基于非期望 SBM-SVM 改进模型的投资有效性预测
——以重庆市工业行业为例

徐 杰,陈义安
重庆工商大学 数学与统计学院,重庆 400067

摘 要:对工业行业进行有效投资在一定程度上对经济发展有重要影响,作为能源消耗和环境污染的主要源头,为落实绿色发展理念,推动节能减排工作,提出对绿色工业投资的有效性研究。基于非期望 SBM-SVM 模型并对其改进,选取重庆市 2011—2020 年工业企业相关指标作为样本数据,将通过非期望 SBM 模型得到的评价效率分为有效和无效两类作为结果变量,投入和产出指标作为特征变量,构建 SVM 模型,对工业投资有效性进行分类预测研究,通过“试错法”、PSO、GA 智能优化算法对 SVM 模型的惩罚因子 C 和核函数参数 g 进行寻优。结果显示:PSO 方法的寻优效果最佳,准确率从 71.88%提高到了 88.66%;构建的新非期望 SBM-SVM 模型在对其改进优化后,进行工业投资有效性分类,具有一定的可行性和适用性。

关键词:SVM;非期望 SBM;投资有效性
中图分类号:O643 **文献标识码**:A doi:10.16055/j.issn.1672-058X.2023.0001.016

Investment Effectiveness Prediction Based on Undesirable SBM-SVM Improvement Model: A Case Study of Industrial Sector in Chongqing

XU Jie, CHEN Yian
School of Mathematics and Statistics, Chongqing Technology and Business University, Chongqing 400067, China

Abstract: The industrial sector is the main source of energy consumption and environmental pollution, and effective investment in the industrial sector has an important impact on economic development to a certain extent. In order to implement the concept of green development and promote energy conservation and emission reduction, a study on the effectiveness of investment in green industry was proposed. Based on the undesirable SBM-SVM model and its improvement, the relevant indicators of industrial enterprises in Chongqing from 2011 to 2020 were selected as the sample data, and the evaluation efficiency obtained by the undesirable SBM model was divided into two categories: effective investment and ineffective investment as the outcome variables. The input and output indicators were used as characteristic variables to construct an SVM model to study the classification and prediction of industrial investment effectiveness. Through trial-and-error method, PSO, and GA intelligent optimization algorithms, the penalty factor C and kernel function parameter g of the SVM model were optimized. The results showed that the optimization effect of the PSO method was the best, and the accuracy rate was increased from 71.88% to 88.66%. The constructed new undesirable SBM-SVM model has certain feasibility and applicability to classify the effectiveness of industrial investment after improvement and optimization.

Keywords: SVM; undesirable SBM; investment effectiveness

收稿日期:2022-01-27 **修回日期**:2022-04-18 **文章编号**:1672-058X(2023)01-0097-08
基金项目:2021 年重庆工商大学研究生创新型科研项目(YJSCXX2021-112-106).
作者简介:徐杰(1997—),女,重庆人,硕士研究生,从事经济统计研究.
引用格式:徐杰,陈义安. 基于非期望 SBM-SVM 改进模型的投资有效性预测——以重庆市工业行业为例[J]. 重庆工商大学学报(自然科学版),2023,40(1):97—104.
XU Jie, CHEN Yian. Investment effectiveness prediction based on undesirable SBM-SVM improvement model: a case study of industrial sector in Chongqing[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2023, 40(1): 97—104.

1 引 言

在国民经济中,工业行业是最重要的经济和生产技术综合体,工业行业分为采矿业,制造业,电力、热力、燃气及水生产和供应业3个门类,总共划分为41个行业大类,工业内部各行业间的投资结构对工业发展有一定的影响,行业间的有效投资有利于工业的健康发展。同时,工业也是能源消耗、环境污染的主要源头,近年来,工业对能源的消耗和产生的环境污染情况也引起了重视,提出的新发展理念包括绿色发展,即在经济发展过程中,要做到对保护环境、节约资源的落实,要做到对节能减排目标的完成,以此来促进经济转型的发展。因此,在研究工业行业的投资有效性时,加入环境污染这一非期望产出指标,将工业行业细分为39个行业,分析研究工业行业间的经济投资相对有效性情况,并构建出分类模型。

近年来将DEA方法与机器算法结合构建新的预测方法引起了众多学者的研究,Song[1]、冉茂盛[2]、李宁等[3]学者将DEA方法中的输入、输出数据以及结果作为训练集训练SVM模型,剩余的决策单元作为测试集进行测试,分别对企业绩效、上市公司的经营效率以及企业的平行部门绩效进行评价与分类预测,最终依据预测分类准确率验证了该方法具有可用性、有效性和实用性。Zhang[4]则是通过构建IG-SVM模型对DEA模型输入和输出数据的最小值、平均值、最大值进行预测,并利用DEA模型计算出决策单元的未来效率值,通过实例表明方法的可行性和适用性。同样,李玉龙[5]和Zhu等[6]学者建立了DEA与神经网络集成模型及机器学习(ML)算法之间的联系,分别对基础设施的投资有效性和中国制造业上市公司的绩效进行预测,证明了方法的适用性。通过上述众多学者的研究及实证证明:将DEA与机器学习算法以及SVM模型相结合构建的新模型对投资有效性研究具有可行性。但是在学者们的研究中,并没有考虑在投入产出过程中,会存在非期望产出的情况,因此本文引入非期望产出指标,利用非期望产出SBM方法与SVM方法相结合生成一个新的模型,以此研究非期望产出时的投资有效性。同样,对SVM模型参数优化,也是众多学者研究讨论的话题,通过对SVM模型的惩罚因子(C)和核函数参数(g)寻优,找到SVM模型的最优参数,提高预测准确率和模型的可用性。徐晓明[7]、颜薇等[8]分别利用智能优化算法、AGA模型对支持向量机的参数C和g进行优化,通过数值实验结果得出优化后的效果,使得预测效果更好。对于SVM模型,可以对其参数寻优,使得预测效果更佳,因此本文同样考虑运用智能优化算法对支持向量机的惩罚参数C和核函数g进行寻优,找到非期望产出SBM-SVM模型的最优C和g。本文将非期望产出SBM与SVM结合建立一个新的有效性分类方法,并对此进行改进优化,得到更佳的分类效果,此方法可以加入非期望产出指标,有利于对绿色发展、绿色投资的有效性等方向的研究。因此,本文探讨非期望SBM模型和SVM模型结合构建新的有效性分类方法及其对其优化是否具有可行性,并利用实证进行研究。

重庆市的经济正处于增长阶段,2020年重庆市全年GDP为25 002.79亿元,其中工业生产总值占国内比重28%,且近10 a来,一直维持在27.8%~37%的占比,为全行业最高。全年工业增加值为6 990.77亿元,比上年增长5.3%,规模以上工业增加值比上年增长5.8%,工业固定资产投资也呈现逐年上升的趋势,2020年比上年增长5.8%、研究重庆市工业各行业的投资有效性可以更好地分析优化产业结构,并促进工业长期稳定发展。本文选择非期望产出SBM模型对重庆市工业各行业投资效率的研究,并基于SVM模型构建出非期望SBM-SVM模型对投资有效性进行分类,运用优化模型对SVM方法参数寻优,根据结果情况得出结论。

2 非期望SBM-SVM有效性方法

Tone[9]在2001年提出了非径向非角度的SBM模型,此方法作为DEA的衍生模型,很好地解决了DEA方法由于径向距离函数以及角度模型所出现的在效率评估中的缺陷。

在投入产出过程中会产生负面效应,可以分为期望和非期望产出。于本文而言,节能减排是近年来备受关注的话题,我们期望在投入的过程中能排放出更少的环境污染物,因此加入非期望产出指标,能够更加科学、绿色健康地评价。

假设有s个决策单元,每一个决策单元有m个投入单元、n_1个期望产出单元和n_2个非期望产出单元,设定X_{ik}代表第k个决策单元的第i种投入,Y^e_{rk}代表第k个决策单元的第r种期望产出,Y^u_{rk}代表第k个决策单元的第r种非期望产出,非期望SBM模型的基本形式为

$$\min \rho^* = \frac{1-\frac{1}{m}\sum_{i=1}^{m} s_i^- / X_{ik}}{1+\frac{1}{n_1+n_2}\left(\sum_{r=1}^{n_1}\frac{s_r^+}{Y^e_{rk}}+\sum_{r=1}^{n_2}\frac{s_r^{u^-}}{Y^u_{rk}}\right)}$$

$$\begin{cases} \text{s.t. } X\lambda + s^- = X_k \\ Y^e\lambda - s^+ = Y^e_k \\ Y^u\lambda + s^{u^-} = Y^u_k \\ \lambda, s^+, s^-, s^{u^-} \geqslant 0 \end{cases}$$

其中,s^-为投入指标松弛变量,s^+为期望产出松弛变量,s^{u^-}为非期望产出松弛变量,投入要素权重为 λ,最终求解的ρ^*为决策单元的效率值,其取值在[0,1]。如果$\rho^*=1$,则称决策单元为 DEA 相对有效;若$\rho^*<1$,则称决策单元为非 DEA 相对有效。

将有效单元设为 $y_k=+1$,反之则设 $y_k=-1$,DMU 的输入和输出指标合并为 $x_k=(X_k,Y_k^e,Y_k^u)$,则 s 个决策单元可以构成(x_k,y_k);$k=1,\cdots,s$;$y\in\{+1,-1\}$。选取 p 个数量的决策单元作为训练集 $D=\{(x_k,y_k)\mid k=1,2,\cdots,p\}$,满足 $y_k[(\boldsymbol{\omega}^{\mathrm{T}}x_k+b)]-1\geqslant 0$,则剩余$(s-p)$个决策单元为测试集。设定可以将有效单元和无效单元两类分类的最优分类超平面表达式为 $D(x)=\boldsymbol{\omega}\cdot x+b$,$\boldsymbol{\omega}$ 为权重量,b 为分类阈值。每个样本点到平面的距离为$\gamma=\dfrac{|\boldsymbol{\omega}^{\mathrm{T}}x+b|}{\|\boldsymbol{\omega}\|}$,则有效单元与无效单元样本点的距离为$2\gamma$,分隔间隔为 $d=\dfrac{2}{\|\boldsymbol{\omega}\|}$,当$\dfrac{1}{\|\boldsymbol{\omega}\|}$为最大值时最优,即 $\|\boldsymbol{\omega}\|$ 取值最小时。

当训练集 D 线性可分时,将分类问题转化成了带约束条件的二次规划问题:

$$\min\varphi(x)=\frac{1}{2}\|\boldsymbol{\omega}\|^2=\frac{1}{2}\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\omega}$$

$$\text{s.t. } y_k(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x_k}+b)\geqslant 1,k=1,2,\cdots,p$$

将二次规划问题转化为对偶函数,引入每个样本对应的拉格朗日乘子 α_k,$\alpha_k\geqslant 0$,$k=1,2,\cdots,p$,可以得到:

$$L(\boldsymbol{\omega},b,\alpha)=\frac{1}{2}\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\omega}-\sum_{i=1}^{p}\alpha_k[y_k(\boldsymbol{\omega}^{\mathrm{T}}\cdot\boldsymbol{x_k}+b)-1] \tag{1}$$

对 $\boldsymbol{\omega}$ 和 b 求导数,并令其为零,可以求得 $\boldsymbol{\omega}$ 和 b 的最小值:

$$\frac{\partial L}{\partial\boldsymbol{\omega}}=\boldsymbol{\omega}-\sum_{k=1}^{p}y_k\alpha_k x_k=0 \tag{2}$$

$$\frac{\partial L}{\partial b}=\sum_{k=1}^{p}y_k\alpha_k=0 \tag{3}$$

求解式(2)(3),得:

$$\boldsymbol{\omega}=\sum_{k=1}^{p}y_k\alpha_k x_k,\ \sum_{k=1}^{p}y_k\alpha_k=0$$

将结果代入式(1)有:

$$L(\boldsymbol{\omega},b,\alpha)=\sum_{k=1}^{p}\alpha_k-\frac{1}{2}\sum_{k,j=1}^{p}\alpha_k\alpha_j y_k y_j\boldsymbol{x_k}^{\mathrm{T}}\boldsymbol{x_j} \tag{4}$$

设定核函数 $K(\boldsymbol{x_k},\boldsymbol{x_j})$ 代入式(4)中:

$$L(\boldsymbol{\omega},b,\alpha)=\sum_{k=1}^{p}\alpha_k-\frac{1}{2}\sum_{k,j=1}^{p}\alpha_k\alpha_j y_k y_j K(\boldsymbol{x_k},\boldsymbol{x_j})$$

通过求解,可得到二次规划:

$$\max Q(\alpha)=\sum_{k=1}^{p}\alpha_k-\frac{1}{2}\sum_{k,j=1}^{p}\alpha_k\alpha_j y_k y_j K(\boldsymbol{x_k},\boldsymbol{x_j})$$

$$\begin{cases}\text{s.t. }\displaystyle\sum_{k=1}^{p}\alpha_k y_k=0\\ \alpha_k\geqslant 0,k=1,\cdots,p\end{cases}$$

根据求解二次规划问题原理,该二次规划具有唯一解,最优分类函数为

$$f(x)=\operatorname{sgn}(\boldsymbol{\omega}^{\mathrm{T}}x+b)=\operatorname{sgn}\Big(\sum_{k=1}^{p}y_k\alpha_k^*K(\boldsymbol{x_k},\boldsymbol{x_j})+b^*\Big)$$

根据 KKT 条件,分类面为最优分类超平面的充要条件是

$$\alpha_k^*[y_k(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x_k}+b^*)-1]=0,k=1,2,\cdots,p \tag{5}$$

根据互补松弛条件可以看出,式(5)中的 α_k^* 和$[y_k(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x_k}+b^*)-1]$不能同时为 0。就可以得到任一样本$(x_s,y_s)$,有 $b^*=y_s-\boldsymbol{\omega}^{\mathrm{T}^*}\cdot x_s$,任一样本的 $y_s^2=1$,将右式的 1 用 y_s^2 代替,再根据$\boldsymbol{\omega}^*=\sum_{k=1}^{p}\alpha_k^*y_k\boldsymbol{x_k}$,可知:

$$b^*=y_s-\sum_{k=1}^{p}y_k\alpha_k^*\boldsymbol{x_k}^{\mathrm{T}}x_s$$

训练集 D 为线性不可分时,引入了松弛变量 $\zeta_i\geqslant 0$,惩罚系数为 C,构建出最大化分类间隔,使得原始带约束的二次规划问题变成以下表述:

$$\min\varphi(x)=\frac{1}{2}\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\omega}+C\sum_{k=1}^{p}\zeta_k$$

$$\text{s.t. } y_k(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x_k}+b)\geqslant 1-\zeta_k,\zeta\geqslant 0,k=1,2,\cdots,p$$

同样将二次规划问题转化为对偶问题,引入 α_k 为每个样本的拉格朗日乘子,并引入新的拉格朗日乘子 μ_k,得到式(6):

$$L(\boldsymbol{\omega},b,\alpha)=\frac{1}{2}\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\omega}+C\sum_{k=1}^{p}\zeta_k-\sum_{k=1}^{p}\alpha_k[y_k(\boldsymbol{\omega}^{\mathrm{T}}\cdot\boldsymbol{x_k}+b)-1+\zeta_k]-\sum_{k=1}^{p}\mu_k\zeta_k \tag{6}$$

对 $\boldsymbol{\omega},b,\zeta_k$ 求导数,并令其为零,可以求得 $\boldsymbol{\omega}$ 和 b 的最小值:

$$\frac{\partial L}{\partial\boldsymbol{\omega}}=\boldsymbol{\omega}-\sum_{k=1}^{p}y_p\alpha_p x_p=0$$

$$\frac{\partial L}{\partial b}=\sum_{k=1}^{p}y_k\alpha_k=0$$

$$\frac{\partial L}{\partial\zeta_k}=C-\alpha_k-\mu_k=0$$

设可以将线性不可分的两类点变成线性可分的核函数 $K(\boldsymbol{x}_k,\boldsymbol{x}_j)$,则对偶函数为

$$\max\ Q(\alpha)=\sum_{k=1}^{p}\alpha_k-\frac{1}{2}\sum_{k,j=1}^{p}\alpha_k\alpha_j y_k y_j K(\boldsymbol{x}_k,\boldsymbol{x}_j)$$

$$\begin{cases}\text{s.t.}\ \sum_{k=1}^{p}\alpha_k y_k=0\\ 0\leqslant\alpha_k\leqslant C,k=1,\cdots,p\end{cases}$$

求解上述函数,分下列情况讨论:

$0<\alpha_k<C$ 时,有 $\alpha[y_k(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}_k+b)-1+\zeta_k]=0$。因为 $C=\alpha_k+\mu_k$,则 $\alpha_k<C$,所以 $\mu_k>0$,同时因为 $\mu_k\cdot\zeta_k=0$,则 $\zeta_k=0$,$y_k(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}_k+b)=1$。

$\alpha_k=C$ 时,有 $\alpha[y_k(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}_k+b)-1+\zeta_k]=0$。因为 $C=\alpha_k+\mu_k$ 且 $\alpha_k=C$,所以 $\mu_k=0$,$\zeta_k\geqslant0$,则 $y_k(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}_k+b)\leqslant1$。

$\alpha_k=0$ 时,有 $\alpha[y_k(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}_k+b)-1+\zeta_k]=0$。因为 $\alpha_k=0$ 且 $\alpha_k+\mu_k=C$,所以 $\mu_k=C$,同时因为 $\mu_k\cdot\zeta_k=0$,则 $\zeta_k=0$,$y_k(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}_k+b)\geqslant1$。

综上所述,所有样本必须满足:

$$\begin{cases}y_k(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}_k+b)\geqslant1,\text{当}\ \alpha_k=0\\ y_k(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}_k+b)\leqslant1,\text{当}\ \alpha_k=C\\ y_k(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}_k+b)=1,\text{当}\ 0<\alpha_k<C\end{cases}$$

根据 KKT 条件,满足 $0<\alpha_k<C$ 的样本点就是支持向量,它们在侧面上。对于任一样本 (x_s,y_s) 中 $y_s^2=1$,可以得到:

$$b^*=y_s-\sum_{k=1}^{p}y_k\alpha_k^*\boldsymbol{x}_k^{\mathrm{T}}x_s$$

$\bar{b}^*$ 为所有支持向量结果的平均值,其中 sv 是支持向量集,$|sv|$ 为 sv 的基数:

由 $\boldsymbol{\omega}^*=\sum_{k=1}^{p}\alpha_k^*y_k\boldsymbol{x}_k$,得到决策函数:

$$f(x)=\operatorname{sgn}\Big(\sum_{k\in sv}y_k\alpha_k^*K(\boldsymbol{x}_k,\boldsymbol{x}_j)+\bar{b}^*\Big)$$

输入 α_k^*,b^* 和测试向量 x_u 到决策函数中的输出 $f(x)$,如果任意一个 $f(x_u)=y_u$,则表示所有核函数、训练集、测试集和惩罚系数的选择都是合适的,输出最终参数 α_k^* 和 b^*,建立最优分类函数,即最优超平面方程决策函数,随后输入剩余的 $(s-p)$ 个决策单元样本数据到决策函数中,并输出其分类结果。以此构建出非期望 SBM-SVM 的一个新分类方法。

3 参数优化及模型评估方法

3.1 优化模型

支持向量机的性能和预测精确度受到惩罚因子 C 和核函数参数 g 取值大小的影响,本文选用智能优化算法调整 SVM 模型中的惩罚因子和核函数参数值,运用“试错法”、粒子群算法(PSO)、遗传算法(GA)对非期望 SBM-SVM 模型进行优化改进。粒子群算法(PSO)源于模拟鸟群捕食行为,通过群体中个体间的协作和信息共享来寻找最优解,粒子通过跟踪个体极值(p_{best})和全局极值(g_{best})两个极值来更新自己,找到两个最优值后,通过下列公式来更新自己的速度和位置[7]:

速度更新公式:

$$V_i=\boldsymbol{\omega}\times V_i+c_1\times rand(\,)\times(p_{\mathrm{best}_i}-X_i)+c_2\times rand(\,)\times(g_{\mathrm{best}}-X_i)$$

其中,$\boldsymbol{\omega}$ 为惯性因子,V_i 为第 i 个粒子的速度,X_i 为第 i 个粒子的位置,c_1 和 c_2 为学习因子,p_{best_i} 为第 i 个粒子的历史最优位置,g_{best} 为粒子群的历史最优位置。

位置更新公式:$X_i=X_i+V_i$。

其更新步骤如下:首先对粒子群进行初始化处理;计算每个粒子的适应度值;比较每个粒子的适应度值和个体极值,若适应度值大于个体极值,则用适应度值替换;再对每个粒子的适应度值和全局极值进行比较,若适应度值大于全局极值,则用适应度值替代;根据速度更新公式和位置更新公式更新粒子的速度和位置;经过上述操作,直到到达设定的迭代次数 T,终止运算,最终输出最优解 C 和 g。

粒子群算法与非期望 SBM 方法和 SVM 模型结合的流程图如图 1 所示。

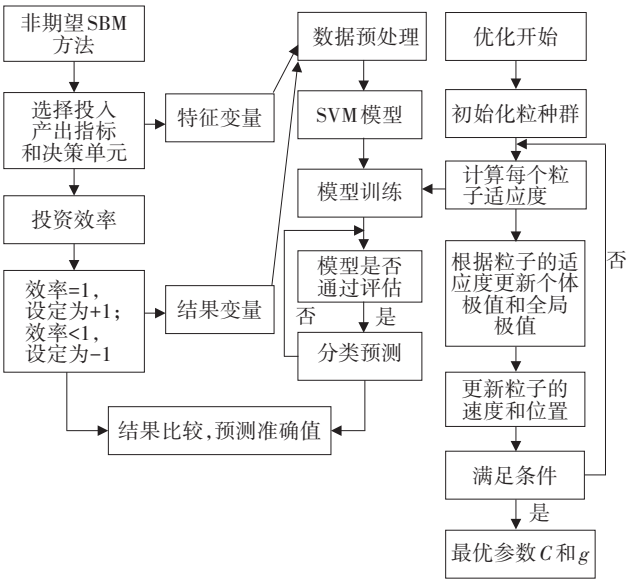

图 1 PSO 优化非期望 SBM-SVM 模型的流程图

Fig. 1 Flow chart of PSO optimizing undesirable SBM-SVM model

3.2　模型评估方法

在做分类时,需要对模型效果好坏进行评估,本文选择预测准确率、ROC 曲线图、准确率和召回率来反映 SVM 模型分类结果的情况,首先需要构建混淆矩阵,如表 1。

表 1　混淆矩阵

Table 1　Confusion matrix

预　测	实际无效	实际有效
预测无效	*TN*	*FN*
预测有效	*FP*	*TP*

(1) 预测准确率。预测准确率指准确判定出决策单元属于预期某一类结果的概率,其值越高,则表示准确预测的概率就越好,分类效果就越好。在表 1 的混淆矩阵中,其正确预测和错误预测的决策单元个数为矩阵交叉项,正确预测个数为 $TN+TP$,预测准确率值为

$$A_{\text{ccuracy}}=\frac{TN+TP}{TN+FN+FP+TP}$$

(2) ROC 曲线图及 AUC 值。对于分类器的优劣评价,通常会采用 ROC 曲线及 AUC 值,其中 AUC 值是 ROC 曲线下的面积,表示分类器对判断预期结果的预测能力,ROC 曲线下的面积越大,则值越大,说明预期结果的判断能力越好,分类器的效果也就越好。

(3) 精确率和召回率。在评估分类模型时,仅使用预测准确率不能完全判定模型的优劣。依据分类数据集的结构情况,判断分类的精确率和召回率,分别对投资有效性和投资无效性预测进行验证。

无效性的精确率为 $P_1=\frac{TN}{TN+FN}$,召回率为 $R_1=\frac{TN}{TN+FP}$;有效性的精确率为 $P_2=\frac{TP}{FP+TP}$,召回率为 $R_2=\frac{TP}{TP+FN}$。为了同时兼顾精确率和召回率,考虑使用调和平均数作为考量两者平衡的综合性指标,计算出每种类别的调和平均数,其值在 0~1 之间,值越接近于 1 越好;$F_{\text{measure}}=\frac{2\times P_i\times R_i}{P_i+R_i}\times 100\%$。

4　实证研究

4.1　指标的建立与决策单元的选取

首先构建非期望 SBM 模型,建立 SBM 模型的输入、输出指标和决策单元的输入、输出指标。从对工业的投入和产出两个方面出发,选取 2011—2020 年重庆市相关指标,投入包括资本投入、劳动投入和能源投入,由固定资产代表资本的投入,应付职工薪酬代表劳动资金的投入,综合能源消费量作为能源的投入。由工业各行业的总产值、利润总额反映期望产出的情况。在进行劳务活动和产品产出的过程中,会产生对环境污染的排放。其中包括废水、废气、废弃物的排放,选择用重庆市工业各行业产生的废水排放量表示废水排放情况,二氧化硫和粉尘排放量表示废气排放情况,工业固体废物产生量表示废弃物排放情况,并用熵值法将以上指标综合为一个环境污染综合指数来代表工业环境污染排放物产出情况,此环境污染综合指数为非期望产出指标。

依据《国民经济行业分类标准》,将工业行业分为 41 个大类,其中采矿业中的开采辅助活动和其他采矿业近十年来未有相关数据发布,剔除这两个行业,本文将工业分为 39 个行业作为决策单元。要求决策单元的数量大于或等于投入和产出指标的数量之和的两倍,此决策单元的选取符合模型的要求。

由于工业小企业缺乏清晰完整的财务报表,开展相关的统计工作较为困难,一年的主营业务收入为 2 000 万元及其以上的工业单位称为规模以上工业企业,此类企业的投资、成本、收入、耗能、排放量等相关数据具有准确性和完整性,并且统计局发布的工业行业数据都为规模以上工业企业的数据,因此,以上输入输出指标都选取为重庆市规模以上工业企业的相关数据。

投入指标:按行业分固定资产 X_1(万元)、按行业分应付职工薪酬 X_2(万元)、按行业分综合能源消费量 X_3(吨标准煤)。

产出指标(期望):按行业分总产值 Y_1(万元)、按行业分利润总额 Y_2(万元)。

产出指标(非期望):按行业分工业环境污染综合指数 Y_3。

决策单元:重庆市工业各行业分类依据《国民经济行业分类》分为 39 类。

为了剔除价格变动的影响,将以上价格相关指标以 2011 年为基期进行平减处理,对固定资产指标进行平减的指数为固定资产价格指数;对应付职工薪酬进行平减处理的指数采用居民消费价格指数;采用工业生产者出厂价格指数分别对总产值和利润总额进行平减处理。

4.2　非期望 SBM 模型

本文选择非期望产出的非径向非角度 SBM 模型来评价工业各行业的 DEA 有效性,基于上文的 SBM 模型介绍,运用 Stata16.0 软件,测算出 2011—2020 年重庆市工业各行业的投资效率,表 2 列出 2020 年重庆市工业各行业的投资效率情况。

表2 重庆市2020年工业行业投资效率评价结果

Table 2 Evaluation results of industrial investment efficiency in Chongqing in 2020

行 业	投资效率	行 业	投资效率
煤炭开采和洗选业	0.058 665	化学纤维制造业	1
石油和天然气开采业	1	橡胶和塑料制品业	0.397 275
黑色金属矿采选业	1	非金属矿物制品业	0.397 825
有色金属矿采选业	0.348 366	黑色金属冶炼和压延加工业	1
非金属矿采选业	0.438 099	有色金属冶炼和压延加工业	0.407 647
农副食品加工业	1	金属制品业	0.394 145
食品制造业	0.399 181	通用设备制造业	0.501 876
酒、饮料和精制茶制造业	1	专用设备制造业	1
烟草制品业	0.214 879	汽车制造业	0.533 634
纺织业	0.293 262	铁路、船舶、航空航天和其他运输设备制造业	0.399 415
纺织服装、服饰业	0.288 712	电气机械及器材制造业	1
皮革、毛皮、羽毛及其制品和制鞋业	0.362 051	计算机、通信和其他电子设备制造业	1
木材加工和木、竹、藤、棕、草制品业	1	仪器仪表制造业	0.425 321
家具制造业	0.402 497	其他制造业	0.169 246
造纸及纸制品业	0.310 134	废弃资源综合利用业	0.648 382
印刷和记录媒介复制业	0.775 849	金属制品、机械和设备修理业	1
文教、工美、体育和娱乐用品制造业	0.492 845	电力、热力的生产和供应业	0.089 447
石油、煤炭及其他燃料加工业	0.513 054	燃气生产和供应业	0.592 639
化学原料和化学制品制造业	0.347 153	水的生产和供应业	0.141 934
医药制造业	0.369 614		

数据来源:《重庆统计年鉴》、重庆市统计局。

表2所列是对决策单元的资源配置能力、资源使用效率等多方面能力的综合衡量与评价。将工业划分为39个行业,有11个行业投资效率达到1,其中包括采矿业中2个行业,制造业中9个行业。电力、燃气及水的生产和供应业中没有达到1的行业。Wu[10]和郑建锋[11]等将DEA得出的效率分为4类,效率在0.98~1之间的决策单元为强相对有效,即此类决策单元稍作修改,便可以达到最佳组合配置;效率在0.8~0.98之间的决策单元为相对有效,此类决策单元比上一类决策单元需要多做修改,才能达到最佳组合配置;效率在0.5~0.8之间的决策单元为相对低效,这类决策单元需要重新调整资源配置或者投入产出结构,并需要一定时间来适应新配置;效率在0~0.5之间的决策单元为低效率单元,这一类决策单元需要花费大量的精力和时间来重新修改和调整投入产出。重庆市2020年的工业内部各行业投资效率,强相对有效决策单元有11个,相对低效决策单元有6个,低效率决策单元有22个,煤炭开采和洗选业为整个工业行业中效率最低的决策单元,效率值仅为0.058 665。说明重庆市2020年工业各行业间的差距较大,需要对低效率、相对低效率行业资源投入规模和配置进行调整修改。

从表2可以看出:投资效率最终呈现的评价结果取值范围在0~1之间。软件运算结果显示:有几项决策单元的效率值为0.999 99,近似等于1,因此本文将效率评价结果在0.98~1之间的决策单元设定为投资有效,运用名义数值1代表;将效率评价结果在0~0.98之间的决策单元设定为投资无效,运用名义数值0代表。

4.3 SVM模型

根据上文非期望产出SBM模型的指标选取及效率评价结果,此处选取SBM模型的输入与输出指标为SVM的特征变量指标,即重庆市2011—2020年规模以上工业按行业分的固定资产、应付职工薪酬、总产值、利润总额、工业环境污染综合指数,以上文得出的含有非期望产出的投资效率结果为SVM模型的结果变量,其中将投资效率大于0.98的决策单元标为DEA相对有效,投资效率小于0.98的决策单元标为非DEA相对有效。采用归一化对上列6个因变量指标进行标准化预处理以消除变量间的影响[12],公式如下:

$$A_i=\frac{x_i-\min(x_i)}{\max(x_i)-\min(x_i)}$$

本文运用R语言软件工具,利用重庆市2011—2020年工业各行业指标构建SVM模型,对重庆市工业各行业投资有效性进行分类,采用分层随机抽样进行数据划分,从结果变量的各层面随机抽取75%的数据,从而组合成训练集,则剩余25%的数据为测试集。高

斯核函数在 SVM 模型中应用最为广泛,因此本文选用高斯核函数,将惩罚参数 C 设定为 1,核函数参数 g 设定为默认值,为特征变量维数的倒数,预测结果如表 3 和图 2 所示。

表 3　重庆市投资有效性分类预测结果表

Table 3　Classification prediction results of investment effectiveness in Chongqing

预　测	投资无效(0)	投资有效(1)
投资无效(0)	63	24
投资有效(1)	3	6
准确率	0.718 8	

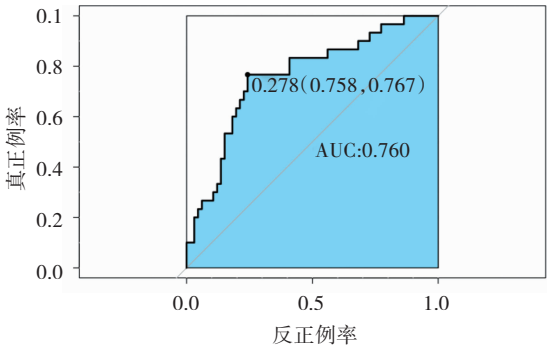

图 2　ROC 曲线图

Fig. 2　ROC curve

最终,基于非期望 SBM-SVM 模型的重庆市工业投资有效性分类预测准确率结果为 71.88%,ROC 曲线中 AUC 值为 0.76,投资有效性精确率为 66.7%,召回率为 20%,其调和平均数为 30.77%;投资无效性的精确率为 72.4%,召回率为 91.3%,其调和平均数为 80.75%。其中分类准确率和 AUC 值还有增大的空间,投资有效性的精确率和召回率都过低,特别是投资有效性的召回率仅为 20%,因此需要对模型进行优化。

投资有效性的召回率仅为 20%,发现样本数据经过有效和无效分类后,存在不平衡情况,因此对样本数据进行人工数据合成处理,将样本分类达到平衡状态。随后加入优化模型对 SVM 模型的惩罚因子 C 和核函数参数 g 进行寻优。本文选用“试错法”、粒子群算法(PSO)、遗传算法(GA)分别对 C 和 g 优化选取最优值,根据众多学者的研究及经验,在“试错法”对支持向量机参数寻优时,惩罚参数 C 和高斯核函数参数 g 取值范围为 $2^{-10}\leqslant C\leqslant 2^{10}$,$2^{-10}\leqslant g\leqslant 2^{10}$,构建出不同的参数组合,采用交叉验证的方法来获得每次组合的错误偏差,最终选取误差最优的参数组合;粒子群算法优化时,参数取值范围为 $0.1\leqslant C\leqslant 10$,$0.1\leqslant g\leqslant 10$;遗传算法优化时,取值范围为 $0.1\leqslant C\leqslant 100$,$0.01\leqslant g\leqslant 10$,染色体数目为 200,交配概率为 0.4,突变概率为 0.01,繁殖次数即循环次数为 100。表 4 和图 3 分别为 3 种方法寻找的参数 C 和 g 的最优值和准确率以及优化后的 ROC 图。

表 4　不同优化模型的预测效果

Table 4　Prediction effects of different optimization models

		试错法	粒子群算法	遗传算法
最优值	C	1 024	3.575 255 8	14.488 004
	g	0.5	0.832 143 1	1.144 978
准确率		0.866	0.887	0.866

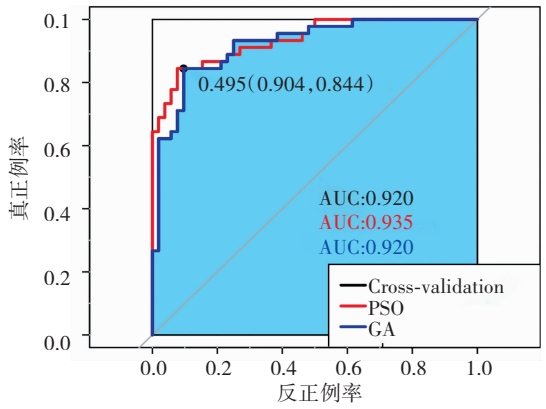

图 3　优化后 ROC 曲线图

Fig. 3　ROC curve after optimization

从表 4 和图 3 的 ROC 图及 3 种优化方法的结果对比可以看出:3 种优化方法寻找 SVM 模型的惩罚参数 C 和核函数 g 对于预测效果具有有效性,3 种方法都提高了预测准确率。试错法寻优后,寻优到的最佳惩罚因子 $C=1\ 024$,核函数参数 $g=0.5$,优化后准确率为 86.6%,ROC 曲线图中的 AUC 值为 0.92;投资有效性的精确率为 100%,召回率为 71.1%,其调和平均数为 83.1%;投资无效性的精确率为 80%,召回率为 100%,其调和平均数为 88.9%。PSO 方法优化后,寻优到的最佳 $C=3.575\ 255\ 8$,$g=0.832\ 143\ 1$,优化后准确率为 88.66%,ROC 曲线图中的 AUC 值为 0.935;投资有效性的精确率为 90.5%,召回率为 84.4%,其调和平均数为 87.3%;投资无效性的精确率为 87.3%,召回率为 92.3%,其调和平均数为 89.7%。GA 方法优化后,寻优到的最佳惩罚因子 $C=14.488\ 004$,核函数参数 $g=1.144\ 978$,准确率为 86.6%,ROC 曲线图中的 AUC 值为 0.92;投资有效性的精确率为 88.1%,召回率为 82.2%,其调和平均数为 85%;投资无效性的精确率为 85.5%,召回率为 90.4%,其调和平均数为 87.9%。其中,PSO 方法寻优效果最佳,分类准确率提高了 16.78%,AUC 值提高了 17.5%,投资有效性精确率与召回率的调和平均数提高了 56.53%,投资无效性精确率与召回率的调和平均数提高了 8.95%,且最优效果

准确率为88.66%,得到了比较理想的效果。说明新构建的非期望SBM-SVM模型通过智能优化算法改进后对与重庆市工业投资有效性的分类研究是具有有效性和实用性的。

5 结 论

本文通过智能优化算法对非期望SBM-SVM模型改进,对惩罚因子C和核函数参数g进行寻优处理,从而提高了模型的分类准确率,提升了模型的性能;随后基于非期望SBM-SVM模型及对其的改进,建立对工业行业投资有效性分类研究的新模型,选取重庆市2011—2020年规模以上工业企业的投资相关指标和工业产出以及环境污染物排放相关指标作为样本数据,将重庆市工业行业划分为39个行业作为决策单元,通过非期望SBM模型得到工业内部各行业的评价效率,将投资效率在0.98到1之间为名义数值1,作为投资有效代表,0到0.98之间为名义数值0,作为投资无效代表,把效率分为DEA相对有效和非DEA相对有效两类,由非期望SBM模型的投入和产出指标作为特征变量,两类评价效率作为结果变量,构建SVM模型,对重庆市工业投资有效性进行分类研究,SVM模型高斯核函数的预测结果为71.88%;本文选择"试错法"、粒子群算法(PSO)、遗传算法(GA)优化模型选取SVM模型的惩罚因子C和核函数参数g的最优值,最终结果表示PSO算法寻优的效果最佳,预测准确度优化到了88.66%。对非期望SBM-SVM模型改进后,模型的准确率、AUC值、精确率和召回率及调和平均值都得到了提升,并达到了平衡,说明通过智能优化算法对模型的改进提升了模型的性能。投资有效性预测结果表明:采用构建的新的非期望SBM-SVM模型对其改进优化后,进行工业行业间投资有效性分类,具有一定的可行性和适用性。

参考文献(References):

[1] SONG J, ZHANG Z. Oil refining enterprise performance evaluation based on DEA and SVM[C]//Second International Workshop on Knowledge Discovery and Data Mining. IEEE Computer Society, 2009: 423—426.

[2] 冉茂盛,周姝,黄凌云. 基于DEA指标的SVM模型在财务预警中的应用[J]. 统计与决策,2009(20):143—145.
RAN Mao-sheng, ZHOU Shu, HUANG Ling-yun. Application of SVM model based on DEA index in financial early warning [J]. Statistics & Decision, 2009 (20): 143—145.

[3] 李宁,杨印生. 基于SVM分类器的平行链式DEA企业绩效评价模型与应用研究[J]. 工业工程,2013,16(4):56—61.
LI Ning, YANG Yin-sheng. Enterprise performance evaluation by using parallel DEA model and SVM classifier[J]. Industrial Engineering Journal, 2013, 16 (4): 56—61.

[4] ZHANG Q, WANG C. DEA efficiency prediction based on IG-SVM [J]. Neural Computing and Applications, 2019, 31(12): 8369—8378.

[5] 李玉龙,李忠富. 基于DEA和神经网络集成模型的我国基础设施投资有效性预测研究[J]. 运筹与管理,2011,20(6):88—98.
LI Yu-long, LI Zhong-fu. Combined DEA and neural network for predicting investment validity of infrastructure on China [J]. Operations Research And Management Science, 2011, 20 (6): 88—98.

[6] ZHU N, ZHU C, EMROUZNEJAD A. A combined machine learning algorithms and DEA method for measuring and predicting the efficiency of Chinese manufacturing listed companies[J]. Journal of Management Science and Engineering, 2020, 6(4): 435—448.

[7] 徐晓明. SVM参数寻优及其在分类中的应用[D]. 大连:大连海事大学,2014.
XU Xiao-ming. SVM parameter optimization and its application in classification[D]. Dalian: Dalian Maritime University, 2014.

[8] 颜薇. 支持向量机优化模型及其应用[D]. 长沙:湖南师范大学,2016.
YAN Wei. Support vector machine optimization model and its application[D]. Changsha: Hunan Normal University, 2016.

[9] TONE K. A slacks-based measure of super-efficiency in data envelopment analysis [J]. European Journal of Operational Research, 2002, 143(1): 32—41.

[10] WU D, YANG Z, LIANG L. Using DEA-neural network approach to evaluate branch efficiency of a large Canadian bank[J]. Expert Systems with Applications, 2006, 31(1): 108—115.

[11] 郑建锋,王应明. 基于DEA-BP神经网络的效率置信区间预测模型研究[J]. 计算机工程与应用,2021,57(3):273—278.
ZHENG Jian-feng, WANG Ying-ming. Research on efficiency confidence interval prediction model based on DEA-BP neural network[J]. Computer Engineering and Applications, 2021, 57 (3): 273—278.

[12] 吴荣火,钟德炎,范明丽,等. 基于SVM的广西宜居城市分类预测模型及R语言实现[J]. 玉林师范学院学报,2019,40(5):26—34.
WU Rong-huo, ZHONG De-yan, FAN Ming-li, et al. Classification prediction model of livable cities in Guangxi and implementation of R language based on SVM [J]. Journal of Yulin Normal University, 2019, 40 (5): 26—34.

责任编辑:李翠薇