

基于 PCA-SVR 模型中国工业固废产生量预测研究

刘炳春, 齐鑫

(天津理工大学 管理学院, 天津 300384)

摘要:依据国家统计局及中国统计年鉴数据,选取国内生产总值(GDP)、工业增加值、财政收入、固定资产投资、原煤产量、原油产量、发电量、年末总人口、我国工业企业单位数量等 9 个指标作为输入指标,构建了 PCA-SVR(主成分分析-支持向量回归)中国工业固废产生量预测模型.并与支持向量回归(Support Vector Regression, SVR)、岭回归(Ridge Regression, RDG)、决策树(Decision Tree, DT)、提升树回归(Gradient Boosting Regression, GBR)多种单一模型的预测结果进行对比.实验结果表明,PCA-SVR 组合模型的平均绝对百分误差(MAPE)为 0.063 0,均方根误差(RMSE)为 2.671 8,其预测误差最小.

关键词:工业固废产生量;PCA-SVR;预测;政策引导

中图分类号:X825

文献标志码:A

工业固废的产生数量伴随中国经济的高速发展而激增.全国 246 个大、中城市发布了固体废物污染环境防治信息情况,其中一般工业固体废物产生量为 14.8 亿 t,工业固体危险废物产生量为 3 344.6 万 t^[1].十八大以后,国家环保压力趋紧,工业固废治理逐步加强,治理环境污染投资有上升的趋势,但仍有大量的工业固废未经处理直接排放,消耗土地资源,加剧了对环境承载力的考验.工业固废的处置面临资金、制度、技术等方面的掣肘,建立模型对工业固废产生量进行预测,梳理工业固废产生趋势条理,可为国家的调控与治理提供决策依据.对于工业固废的研究,主要集中在应用数学模型对工业固废产生量进行预测以及对工业固废处理处置两个方面.通过灰色模型对工业固废产生量进行预测成为许多学者采用最多的研究方法^[2].另外一些学者探索不同方法的建模方向,对工业固废产生量的预测进行了积极的探索.李惠萌等^[3]构建等位灰数递补残差修正模型,运用 MATLAB 工具预测了湘江流域地区的工业固废产生量.李峰等^[4]进行微分方程演化建模,发挥了该模型算法的自适应、自组织、自学习的特性,预测了山东省的工业固废产生量,其拟合精度高于 GM(1,1)模型.赵婉君等^[5]将统计分析方法中的聚类分析与灰色模型相结合,预测了全国 30 个主要城市的工业固废产生量,其结果较为精准.此外,有学者在不同的方面对工业固废进行了研究.LIU 等^[6]研究了典型工业固体废弃物污染现状及环境无害化管理(EMS)方式.ZHANG 等^[7]建立有害废物的管理与跟踪的清单制度,对天津经济技术开发区的一般固体废物进行管理.文献^[8]运用多层次决策技术,研究了伊朗西南部省份工业单位排放的工业固废类别及数量,探究不同的工业固所占总体污染排放的比例.宋小龙等^[9]运用生命周期管理方法对工业固体废物进行分析,得出不同情境的环境负荷评价结果.逯馨华等^[10]通过区域物质流运动的角度来研究工业固废链,揭示工业产业与区域经济和环境的相互影响机制.赵丽娜等^[11]转而对工业固废的产生特征进行研究,探讨了我国工业固体废物管理存在的问题,并提出了相应的对策和建议.

基于以上文献探讨,工业固废的预测研究大部分集中在通过应用单一的计算模型对不同地区和城市的固废产生量进行预测,输入指标多为一维工业固废产生量,输出指标为未来的工业固废产生量.单一的计算模型忽略了多种因素对工业固废产生量的影响,进而影响预测的准确性.工业固废产生量受到多种因素的影响,进行多种输入指标预测模型构建可降低单一变量或低影响因素掺杂下的预测误差,为工业固废产生量预

收稿日期:2019-01-02;修回日期:2019-04-01.

基金项目:天津市教委社会科学重大项目(2017JWZD16)

作者简介(通信作者):刘炳春(1980—),男,天津人,天津理工大学副教授,主要研究方向为循环经济与能源经济, E-mail: tjutlbc@tjut.edu.cn.

测研究开辟新的方向.本文通过搜集历史数据,通过9项指标,应用PCA-SVR组合模型对工业固废的产生量进行预测,以期减小预测误差,为相关部门出台环境治理政策提供参考依据.

1 研究方法

1.1 主成分分析

主成分分析(Principal Component Analysis, PCA)是一种经典的数据降维方法,能够从多维数据中提取相关信息,并通过线性变换提取数据中的主成分,以此来降低多维数据的维度,去除数据中的干扰点^[12].因此PCA可用于多种类别的数据处理^[13-15],提高数据的输入质量^[16].PCA的实现步骤如下^[17]:

构建数据矩阵 $\mathbf{X}_{(n \times p)}$,其中 n 为样本个数, p 为变量个数.将原始数据矩阵进行标准化,得到标准化矩阵 $\mathbf{X}_{(n \times p)}^*$.计算 $\mathbf{X}_{(n \times p)}^*$ 的协方差矩阵 \mathbf{R} ,其中, \mathbf{X}^T 为转置矩阵,公式如下:

$$\mathbf{R} = \frac{1}{n-1} \mathbf{X}^T \mathbf{X}. \quad (1)$$

计算 \mathbf{R} 的特征向量矩阵 \mathbf{M} 以及特征矩阵 \mathbf{N} ,公式如下:

$$\mathbf{R}\mathbf{M} = \mathbf{M}\mathbf{N}. \quad (2)$$

根据矩阵 \mathbf{N} 和 \mathbf{M} ,计算输入数据的主成分贡献率 Q_k 以及累计贡献率 Q .其计算公式如下:

$$Q_k = \frac{\lambda_k}{\sum_{k=1}^p \lambda_k} \times 100\%; Q = \sum_{k=1}^p Q_k,$$

其中 λ_k 为特征值,最后根据主成分的贡献率选择数据.

1.2 支持向量回归

支持向量回归(Support Vector Regression, SVR)是支持向量机(SVM)的延伸应用.SVR是以结构风险最小化为回归目标的算法,其拥有处理非线性数据的能力,能够有效克服神经网络只能达到局部最小的缺点,从而达到全局最优化^[18],是一种重要的解决回归问题的方法^[19],被广泛应用^[20].输入的非线性过程可以通过回归来描述^[21].其过程如下:

假设 $f(x) = w\varphi(x) + b$,其中 w 和 b 是调整系数, $\varphi(*)$ 表示映射函数, x 为输入数据向量.

$$\begin{cases} \min C \sum_{i=1}^n \xi_i^+ + \xi_i^- + \frac{1}{2} \|w\|^2, \\ \text{s.t. } f(x_i) - y_i \leq \xi_i^+ + \epsilon, i = 1, \dots, n, \\ f(x_i) - y_i \geq \xi_i^- + \epsilon, i = 1, \dots, n, \\ \xi_i^+, \xi_i^- \geq 0, i = 1, \dots, n, \end{cases} \quad (3)$$

其中, C 为惩罚函数, ϵ 为不敏感损失函数参数, ξ_i^+ , ξ_i^- 为松弛变量, y_i 为第 i 个输出.

通过引入拉格朗日函数求解(3)式,可以得知: $w = \sum_{i=1}^n (\beta_i^* - \beta_i) \varphi(x_i)$,其中 β_i^* 和 β_i 为拉格朗日系数.

最后求得: $f(x) = \sum_{i=1}^n (\beta_i^* - \beta_i) K(x_i, x_j) + b$, (4)

$K(x_i, x_j)$ 为参数选择中的核函数.

2 实验与分析

2.1 数据来源

本文选用了国家统计局及中国统计年鉴数据.实验选取1980—2015年共35年的数据,其中提取国内生产总值(GDP)、工业增加值、财政收入、固定资产投资、原煤产量、原油产量、发电量、年末总人口、我国工业企业单位数量共9个指标作为输入指标以及工业固废产生量作为输出指标.本文选取1980—2008年的29笔数据作为训练样本(train date),2009—2015年的7笔数据进行测试验证(test date).训练样本用于矫正模型的误差,测试样本用于验证模型的准确性,以修正预测模型.

2.2 构建预测模型

为提升预测模型精度,本文运用 PCA 对数据进行降维处理,并构建 PCA-SVR 组合模型,对我国工业固废的产量进行预测.其流程如图 1 所示,PCA 处理结果如表 1 所示.

其中, X_i 为输入特征值, $C_{k_1} - C_{k_n}$ 为隐藏层, T 作为输出结果.本文输入值选取 X_1 国内生产总值、 X_2 工业增加值、 X_3 财政收入、 X_4 固定资产投资、 X_5 原煤产量、 X_6 原油产量、 X_7 发电量、 X_8 年末总人口、 X_9 我国工业企业单位数, T 工业固废产生量作为输出变量,并将其运用科学计数法进行标准化.选取均方根误差(RMSE)、平均绝对百分误差(MAPE)进行误差比对.预测模型关键参数如下:核函数类型为线性核函数,惩罚函数 $cost = 10$,开启收缩启发式,停止训练误差精度 $tol = 0.001$.

由表 1 可知,前 4 项的主成分贡献率高于 2%,累计贡献率大于 99%,结果表明其携带原始数据的主要信息,因此提取前 4 项主成分构建 PCA-SVR 预测模型.

表 1 PCA 结果

Tab.1 The PCA results

序号	特征值	贡献率/%	累计贡献率/%	序号	特征值	贡献率/%	累计贡献率/%
1	6.779 7	75.330 4	75.330 4	6	0.002 3	0.025 9	99.973 4
2	1.029 9	11.444 3	86.774 8	7	0.001 7	0.019 5	99.993 0
3	0.927 4	10.305 3	97.080 2	8	0.000 5	0.005 8	99.998 9
4	0.224 1	2.490 9	99.571 1	9	0.000 0	0.001 0	100.000 0
5	0.033 8	0.376 3	99.947 5				

为达到实验目的,增加预测的准确性,本文构建多种预测模型,并将结果进行比对,选取测试样本中误差最低的模型进行预测.本文选用支持向量回归(Support Vector Regression, SVR)^[22]、岭回归(Ridge Regression, RDG)^[23]、决策树(Decision Tree, DT)^[24]、提升树回归(Gradient Boosting Regression, GBR)^[25]模型与 PCA-SVR 模型比对,其结果见表 2.

通过实验,得到 PCA-SVR 的组合模型 MAPE 为 0.063 0, RMSE 为 2.671 8,误差低于其余比对工业固废预测模型.因此本文构建 PCA-SVR 组合模型进行工业固废产生量的预测,PCA-SVR 组合模型与其余模型训练数据和测试样本实验输出结果见图 2.

2.3 基于 PCA-SVR 的工业固废产生量预测

上述实验表明,基于 PCA-SVR 组合模型对中国未来的工业固废产生量进行预测是可行的.本文通过假设不同情境对中国工业固废的产生量进行预测.输入特征变量中的人口数量采用国务院印发的《国家人口发展规划(2016—2030 年)》中的数据,到 2020 年我国总人口达到 14.2 亿左右.根据我国 2015 年人口数量 13.74 亿人,推算出人口的年平均增长率以及每年的人口数量.根据十八大报告中提出的要求,2020 年中国 GDP 要比 2010 年翻一番,其最低年增长率要达到 6.5%.其余输入指标存在差异性,按照我国发展现状将分为 3 种情境,并分别计算其增长率进行预测,见表 3.

本文通过 PCA-SVR 模型预测,结果显示,我国的工业固废产生量成增加态势.情境 1 的情况下,我国工业固废产生量将增加到 2021 年的 42.981 6 亿 t;情境 2 中,我国工业固废产生量 2015 年 32.707 9 亿 t 增长到

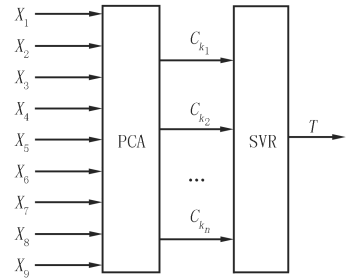


图 1 PCA-SVR 组合模型流程图
Fig.1 Flow chart of PCA-SVR composite mode

表 2 工业固废预测模型误差比较

Tab.2 Comparison of error model of industrial solid wastes

模型	MAPE	RMSE
DT	0.363 4	12.667 0
GBR	0.345 2	11.875 8
RDG	0.138 1	5.549 1
SVR	0.189 1	7.597 8
PCA-SVR	0.063 0	2.671 8

2021年的58.4438亿t;情境3中,我国工业固废产生量增速较快,将在2021年达到174.5816亿t.不同情境的预测结果见图3.

表3 情境分类表

Tab.3 Situation classification

预测情境	情境1	情境2	情境3
工业固废产生量	初始增长率	平均增长率	最大增长率

2.4 预测结果分析与对策

通过对预测结果的分析,中国的工业固废产生量成增长态势,工业固废产生量和中国的工业化进程具有关联性,而资源型企业是产生工业固废的主力.资源型企业生产规模化、产地化以及生产技术升级提升了企业生产效率,但同时也导致了工业固废产生量的增加.此外,工业固废处置仍然面临技术和资金制约,部分工业固体废物已经有循环利用的技术,如煤矸石回填技术、粉煤灰再利用技术等,但因技术掣肘,处理成本高昂,大部分的工业固废仍然以堆存作为处理方式.

工业固废兼具资源性与污染性双重特征,因此我国应以固废循环利用技术为依托,抓住环境治理的机遇,加强工业固废资源化、无害化的处理力度;继续加大环境污染治理投资,支持企业采购工业固废无害化处理设备,研究工业固废处理技术,如尾矿回填技术、粉煤灰回收利用技术等.充分宣传引导,发挥工业固废循环利用龙头企业示范带动作用,总结工业固废处理资源化、无害化典型案例进行推广;提高高污染、高耗能、高工业固废产生量企业的准入标准,加大企业环保核查力度,推动企业升级转型,力争实现效益与生态共赢的局面;推动示范园区的建设,建设循环经济示范园区,以循环经济为载体,以循环产业链为核心,以技术推动为抓手,构建区域性循环发展模式,推动工业固废资源化.

4 结论

本文采取PCA-SVR组合模型对中国的工业固废产生量进行预测,选用了国内生产总值(GDP)、工业增加值、财政收入、固定资产投资、原煤产量、原油产量、发电量、年末总人口、我国工业企业单位数量9个特征变量进行预测模型的构建.通过实验可以得到以下结论:(1)基于PCA-SVR组合模型进行中国工业固废产生量的预测精度比单一SVR模型精度高,MAPE可以从0.1891提升到0.0515.(2)中国的工业固废产生量和上述9个特征变量息息相关,PCA降维结果中,GDP、工业增加值、财政收入、固定资产投资携带信息量较大.(3)工业固废产生量成增加态势,与中国工业化进程相符,政府应继续完善环境治理和保护的法相关规,

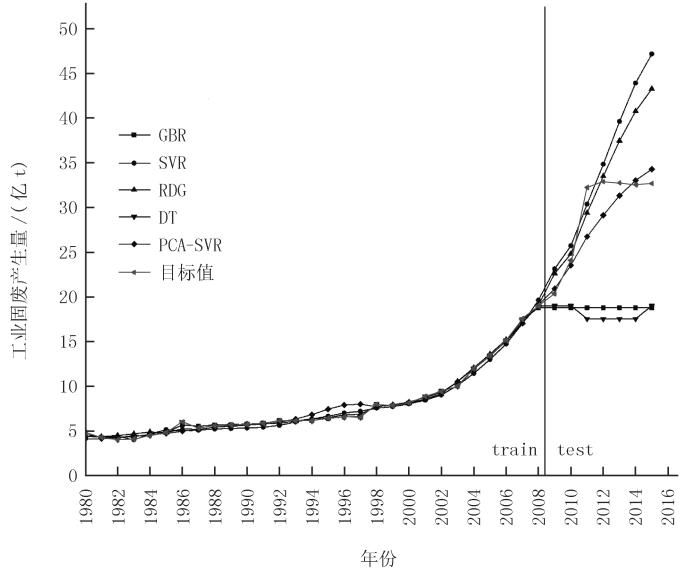


图2 工业固废产生量预测

Fig.2 PCA-SVR prediction chart

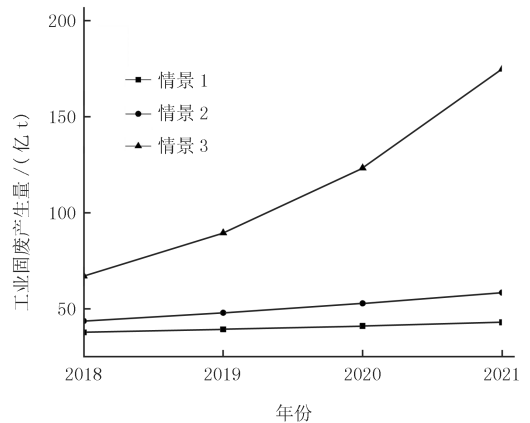


图3 基于PCA-SVR模型的工业固废产生量预测结果

Fig.3 PCA-SVR model based on the result of predicting the amount of industrial solid waste

加强工业固废循环利用能力。

本文选取多个特征变量进行工业固废产生量的预测,拓宽了仅通过工业固废产生量单一特征值进行预测的模型的研究方向。此外,由于数据精度限制和模型自身局限性制,预测误差会在一定范围内波动。工业固废产生量的预测还需要更深入的建模、研究与探讨。针对工业固废预测的研究方向,还可以拓展到如下两个方面。第一,加大工业固废的研究地区的范围,探求工业固废产生的区域关联性;第二,探索不同的预测模型构建,加强预测算法与程序架构的研究,增加预测的准确性。

参 考 文 献

- [1] 2017 年全国大、中城市固体废物污染环境防治年报[J].环境保护,2018,46(Z1):90-106.
2017 Annual Report on Prevention and Control of Solid Waste in China's Large and Medium-Sized Cities[J].Environmental Protection, 2018,46(Z1):90-106.
- [2] 邓琪,王琪,黄启飞.GM(1,1)在工业固体废物产生量预测中的应用[J].环境科学与技术,2012,35(6):180-183.
DENG Q,WANG Q,HUANG Q F.Application of GM(1,1)in Forecasting Quantity of Industrial Solid Waste[J].Environmental Science & Technology,2012,35(6):180-183.
- [3] 李惠萌,袁兴中,曾光明,等.基于 MATLAB 的湘江流域工业固体废物灰色预测[J].环境科学与技术,2008(8):136-140.
LI H M,YUAN X Z,HUANG Q M,et al.Grey Forecasting on Industrial Solid Waste in Xiangjiang Valley Based on MATLAB[J].Environmental Science & Technology,2008(8):136-140.
- [4] 李峰,李永干.微分方程演化建模用于工业固废产量的研究[J].滨州师专学报,2000(2):31-33.
Li F,Li Y G.Evolutionary Modeling of Differential Equations for Study on Year's Output of Industrial Solid Wastes[J].Journal of Binzhou Teachers College,2000(2):31-33.
- [5] 赵婉君,江浩芝,彭开鲜.基于统计分析的固废产生量预测方法初探[J].广东化工,2015,42(12):55.
ZHAO W J,JIANG H Z,PENG K X.Analysis of Solid Waste Generation Prediction Method Based on Statistics[J].Guangdong Chemical Industry,2015,42(12):55.
- [6] LIU Y Q,GUO D W,LU D,et al.Pollution Status and Environmental Sound Management(ESM)Trends on Typical General Industrial Solid Waste[J].Procedia Environmental Sciences,2016,31:615-620.
- [7] ZHANG M,WANG Y H,SONG Y Y,et al.Manifest system for management of non-hazardous industrial solid wastes; results from a Tianjin industrial park[J].Journal of Cleaner Production,2016,133:252-261.
- [8] KARAMOUZ M,ZAHRAIE B,KERACHIAN R,et al.Development of a master plan for industrial solid waste management[J].International Journal of Environmental Science & Technology,2006,3(3):229-242.
- [9] 宋小龙,徐成,杨建新,等.工业固体废物生命周期管理方法及案例分析[J].中国环境科学,2011,31(6):1051-1056.
SONG X L,Xu C,YANG J X,et al.A method for life cycle management of industrial solid waste and its case study[J].China Environmental Science,2011,31(6):1051-1056.
- [10] 逯馨华,杨建新,陈波,等.工业固废生态链的构建对区域物质流的影响[J].中国人口·资源与环境,2010,20(11):147-153.
LU X H,YANG J X,CHEN B,et al.Effects of Industrial Solid Waste Exchange Chain on Regional Material Flow[J].China Population, Resources and Environment,2010,20(11):147-153.
- [11] 赵丽娜,姚芝茂,武雪芳,等.我国工业固体废物的产生特征及控制对策[J].环境工程,2013,31(S1):464-469.
ZHAO L N,YAO Z M,WU X F,et al.Generation Characteristics and Control Countermeasures of Industrial Solid Waste in China[J].Environmental Engineering,2013,31(S1):464-469.
- [12] 许爱东,李昊飞,程乐峰,等.PCA-PSO-ELM 配网供电可靠性预测模型[J].哈尔滨工程大学学报,2018,39(6):1116-1122.
XU A D,LI H F,CHENG L F,et al.Prediction model for power supply reliability of distribution network using PCA-PSO-ELM[J].Journal of Harbin Engineering University,2018,39(6):1116-1122.
- [13] 黄忠山,田凌,向东,等.基于 PCA 和 SPC-动态神经网络的风电机组齿轮箱油温趋势预测[J].清华大学学报(自然科学版),2018,58(6):539-546.
HUANG Z S,TIAN L, XIANG D,et al.Prediction of oil temperature variations in a wind turbine gearbox based on PCA and an SPC-dynamic neural network hybrid[J].Journal of Tsinghua University(Science and Technology),2018,58(6):539-546.
- [14] 吴圣超,刘太昂,葛炯,等.化学成分-朴素贝叶斯分类算法的烟叶产地模式识别[J].河南师范大学学报(自然科学版),2018,46(1):77-83.
WU S C,LIU T A,GE J,et al.Pattern recognition of the producing areas of flue-cured tobacco based on naive bayesian classifier algorithm base on the contents of chemical components[J].Journal of Henan Normal University(Natural Science Edition),2018,46(1):77-83.
- [15] 徐久成,黄方舟,穆辉宇,等.基于 PCA 和信息增益的肿瘤特征基因选择方法[J].河南师范大学学报(自然科学版),2018,46(2):104-110.
XU J C,HUANG F Z,MU H Y,et al.Tumor feature gene selection method based on PCA and information gain[J].Journal of Henan Normal University(Natural Science Edition),2018,46(2):104-110.

- [16] 聂敏,刘志辉,刘洋,等.基于PCA和BP神经网络的径流预测[J].中国沙漠,2016,36(4):1144-1152.
NIE M,LIU Z H,LIU Y,et al.Runoff Forecast Based on Principal Component Analysis and BP Neural Network[J].Journal of Desert Research,2016,36(4):1144-1152.
- [17] 王卫红,卓鹏宇.基于PCA-FOA-SVR的股票价格预测研究[J].浙江工业大学学报,2016,44(4):399-404.
WANG W H,ZHUO P Y.Research on stock price prediction based on PCA-FOA-SVR[J].Journal of Zhejiang University of Technology,2016,44(4):399-404.
- [18] ESMAEILI M,SALIMI A,DREBENSTEDT C,et al.Application of PCA,SVR,and ANFIS for modeling of rock fragmentation[J].Arabian Journal of Geosciences,2014,8(9):6881-6893.
- [19] KHODAPANAH M,ZOBAA A F,ABBOD M.Estimating power factor of induction motors at any loading conditions using support vector regression(SVR)[J].Electrical Engineering,2018,100(1):1-10.
- [20] YIN Z L,QI F,WEN X H,et al.Design and evaluation of SVR,MARS and M5Tree models for 1,2 and 3-day lead time forecasting of river flow data in a semiarid mountainous catchment[J].Stochastic Environmental Research & Risk Assessment,2018,32(9):2457-2476.
- [21] Müller K R,SMOLA A J,Rätsch G,et al.Predicting time series with support vector machines[C]//International Conference on Artificial Neural Networks.Berlin:Heidelberg,1997:999-1004.
- [22] 姚卫红,方仁孝,张旭东.基于混合人工鱼群优化SVR的交通流量预测[J].大连理工大学学报,2015,55(6):632-637.
YAO W H,FANG R X,ZHANG X D.Traffic flow forecasting based on optimized SVR with hybrid artificial fish swarm algorithm[J].Journal of Dalian University of Technology,2015,55(6):632-637.
- [23] 董小刚,刁亚静,李慧玲,等.岭回归、LASSO回归和Adaptive-LASSO回归下的财政收入因素分析[J].吉林师范大学学报(自然科学版),2018,39(2):45-53.
DONG X G,DIAO Y J,LI H L,et al.The analysis of the fiscal revenue factors under the ridge regression,LASSO regression and the Adaptive-LASSO regression[J].Jilin Normal University Journal(Natural Science Edition),2018,39(2):45-53.
- [24] 安璐,周思瑶,余传明,等.突发传染病微博影响力的预测研究[J].情报科学,2017,35(4):27-31.
AN L,ZHOU S Y,YU C M,et al.Predicting the Influence of Microblog Entries on Emergent Infectious Diseases[J].Information Science,2017,35(4):27-31.
- [25] 张加龙,胥辉,陆驰.应用Landsat8 OLI和GBRT对高山松地上生物量的估测[J].东北林业大学学报,2018,46(8):25-30.
ZHANG J L,XU H,LU C.Estimating Above Ground Biomass of Pinus densata Based on Landsat8 OLI and Gradient Boost Regression Tree[J].Journal of Northeast Forestry University,2018,46(8):25-30.

Research on the forecast of industrial solid waste generation in China based on PCA-SVR

Liu Bingchun, Qi Xin

(School of Management, Tianjin University of Technology, Tianjin 300384, China)

Abstract: In this paper, based on the information from China Statistical Yearbook of 1980-2015, data of gross domestic product(GDP), industrial added value, fiscal revenue, fixed asset investment, output of raw coal, crude oil production, electricity generation, population at the Year-end, the number of industrial enterprises in our country were selected as input features. The PCA-SVR(principal component analysis-support vector regression) prediction model for solid waste production in China is established. It was compared with the prediction results of a variety of single models including Support Vector Regression(SVR), Ridge Regression(RDG), Decision Tree(DT) and Gradient Boosting Regression(GBR). The experimental results showed that the mean absolute percentage error(MAPE) and root mean square error(RMSE) of PCA-SVR model are 0.063 0 and 2.671 8 respectively, and the prediction error is the smallest.

Keywords: industrial solid waste generation; PCA-SVR; forecast; policy guidance

[责任编辑 赵晓华 陈留院]