_{i-1}^t, & i=2,3,\dots,NP \\ \alpha_1 \cdot (X_{best}^t + \beta \cdot |X_{best}^t - X_i^t|) + \alpha_2 \cdot X_i^t, & i=1, \text{if } rand \geq \frac{t}{t_{max}} \\ \alpha_1 \cdot (X_{best}^t + \beta \cdot |X_{best}^t - X_i^t|) + \alpha_2 \cdot X_{i-1}^t, & i=2,3,\dots,NP \end{cases} \quad (4)$$

其中, X_i^{t+1} 表示 $t+1$ 次迭代时的第 i 个个体, X_{best}^t 表示当前最佳个体, α_1 表示金枪鱼初始时追随最佳个体的权重系数, 如式(5)所示, α_2 表示跟随前一个个体的趋势, 如式(6)所示, t_{max} 表示最大迭代次数, b 表示均匀分布在 $[0,1]$ 的常数。

$$\alpha_1 = \delta + (1 + \delta) \cdot \frac{t}{t_{max}} \quad (5)$$

$$\alpha_2 = (1 - \delta) + (1 - \delta) \cdot \frac{t}{t_{max}} \quad (6)$$

金枪鱼的第二种觅食策略是抛物线觅食, 除了螺旋觅食外, 金枪鱼还会以自己的食物作为参照点, 以一种抛物线的形状捕食, 并会在自己的附近捕食, 假设在 50% 的概率下这两种捕食模式同时进行, 如式(7)所示。

$$X_i^{t+1} = \begin{cases} X_{best}^t + rand \cdot (X_{best}^t - X_i^t) + TF \cdot p^2 \cdot (X_{best}^t - X_i^t), & \text{if } rand < 0.5 \\ TF \cdot p^2 \cdot X_i^t, & \text{if } rand \geq 0.5 \end{cases} \quad (7)$$

其中, TF 表示种群开发方向的随机数, 取值范围为 $[-1, 1]$, p 表示影响种群开发幅度的参数, 计算公式如式(8)所示。

$$p = (1 - \frac{t}{t_{max}})^{\frac{t}{t_{max}}} \quad (8)$$

2.3 LSSVR 模型

LSSVR 模型适合于小样本数据的预测, 通过非线性映射函数 $\varphi(\cdot)$ 将输入空间映射到高维特征空间, 从而实现非线性回归。根据最小化结构风险来进行模型选择, 可有效避免过拟合问题, 从而提高模型的泛化能力。LSSVR 是存在约束条件的优化问题。

$$\begin{cases} \min J(\omega, e) = \frac{1}{2} \|\omega\|^2 + \frac{1}{2} \gamma \sum_{i=1}^l e_i^2 \\ \text{s. t. } y = \omega^T \varphi(x_i) + b + e_i \end{cases} \quad (9)$$

式中 ω 表示权值向量, e 表示误差变量, γ 表示正则化参数, 用来评价拟合曲线的平滑程度, 实现最小拟合误差, 正则化参数过高或过低都会影响模型的泛化能力。LSSVR 建立的回归函数为:

$$y(x) = \sum_{i=1}^l \alpha_i k(x, x_i) + b \quad (10)$$

式中 α_i 表示 Lagrange 乘子, b 表示偏置向量, $k(x, x_i)$ 表示核函数。选取拟合精度较高且适用于小样本的高斯函数为核函数。

$$k(x_i, x_j) = \exp \left(-\frac{\|x_i - x_j\|^2}{2\sigma^2} \right) \quad (11)$$

式中 σ 为核函数宽度。如果核函数宽度过小, 可能会出现过拟合的现象。过高的数值则会降低模型的泛化能力。为改善 LSSVR 预测效果, 利用 TSO 算法寻找两个适当的参数, 具体步骤如图 2 所示。

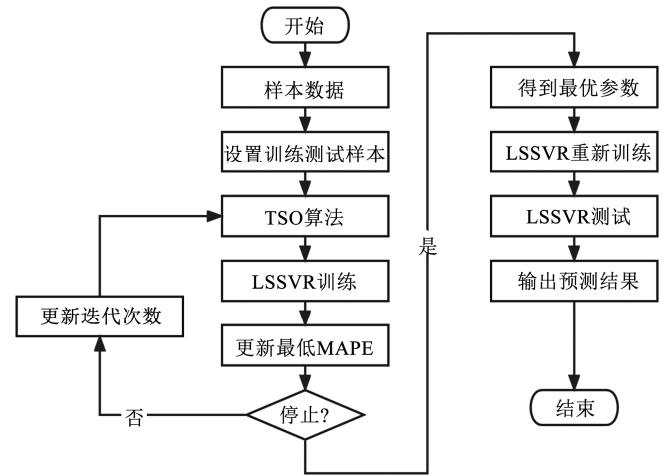


图 2 TSO 算法优化 LSSVR 参数的过程

Fig. 2 Optimization of LSSVR parameter process with TSO algorithm

3 实证分析

3.1 数据来源与预处理

采用的数据主要分为两部分, 一部分是我国季度 GDP 值, 选自国家统计局, 另一部分为网络搜索关键词, 选自百度指数官方网站。为了获取更为全面的网民搜索行为数据, 选取的关键词搜索指数既包括 PC 端搜索指数(2006 年开始), 也包括移动端搜索指数(2011 年开始), 样本跨度为 2011Q1–2023Q1。

由于重要节日、年末绩效等因素, 我国季度 GDP 序列呈锯齿状周期性波动上升趋势, 其基本变动趋势往往会被掩盖, 使经济发展中的其他客观变化要素变得模糊, 故很难对经济规律进行深入研究。使用 X-13A-S 季节调整方法, 对我国季度 GDP 进行季节调整。在 X-13A-S 方法中考虑一些移动假日(如春节、清明节、端午节)的影响, 从而提高季节调整的精度。

针对网络搜索关键词, 主要从以下几方面进行处理, 排除未被百度收录和指数较低的关键词。由于百度指数关键词数据是日度数据, 而国家统计局公布的是 GDP 季度数据, 需要把日搜索数据转换成季度平均搜索数据。为消除短期波动, 采用三期移动平均对百度指数数据进行平滑处理。

3.2 网络搜索指数的合成

(1) 网络搜索关键词的确定。由理论分析可知, 宏观经济发展状况在需求端主要受居民消费、投资和净出口的影响, 在供给端主要受劳动力、资本和技术的影响, 考虑到在数据样本范围内我国步入经济结构调整的新常态发展时期, 故在选择关键词时充分考虑了宏观经济发展的政策条件。从需求、供给与政策三方面选取网络搜索关键词。

(2) 拓展关键词。结合百度指数的关键词推荐功能、需求图谱和长尾词挖掘工具, 逐一输入初始关键

词,将初选关键词扩充到 151 个。

(3)筛选网络搜索关键词。利用时差相关分析法,得到每个关键词与 GDP 之间的时差相关系数,考虑后续建模时样本容量与变量个数的关系,剔除掉与 GDP 相关系数小于 0.7 且滞后于 GDP 数据变化的关键词指标。最终选取关键词 43 个,部分关键词相关分析结果如表 1 所示。

表 1 部分网络搜索关键词时差相关系数

Tab. 1 Correlation coefficient of some web search keyword lag

关键词	时差数	相关系数	关键词	时差数	相关系数
酒	-7	0.726	五险一金	-1	0.768
物业费	-1	0.877	设备	-1	0.794
成本	-1	0.878	生产资料	-1	0.879
茅台	-2	0.812	数据	-1	0.878
就业	-4	-0.734	国家政策	-1	0.826

为解决使用网络搜索数据进行综合指标合成过程中共线性较难消除、指标权重不能合理确定等问题,基于数据降维思想,引入主成分分析方法来构建综合指数。

对原始数据进行 KMO 和 Bartlett 检验,发现 Bartlett 球形度检验的显著性值 $p < 0.001$,表明数据的内关联性很高,有助于进一步开展主成分研究。运用特征值大于 1 的提取方法,提取了 4 个主成分,其方差累计贡献率为 89.887%,即 4 个主成分能表示原来 43 个指标 89.887% 的信息。以每个主成分所对应的特征值和得分矩阵为基础,得出各个主成分和综合主成分值,编制出网络搜索指数 F ,见式(12)。

$$F = 0.7524 \times F_1 + 0.0754 \times F_2 + 0.0404 \times F_3 + 0.0307 \times F_4 \quad (12)$$



图 3 网络搜索指数与季度 GDP 对比

Fig. 3 Comparison of web search index and quarterly GDP

通过比较合成的网络搜索指数 F 和我国各季度 GDP 发现, F 的变化趋势和 GDP 的变化趋势基本一致,经皮尔逊相关系数检验,GDP 和网络搜索指数 F 之间的相关性达到 0.939。

3.3 预测模型的建立与对比

选取 2011Q1 – 2023Q1 的网络搜索指数 F 与 GDP 数据进行研究,共 49 个数据,由于对网络搜索数据进行了三期移动平均处理,数据长度变为 2011Q2 –

2022Q4,共 47 个数据,选取 2011Q2 – 2018Q4 的数据作为训练集用于模型训练,共 31 个数据,2019Q1 – 2022Q4 的数据作为测试集用于预测检验,共 16 个数据。

运用网络搜索数据,利用经 TSO 算法优化的 LSSVR 模型预测 GDP。根据预测步骤对 LSSVR 模型的参数初始化,利用 TSO 算法训练 LSSVR 模型以获取最优参数,将其应用于测试集进行预测。采用 MATLAB 软件开展实验,设置搜索代理的数量,即种群的数量为 5,迭代次数为 50。在模型数据的输入方面,使用 mapminmax 函数对数据进行归一化,加速模型收敛性,改善模型预测效果。在预测完成后,再将其逆标准化,获得与之对应的实际预测值。

为了更直观地展示该模型的预测结果,图 4 显示了该模型在训练集样本范围内的拟合曲线,图 5 显示了该模型在测试集样本范围内的拟合曲线。可以看出,TSO-LSSVR& WSD 组合模型拟合效果较好,预测值与真实值的变动趋势较为一致,说明所选择的模型具有较好的预测性能。经过计算,在训练集中,平均绝对百分比误差 MAPE 为 3.7783%,在测试集中为 5.253%,其值小于 10%,表明无论是在训练集还是在测试集中,TSO-LSSVR&WSD 组合模型预测性能均较高,具有很好的泛化能力。

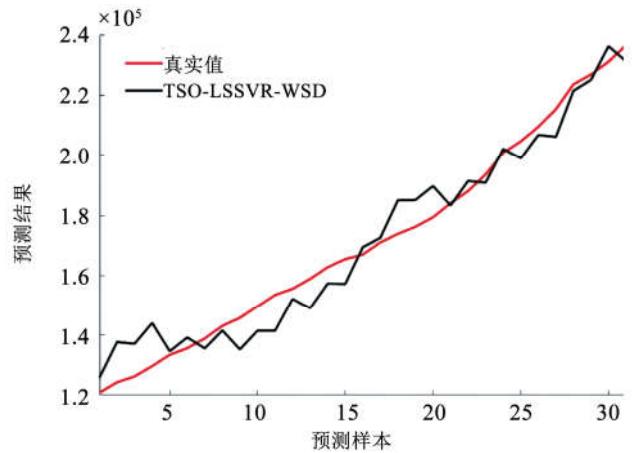


图 4 训练集预测效果对比

Fig. 4 Comparison of the prediction effect of training set

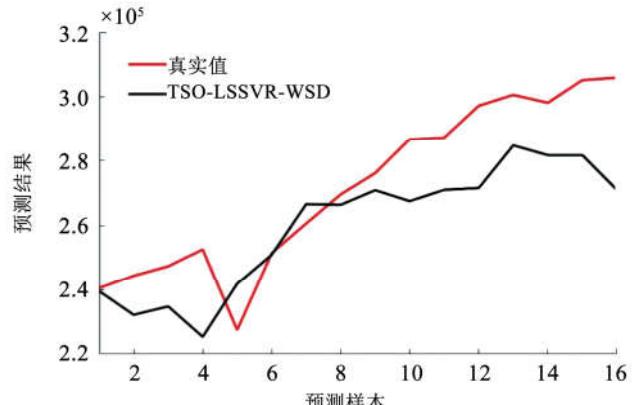


图 5 测试集预测效果对比

Fig. 5 Comparison of the effect of test set prediction

在预测效果判别标准方面,采用了相关系数 R 、平均绝对百分比误差 MAPE 评估模型的预测性能。指标定义如下。

$$R = \frac{\sum_{t=1}^N (y_t - \bar{y})(\hat{y}_t - \bar{\hat{y}})}{\sqrt{(y_t - \bar{y})^2} \sqrt{(\hat{y}_t - \bar{\hat{y}})^2}} \quad (13)$$

$$MAPE = \frac{1}{N} \sum_{t=1}^N \left| \frac{y_t - \hat{y}_t}{y_t} \right| \times 100\% \quad (14)$$

其中, y_t 为真实值, \hat{y}_t 为预测值, N 表示预测期数。

R 对真实值与预测值之间的关联性强弱进行度量, 数值越接近 1 表明关联性越高, 说明模型的预测性能越好。MAPE 反映了真实值和预测值之间的偏差, 值越小说明模型预测精度越高。

表 2 4 个预测模型统计性能指标对比

Tab. 2 Comparison of the statistical performance indicators of 4 prediction models

变量	TSO-LSSVR&WSD	TSO-LSSVR	LSSVR	LSSVR&WSD
R	0.87233	0.76341	0.44187	0.83279
MAPE/%	5.2530	7.7387	12.2124	8.1916

预测结果表明, 在预测我国 GDP 的研究中, 良好的预测结果不仅依赖于网络搜索数据的运用, 还与模型参数的选择息息相关, 构建的 TSO – LSSVR& WSD 组合模型的预测效果是最佳的。

运用不同模型预测得到的统计测量指标见表 2。两种预测性能统计指标的结果表明, TSO – LSSVR&WSD 组合模型的预测性能是最好的。从相关系数 R 来看, TSO – LSSVR&WSD 组合模型 R 值最大, 其预测值与真实值关联性最强。通过对 4 个模型的 MAPE 进行比较发现, TSO – LSSVR&WSD 组合模型的 MAPE 最小。结果表明, 该模型的真实值与预测值之间的偏差程度最小, 预测精度最高。

在单一模型方面, 加入网络搜索数据后的模型即 LSSVR&WSD 模型预测效果要优于未加入网络搜索数据的模型即 LSSVR 模型。在组合模型方面, TSO – LSSVR 模型效果明显优于单一模型 LSSVR, 加入网络搜索数据的 TSO – LSSVR&WSD 使得模型效果进一步提高, 印证了网络搜索数据的加入使得预测结果具有时效性, 引入参数智能优化算法可提高 LSSVR 模型的预测性能。提出的 TSO – LSSVR&WSD 模型充分利用网络搜索数据及组合预测优势, 提高了 GDP 的时效性和预测精度, 可应用于宏观经济指标预测中。

4 结论

从供需端出发, 选取网络搜索关键词并将其与

GDP 进行相关性界定, 筛选后的关键词采用主成分分析法进行降维, 得到网络搜索指数。采用经 TSO 算法优化的 LSSVR 模型对 GDP 进行预测, 将组合预测模型效果与单一模型效果进行比较分析, 得出建立的 TSO – LSSVR&WSD 组合预测模型显著提高了预测效果, 结论如下:

基于国民经济核算与经济增长理论, 从供需端出发, 选取网络搜索关键词, 在一定程度上减少了选词的片面性。网络搜索指数与我国 GDP 之间关系密切, 相关系数高达 0.939, 合成的网络搜索指数能较好地反映 GDP 的波动走势。

利用网络搜索数据使得预测结果更具时效性。通过对比不同预测模型的性能指标发现, 网络搜索数据对预测 GDP 助益, 预测效果和预测精度有赖于对最优模型的选取, 引入参数智能优化算法可改善模型的预测性能, 验证了提出的 TSO – LSSVR&WSD 组合模型是适用于预测我国 GDP 的, 基于最优模型的预测值可以及时捕捉宏观经济走势的变化, 为宏观调控的时机选择及力度把握等提供决策依据。

参考文献:

- [1] 陈飞, 高铁梅. 结构时间序列模型在经济预测方面的应用研究 [J]. 数量经济技术经济研究, 2005(02): 95 – 103.
- [2] 田梓辰, 刘森. 基于改进灰色 GM(1,1) 模型的 GDP 预测实证 [J]. 统计与决策, 2018, 34(11): 83 – 85.
- [3] Jahn M. Artificial neural network regression models in a panel setting: predicting economic growth [J]. Economic Modelling, 2020, 91: 148 – 154.
- [4] Dušan C, Meysam A, Ivan P, et al. A machine learning approach for predicting the relationship between energy resources and economic development [J]. Statistical Mechanics and its Applications, 2018, 30(09): 211 – 214.
- [5] 肖争艳, 刘玲君, 赵廷蓉, 等. 深度学习神经网络能改进 GDP 的预测能力吗? [J]. 经济与管理研究, 2020, 41(07): 3 – 17.
- [6] 谢成兴, 王丰效. 基于 ARIMA – DGM – BP 组合模型的喀什地区 GDP 预测 [J]. 数学的实践与认识, 2020, 50(15): 43 – 48.
- [7] Han H, Li Z, Li Z. Using machine learning methods to predict consumer confidence from search engine data [J]. Sustainability, 2023, 15(04): 3100.
- [8] 刘宽斌, 张涛. 利用网络搜索大数据实现对 CPI 的短期预报及拐点预测——基于混频抽样数据模型的实证研究 [J]. 当代财经, 2018, 408(11): 3 – 15.
- [9] 何强, 董志勇. 利用互联网大数据预测季度 GDP 增速的方法研究 [J]. 统计研究, 2020, 37(12): 91 – 104.
- [10] Xie L, Han T, Zhou H, et al. Tuna swarm optimization: a novel swarm-based metaheuristic algorithm for global optimization [J]. Computational Intelligence and Neuroscience, 2021, 2021: 1 – 22.