文章编号:1672－9331(2021)02－0049－08

基于 RF-PSO-LSSVM 的高层建筑项目工期风险预测

刘伟军,赵　威

(长沙理工大学 交通运输工程学院,湖南 长沙　410114)

摘　要:针对高层建筑项目在工期风险预测时样本数据少且特征维度高的特点,建立了利用随机森林(random forest,RF)算法和粒子群(particle swarm optimization,PSO)算法优化最小二乘支持向量机(least squares support vector machine,LSSVM)的高层建筑项目工期风险预测模型。采用在特征选择方面具有显著优势的 RF 算法筛选出最佳特征子集;利用 PSO 算法对 LSSVM 的正则化参数和核函数参数进行优化;采用精确率、召回率以及 F_{1m} 值对所建立模型的预测性能进行验证与评估。研究结果表明:用所建立的模型对高层建筑项目进行工期风险预测,平均精确率达到了 93.71%,平均召回率达到了 94.04%。该模型能够准确预测高层建筑项目工期的风险等级,进一步完善了高层建筑项目工期风险的预测方法,其预测结果可为高层建筑项目控制工期风险提供一定的参考。

关键词:高层建筑项目;工期风险预测;随机森林算法;粒子群算法;最小二乘支持向量机

中图分类号:TU974;TP399　　**文献标志码**:A

Risk prediction of high-rise building project duration based on RF-PSO-LSSVM

LIU Wei-jun, ZHAO Wei

(School of Traffic and Transportation Engineering, Changsha University of Science & Technology, Changsha 410114, China)

Abstract: Aiming at the characteristics of small sample data and high feature dimension in risk prediction of high-rise building project duration, a risk prediction model of high-rise building project duration was proposed based on least squares support vector machine(LSSVM) optimized by random forest(RF) algorithm and particle swarm optimization(PSO) algorithm. RF algorithm with obvious advantages in feature selection was used to select the best feature subset. PSO algorithm was used to optimize the regularization parameters and kernel function parameters of LSSVM. The precision rate, recall rate and F_{1m} value were used to verify and evaluate the predictive performance of the proposed model. The research results show that the average precision rate reaches 93.71%, and the average recall rate reaches 94.04% to predict the risk of high-rise building project duration using the proposed model. The proposed model can accurately predict the risk level of high-rise building project duration, im-

收稿日期:2020-12-25

基金项目:河南省交通运输厅科技项目(2014G25);湖南省自然科学基金资助项目(2020JJ4629)

通讯作者:刘伟军(1975－),男,副教授,主要从事公路工程造价管理与项目管理方面的研究。

E-mail: liuwesley@163.com

proves the risk prediction method of high-rise building project duration, and the predicted results would provide a reference for risk control of high-rise building project duration.

Key words: high-rise building project; risk prediction of duration; random forest algorithm; particle swarm optimization algorithm; least squares support vector machine

近几十年来,各国迅速增多的高层建筑有效缓解了城市用地紧张的问题。与普通建筑项目相比,高层建筑项目具有规模大、投资多、结构复杂、施工技术与组织难度高等特点,由此导致工期延误的现象时有发生,工期风险较大。因此,对高层建筑项目的工期风险进行预测具有一定的现实意义。

目前,用于工程项目工期风险预测的方法有蒙特卡洛仿真[1]、贝叶斯信念网络[2]等,但这些方法都有一定的模糊性和局限性,易受分析过程的随机性和人为主观因素的影响。近年来,国内外学者已逐步将机器学习算法用于工程项目工期风险的预测。Gondia等[3]提出了基于朴素贝叶斯和决策树的工程项目工期风险预测模型;Yaseen等[4]构建了基于遗传算法和随机森林(random forest,RF)算法的工程项目工期风险预测模型;Leu等[5]基于1 538组工业建筑案例,将主成分分析(principle component analysis,PCA)法和BP(back propagation)神经网络相结合对台湾工业建筑的工期风险进行预测;El-kholy[6]提出了概率神经网络、广义回归神经网络等4种人工神经网络算法,并对高速公路的工期和成本风险进行了预测。上述方法虽能在一定程度上减少人为因素的影响,并能对研究对象进行快速地分类和预测,但在解决具有小样本、高维度等特征的预测或分类问题时仍然存在不足[7-8]。

本研究采用具有较强泛化能力、针对小样本数据即可分类、预测的最小二乘支持向量机(least squares support vector machine,LSSVM)对高层建筑项目的工期风险进行预测。同时采用在数据维度较高时可以发挥优势的RF算法对样本数据进行降维处理,将选出的重要特征作为LSSVM的输入变量。此外,为解决LSSVM参数选择困难的问题,采用具有较强全局优化能力的粒子群(particle swarm optimization,PSO)算法对LSSVM的正则化参数和核函数参数进行寻优,建立一种基于RF-PSO-LSSVM的高层建筑项目工期风险预测模型。通过工程实例分析及与其他模型的性能对比,验证本研究所建模型的准确性和有效性。该模型可为高层建筑项目工期风险的预测提供一种新的思路。

1 工程实例获取及数据预处理

1.1 工程实例获取

本研究采用的48组高层建筑项目工期延误案例均来自文献[9]。首先,为了选取合理的工期风险评价指标,在分析现有研究成果的基础上结合专家实践经验,对高层建筑项目工期风险因素进行识别,并构建高层建筑项目工期风险评价指标体系,其包含9个一级评价指标,36个二级评价指标[9]。然后采用调查问卷方式,由高层建筑项目的管理人员对项目中的每个风险因素发生的可能性和影响程度进行打分。采用李克特5分量表法对问卷中的问项进行打分,1~5分分别表示非常低、低、中等、高和非常高。去除无效问卷,最终共得到48份有效问卷。最后将每个风险因素发生的可能性得分和影响程度得分相乘,得到该风险因素的风险值。对各个风险因素的风险值求平均,得到该项目所有风险因素的平均风险水平。在此基础上,明确风险等级划分标准[3,9],建立风险矩阵,如图1所示。根据工期的延误程度,将上述48组高层建筑项目的工期风险等级划分为3类:低风险(13组)、中等风险(26组)及高风险(9组)。

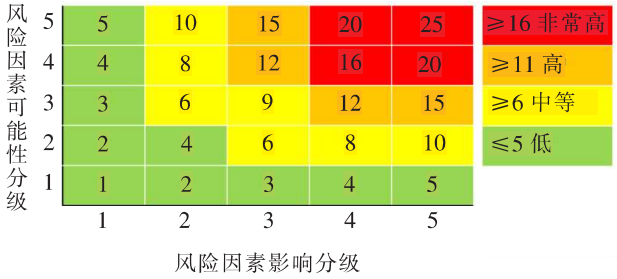

图1 风险矩阵

Fig. 1 Risk matrix

1.2 数据预处理

1.2.1 缺失数据的处理

在48组案例中有小部分数据缺失,为了保留

尽量多的样本，对样本中缺失的数据进行空值填充处理。通过对比均值、众数、中位数等样本特征数的填充效果，最终选择用同类完整样本的均值对缺失数据进行填充。

1.2.2　特征提取和特征选择

由于样本数据的维度较高，在对模型训练前需对数据集进行降维处理。目前主流的降维方法有特征提取和特征选择两类，它们主要的区别在于是否会产生新的特征。PCA 法作为一种经典的特征提取方法常被用于高维数据的降维。它通过对原特征集进行一系列变换产生新的特征空间，将具有相关性的特征变量转换为彼此相互独立的主成分，同时能尽量多地保留原始变量的信息[10]。因此，本研究采用 PCA 法对样本数据进行特征提取，并用 SPSS 软件计算各个主成分对工期风险影响因素的样本数据进行解释的特征值及累积方差贡献率，结果如图 2 所示。由图 2 可知，前 16 个主成分的累积方差贡献率为 86.48%，大于 85%，样本数据的维度由 36 维降至 16 维。可见，在采用 PCA 法降维后数据的维度依然较高。

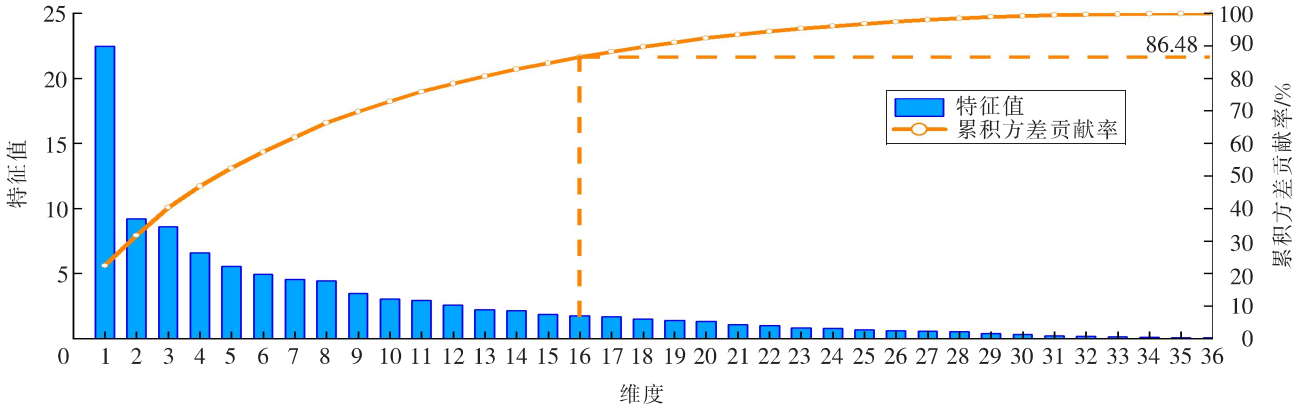

图 2　主成分分析结果

Fig. 2　Analysis results of principal component

与特征提取相比，特征选择是从原特征集中选择相对重要的特征子集，在降低数据维度的同时不改变原始特征的意义。对于本研究的研究对象，采用特征选择方法有助于找出影响高层建筑项目工期风险的重要风险因素。因此，尝试采用 RF 算法对样本数据进行特征选择。

RF 算法是数据降维处理、特征选择的代表算法，它的采样方式被称为 bootstrap 法[11]。在使用 bootstrap 法进行采样的过程中，约有 36.8%的样本数据不会被抽取，这部分样本数据被形象地称为袋外数据（out of bag，OOB）。OOB 可用于评估决策树的性能，计算模型的预测错误率。给某个特征随机添加噪声，若其 OOB 的错误率大幅上升，则表明该特征对分类结果影响较大，是一个比较重要的特征[7]。采用 RF 算法对样本数据进行特征选择的具体步骤如下：

① 假设随机森林中有 k 棵决策树，为随机森林中的每棵决策树选择相应的袋外数据，并计算它们的误差值，记为 $E_{\mathrm{OOB},1}$，$E_{\mathrm{OOB},2}$，…，$E_{\mathrm{OOB},k}$。

② 对 OOB 全部样本的第 i 个特征随机添加噪声，并保证其余特征不变。然后重新计算 OOB 的误差，记为 $E_{i,1}$，$E_{i,2}$，…，$E_{i,k}$。

③ 用下式计算特征重要性，并基于重要性对特征进行排序：

$$S_i=\frac{1}{k}\sum_{m=1}^{k}(E_{i,m}-E_{\mathrm{OOB},m}) \tag{1}$$

随机森林特征重要性排序结果如图 3 所示。为选取最佳的特征子集数目，根据特征重要性排序采用随机森林分类器对特征子集从 top1 至 top36 逐一累加，并进行十折交叉验证测试，然后选出结果最好的一组特征子集。特征变量个数与总体分类精度、Kappa 系数的关系如图 4 所示。

由图 4 可知，随着参与分类的特征变量数目的增加，前期（1～5 个特征）分类精度呈现快速上升的趋势，总体分类精度从单个特征的 66.5%迅速上升至 84.0%。这是因为前 5 个特征的重要性得分较高，特征之间的相关性较小，且冗余特征少。中期（5～10 个特征）分类精度呈现平稳趋势。后期（11～36 个特征）分类精度呈现缓慢下降的趋

势,这是因为在后期冗余特征数目有所增加。当特征变量的数目为5～10个时,分类精度较高(在89.5%上下波动),Kappa系数也较大(在0.82上下波动),样本数据的维度由开始的36维降至5维。由此可见,RF算法对本研究数据集的降维效果明显优于PCA的降维效果。所以,本研究选择RF算法进行特征选择,并优选出5个重要特征作为后续数据处理过程的输入变量。

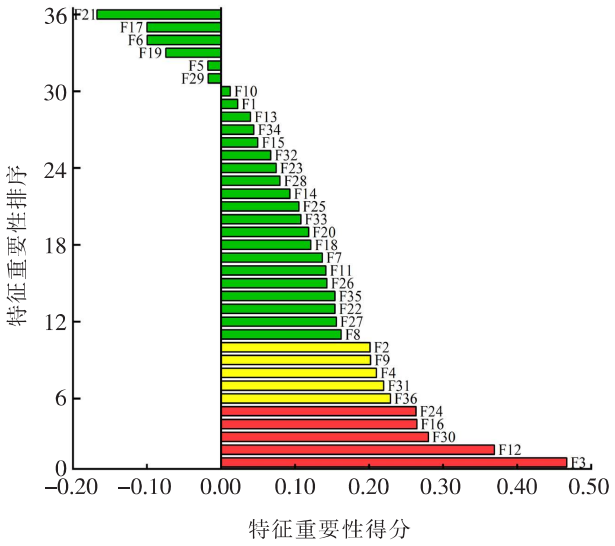

图3 随机森林特征重要性排序结果

Fig. 3 Ranking results of RF feature importance

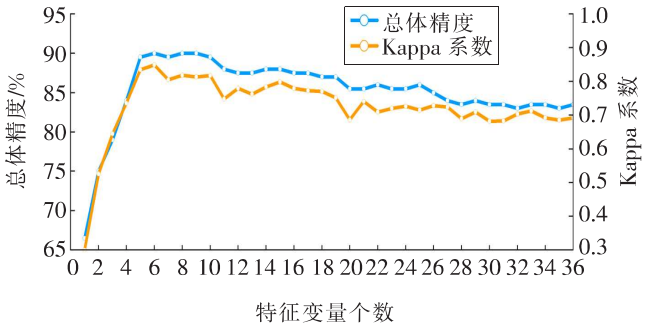

图4 特征变量个数与分类精度、Kappa系数的关系

Fig. 4 Relationship of number of feature variables with classification accuracy and Kappa coefficient

1.2.3 数据归一化和过采样处理

为了消除样本数据的量级给模型预测结果带来的影响,需对样本数据进行归一化处理[10]。归一化处理的公式如下式所示:

$$y=\frac{x-x_{\min}}{x_{\max}-x_{\min}} \tag{2}$$

式中:y 为归一化后的数据;x 为归一化前的数据;$x_{\min}$ 为数据的最小值;$x_{\max}$ 为数据的最大值。

为了提高模型的泛化能力,防止模型过拟合和欠拟合,对归一化后的数据集进行随机打乱处理,并在此基础上以4∶1的比例对数据集进行划分。训练样本用于模型训练,测试样本用于检验、评估已完成训练模型的可靠性和泛化性能。

此外,所收集的数据集中3个类别的数据不均衡,低风险样本数量及高风险样本数量明显少于中等风险样本数量。因此,采用SMOTE(synthetic minority oversampling technique)算法在少数类样本和其近邻样本的连线上进行线性插值,生成新的样本。插值公式如下:

$$x_{\text{new}}=x+r(y_i-x) \tag{3}$$

式中:r 为0～1的一个随机数;$i=1,2,\cdots,N$;x 为少数类样本;x_{new} 为增加的新样本;y_i 为 x 的第 i 个近邻样本。

将新生成的样本加入原训练数据集中,直至训练数据集中3类工期风险的实例频数达到均衡,由此可以解决非均衡数据集给模型训练带来的问题[12]。

2 预测模型

2.1 LSSVM

LSSVM是对标准支持向量机的一种重要改进。它通过将支持向量机中的不等式约束改为等式约束,避免了求解复杂的二次规划问题,在计算过程中能加快收敛速度,并能提高计算结果的精度[13]。具体步骤如下:

① 采用非线性映射函数 $\varphi(x)$ 将训练样本集 $\{x_i,y_i\}$(x_i 为输入变量,y_i 为输出变量,$i=1,2,\cdots,N$)中的数据转换到高维特征空间中。在高维特征空间中,样本集可表示为如下映射关系:

$$f(x)=\boldsymbol{\beta}^{\mathrm{T}}\varphi(x)+b \tag{4}$$

式中:$\boldsymbol{\beta}$ 为权向量;b 为偏置项。

② 依据结构风险最小化原则,将不等式约束问题转换为等式约束问题,即:

$$\min\ g(\boldsymbol{\beta},e)=\frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}+\frac{c}{2}\sum_{i=1}^{n}e_i^2 \tag{5}$$

$$\text{s.t.}\ \ y_i=\boldsymbol{\beta}^{\mathrm{T}}\varphi(x)+b+e_i \tag{6}$$

式中：c 为正则化参数；e_i 为误差变量，且 $e_i \in \mathbf{R}$，$\mathbf{R}$ 为实数集。

③ 对目标函数建立拉格朗日等式：

$$L(\boldsymbol{\beta},b,e,\theta)=g(\boldsymbol{\beta},e)-\sum_{i=1}^{n}\theta_i(\boldsymbol{\beta}^{\mathrm{T}}\varphi(x_i)+b+e_i-y_i) \tag{7}$$

式中：θ_i 为拉格朗日乘子，且 $\theta_i \in \mathbf{R}$。

④ 依据 KKT（Karush-Kuhn-Tucher）条件，对 L 函数中的各变量求偏导，然后消去 $\boldsymbol{\beta}$ 和 e，将目标优化的极小值问题转换成线性优化问题，通过求解，得到最终的决策函数，即：

$$f(x)=\sum_{i=1}^{n}\theta_i K(\boldsymbol{x},\boldsymbol{x}_i)+b \tag{8}$$

式中：$K(\boldsymbol{x},\boldsymbol{x}_i)$ 为核函数。

选择径向基函数作为模型的核函数，其函数表达式为：

$$K(\boldsymbol{x},\boldsymbol{x}_i)=\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}_i\|^2}{\sigma^2}\right) \tag{9}$$

式中：σ 为核函数参数；$\boldsymbol{x}$ 为 m 维输入向量；$\boldsymbol{x}_i$ 为第 i 个径向基函数的参数，且与 $\boldsymbol{x}$ 的维数相同；$\|\boldsymbol{x}-\boldsymbol{x}_i\|$ 为 $\boldsymbol{x}-\boldsymbol{x}_i$ 的范数，表示 $\boldsymbol{x}$ 与 $\boldsymbol{x}_i$ 之间的距离。

LSSVM 的泛化性能受正则化参数及核函数参数取值的影响较大，为了进一步提升 LSSVM 的泛化能力，采用 PSO 算法对正则化参数和核函数参数进行寻优。

2.2　PSO 算法

PSO 算法是 Eberhart 等人受鸟类捕食行为启发而提出的一种智能优化算法[14]。首先利用 PSO 算法对一群随机产生的粒子进行初始化，每个粒子都代表优化问题的一个潜在最优解；然后通过跟踪个体与群体最优解来更新自身的位置及速度，以此来实现全局寻优。假设粒子的运动速度为 $\boldsymbol{V}$、位置为 $\boldsymbol{X}$，决策变量的维数为 d，则第 i 个粒子的参数为：

$$\boldsymbol{X}_i=(x_{i1},x_{i2},\cdots,x_{id}) \tag{10}$$

$$\boldsymbol{V}_i=(v_{i1},v_{i2},\cdots,v_{id}) \tag{11}$$

$$\boldsymbol{V}_{id(t+1)}=\omega\cdot\boldsymbol{V}_{idt}+c_1r_1(\boldsymbol{p}_{idt}-\boldsymbol{X}_{idt})+c_2r_2(\boldsymbol{p}_{gdt}-\boldsymbol{X}_{idt}) \tag{12}$$

$$\boldsymbol{X}_{id(t+1)}=\boldsymbol{X}_{idt}+\boldsymbol{V}_{id(t+1)} \tag{13}$$

式中：$\boldsymbol{p}_{idt}$，$\boldsymbol{p}_{gdt}$ 为 t 时刻个体与群体经历过的最佳位置；ω 为惯性权重；r_1，r_2 为[0,1]中的随机数；c_1，c_2 为加速度常数；t 为当前的迭代次数。PSO 算法优化 LSSVM 参数的流程见图 5。

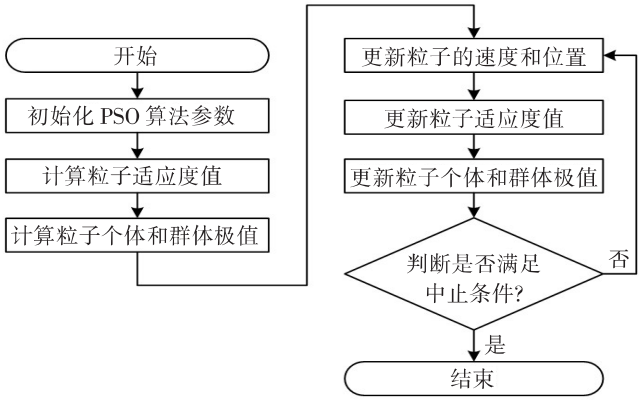

图 5　用 PSO 算法优化 LSSVM 参数的流程

Fig. 5　Optimizing process of LSSVM parameters using PSO algorithm

用 PSO 算法优化 LSSVM 参数的步骤如下：

① 初始化粒子的初始速度与位置以及群体规模、最大迭代次数、加速度常数等参数。

② 依据分类性能的评价函数，计算各个粒子的适应度值。

③ 将每个粒子当前位置的适应度值同其历史最佳位置 $\boldsymbol{p}_{idt}$ 的适应度值进行对比，如果更优，则用当前位置更新粒子最优位置，否则维持不变。

④ 将每个粒子当前位置的适应度值同群体最佳位置 $\boldsymbol{p}_{gdt}$ 的适应度值进行对比，如果更优，则用当前位置更新群体最优位置，否则维持不变。

⑤ 按照式(12)～(13)更新粒子的速度和位置。

⑥ 判断是否满足寻优中止条件，如果满足则求出最优解，如果不满足则转至步骤②。

2.3　预测模型的构建流程

本研究将 RF 算法优化选择特征变量的能力、LSSVM 解决小样本及非线性等问题的优势及 PSO 对 LSSVM 正则化参数和核函数参数的全局寻优能力结合在一起，构建基于 RF-PSO-LSSVM 的高层建筑项目工期风险预测模型，构建流程如图 6 所示。

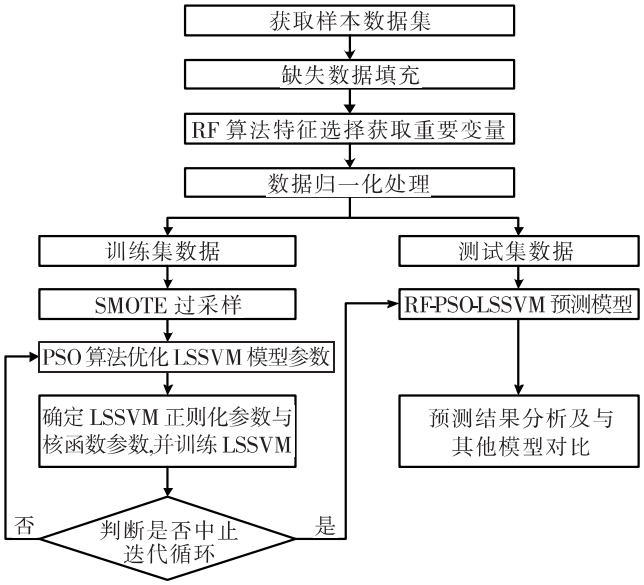

图6 基于RF-PSO-LSSVM的高层建筑项目工期风险预测模型构建流程

Fig. 6 Construction process of duration risk prediction model for high-rise building project based on RF-PSO-LSSVM

3 工程实例分析

3.1 模型的构建及参数设置

基于MATLAB R2018b平台和LSSVM工具箱,载入48组样本数据,按照上述流程随机选取训练样本,构建RF-PSO-LSSVM高层建筑项目工期风险预测模型。为验证模型的有效性,同时构建PSO-LSSVM,PCA-PSO-LSSVM,LSSVM,PCA-LSSVM,RF-LSSVM等模型进行对比分析。PSO算法的参数设置如下:粒子个体的维度为2,种群数量为20,最大迭代次数为200,参数搜索范围为$c\in[0.01,100]$,$\sigma\in[0.1,10]$,加速度常数$c_1=c_2=1.5$,惯性权重$\omega=1$。

3.2 评价指标选取

为了评估分类模型的泛化能力,本研究采用精确率、召回率以及F_1值综合评估模型的性能。精确率表示在所有被预测为正例的样本中实际类别为正例的样本的概率;召回率表示实际类别为正例的样本被预测为正例的概率;F_1值则为两者的调和均值。由于本研究所选取的样本数据共有3个类别,因此需要计算三者的宏平均值或微平均值。本研究选择计算三者的宏平均值,计算公式为:

$$P=\frac{T_P}{(T_P+F_P)} \tag{14}$$

$$P_m=\frac{1}{n}\sum_{i=1}^{n}P_i \tag{15}$$

$$R=\frac{T_P}{(T_P+F_N)} \tag{16}$$

$$R_m=\frac{1}{n}\sum_{i=1}^{n}R_i \tag{17}$$

$$F_1=\frac{2T_P}{(2T_P+F_P+F_N)} \tag{18}$$

$$F_{1m}=\frac{1}{n}\sum_{i=1}^{n}F_{1,i} \tag{19}$$

式中:T_P为分类正确的正例数;F_P为分类错误的正例数;F_N为分类错误的反例数;m表示该值为宏平均值;n为类的数量,在本研究中n值取为3。

3.3 模型预测结果分析

将测试样本载入上述训练好的模型进行预测。为了消除试验随机性引起的偏差,采用10次随机划分、重复进行的试验结果的均值作为最终的预测结果,并根据3种模型的评价指标比较各模型的预测能力,结果如表1所示。其中,A表示模型的输入变量为全部特征,PCA表示模型的输入变量是用PCA法提取的主成分,RF-FS表示模型的输入变量是用RF算法选择的重要特征。

表1 模型预测结果对比

Table 1 Comparison of model prediction results %

模型		精确率P_m	召回率R_m	F_{1m}
PSO-LSSVM	A	88.11	87.63	87.73
	PCA	90.70	90.00	88.67
	RF-FS	93.71	94.04	93.25
LSSVM	A	83.33	83.50	83.29
	PCA	85.10	83.33	80.03
	RF-FS	87.88	85.19	83.81

由表1可知,用RF算法进行特征选择得到的重要特征作为模型的输入变量时,两种模型的泛化性能都得到了明显提升。与采用全部特征作为输入变量相比,RF-PSO-LSSVM模型的精确率提高了5.60%,召回率提高了6.41%,RF-LSSVM模型的精确率提高了4.55%,召回率提高了1.69%;与采用PCA法提取的主成分作为输入变量相比,RF-PSO-LSSVM模型的精确率提高了3.01%,召回率提高了4.04%,RF-LSSVM模型

的精确率提高了 2.78%，召回率提高了 1.86%。可见 RF 算法在有效降低数据维度的同时，大大提升了模型的预测性能，而且采用 RF 算法进行特征选择还有助于项目管理人员找出影响高层建筑项目工期风险的主要风险因素，及时监控和规避相关风险，而 PCA 法生成的主成分具有一定的模糊性，可理解程度不高。

此外，PSO-LSSVM 模型的预测性能明显优于 LSSVM 模型的。在输入变量为全部特征、PCA 法提取的主成分以及 RF 特征选择的重要特征 3 种情景下，与未经 PSO 算法优化的 LSSVM 模型相比，经 PSO 算法优化的 LSSVM 模型的精确率分别提高了 4.78%，5.60% 和 5.83%，召回率分别提高了 4.13%，6.67% 和 8.85%。可见，PSO 算法具有较好的全局搜索能力，能避免过早地陷入局部最优，并能精确地对 LSSVM 的正则化参数和核函数参数进行优化，从而进一步提高 LSSVM 的分类精度。

为了进一步验证 RF-PSO-LSSVM 模型在预测高层建筑项目工期风险方面的优越性，同时采用 BPNN，KNN，RF 算法对高层建筑项目工期风险进行预测。设置 BP 神经网络隐含层的层数为 10，隐含层激活函数选用“Sigmoid”，训练精度为 0.01，学习率为 0.01，最大训练次数为 1 000；设置 KNN 的近邻数 $K=5$，选用欧氏距离评估样本间的距离；设置 RF 算法中决策树的棵数 $n_{tree}=500$。同样采用 10 次随机划分、重复进行的试验结果的均值作为最终的分类结果。结果如表 2 所示。

表 2　本研究模型与 BPNN，KNN，RF 算法预测结果对比

Table 2　Comparison of prediction results of model of this study, BPNN, KNN and RF algorithm

模型		精确率 P_m/%	召回率 R_m/%	F_{1m}/%
PSO-LSSVM	RF-FS	93.71	94.04	93.25
	A	70.38	74.43	68.05
BPNN	PCA	72.22	76.39	73.16
	RF-FS	75.16	77.50	69.83
	A	77.00	80.39	71.78
KNN	PCA	79.72	82.78	78.42
	RF-FS	82.25	82.78	78.72
	A	87.20	87.22	85.25
RF	PCA	88.51	88.95	87.42
	RF-FS	90.71	90.00	89.12

由表 2 可知，本研究提出的 RF-PSO-LSSVM 模型的预测性能最佳，精确率为 93.71%，召回率为 94.04%，F_{1m} 值为 93.25%。其次为 RF 模型的，其预测精确率最高为 90.71%，召回率最高为 90.00%，F_{1m} 值最高为 89.12%。此外，在 3 种不同的输入变量情景下，RF 算法的预测结果上下浮动的范围最小，表现出较高的稳健性。这是因为在模型的构建过程中，RF 算法中的决策树对特征进行随机采样，所以特征数量的变化没有产生很大的负面影响。KNN 模型的总体表现较差，预测精确率最高为 82.25%，召回率最高为 82.78%，F_{1m} 值最高为 78.72%。但 KNN 实现简单，需要调节的超参数较少，在考虑使用更高级的技术之前，此算法是一种较好的基础方法。BPNN 模型的总体表现最差，预测精确率最高为 75.16%，召回率最高为 77.50%，F_{1m} 值最高为 73.16%。这是因为本研究的样本数量无法满足 BPNN 使用大样本数据进行建模的要求，同时 BPNN 设置了较多的参数，而确定这些参数的最优值难度较大。

综上所述，本研究提出的 RF-PSO-LSSVM 模型在高层建筑项目工期风险预测方面具有一定的优越性和较高的准确性，适用于具有小样本和高维度数据特征的高层建筑项目工期风险的预测，其预测结果能为高层建筑项目工期风险预测起到一定的指导作用。

4　结论

针对具有小样本、高维度数据特征的高层建筑项目工期风险的预测问题，本研究提出了一种基于 RF-PSO-LSSVM 的高层建筑项目工期风险预测模型，并得出以下结论：

1）当数据维度较高时，使用 RF 算法进行特征选择可以取得较好的效果。与采用 PCA 法的预测效果相比，使用 RF 算法进行特征选择对数据的降维效果更明显，同时 LSSVM 模型的预测性能也得到了显著提升。

2）采用 PSO 算法对 LSSVM 正则化参数与核函数参数进行寻优，能够有效避免 LSSVM 参数选择的主观性和随机性，有利于提高 LSSVM 对高层建筑项目工期风险预测的准确性。

3）LSSVM 克服了对大样本的依赖性。针对

工程实践中的小样本数据,使用该模型进行工期风险预测具有一定的优越性和更强的适用性。

本研究所建模型可推广到其他类型的工程项目中,可将以往同类已完工程的工期风险数据作为学习样本,训练并构建模型来预测待建项目的工期风险水平,为工期风险预警和控制提供借鉴。

本研究所建模型的局限在于数据预处理过程较烦琐,后续研究可以考虑收集质量更高的数据,或者尝试采用多种集成学习算法进行建模。

〔参考文献〕

[1] Kokkaew N, Wipulanusat W. Completion delay risk management: a dynamic risk insurance approach[J]. Ksce Journal of Civil Engineering, 2014, 18(6): 1 599-1 608.

[2] Leu V T, Kim S Y, Tuan N V, et al. Quantifying schedule risk in construction projects using Bayesian belief networks[J]. International Journal of Project Management, 2009, 27(1): 39-50.

[3] Gondia A, Siam A, El-dakhakhni W, et al. Machine learning algorithms for construction projects delay risk prediction[J]. Journal of Construction Engineering and Management, 2020, 146(1): 1-16.

[4] Yaseen Z M, Ali Z H, Salih S Q, et al. Prediction of risk delay in construction projects using a hybrid artificial intelligence model[J]. Sustainability, 2020, 12(15): 1-14.

[5] Leu S S, Liu C M. Using principal component analysis with a back-propagation neural network to predict industrial building construction duration[J]. Journal of Marine Science & Technology, 2016, 24(2): 82-90.

[6] El-kholy A M. Exploring the best ANN model based on four paradigms to predict delay and cost overrun percentages of highway projects[J]. International Journal of Construction Management, 2019, 19(1): 1-19.

[7] 张雪蕾,汪明,曹寅雪,等. 3种典型机器学习方法在灾害敏感性评估中的对比[J]. 中国安全生产科学技术, 2018, 14(7): 79-85.
ZHANG Xue-lei, WANG Ming, CAO Yin-xue, et al. Comparison of three typical machine learning methods in susceptibility assessment of disasters[J]. Journal of Safety Science and Technology, 2018, 14(7): 79-85.

[8] 姜海龙,李潼清,程浩,等. 基于PSO-LSSVM的高压真空断路器弹簧机构机械故障诊断[J]. 高压电器, 2019, 55(12): 248-255.
JIANG Hai-long, LI Tong-qing, CHENG Hao, et al. Mechanical fault diagnosis for spring mechanism of high-voltage vacuum circuit breaker based on PSO-LSSVM[J]. High Voltage Apparatus, 2019, 55(12): 248-255.

[9] Sanni-anibire M O, Zin R M, Olatunj S O. Machine learning model for delay risk assessment in tall building projects[J]. International Journal of Construction Management, 2020, 20(1): 1-10.

[10] 王首绪,刘嘉瑜. 高速公路子项目对项目群管理绩效的影响[J]. 长沙理工大学学报(自然科学版), 2018, 15(2): 37-42.
WANG Shou-xu, LIU Jia-yu. Influence of expressway's subprojects on programme management performance[J]. Journal of Changsha University of Science & Technology (Natural Science), 2018, 15(2): 37-42.

[11] 张磊,宫兆宁,王启为,等. Sentinel-2影像多特征优选的黄河三角洲湿地信息提取[J]. 遥感学报, 2019, 23(2): 313-326.
ZHANG Lei, GONG Zhao-ning, WANG Qi-wei, et al. Wetland mapping of yellow river delta wetlands based on multi-feature optimization of Sentinel-2 images[J]. Journal of Remote Sensing, 2019, 23(2): 313-326.

[12] 封化民,李明伟,侯晓莲,等. 基于SMOTE和GBDT的网络入侵检测方法研究[J]. 计算机应用研究, 2017, 34(12): 3 745-3 748.
FENG Hua-min, LI Ming-wei, HOU Xiao-lian, et al. Study of network intrusion detection method based on SMOTE and GBDT[J]. Application Research of Computers, 2017, 34(12): 3 745-3 748.

[13] 刘伟铭,雷焕宇,翟聪,等. 基于PSO-LSSVM的高速公路短时行程时间预测[J]. 公路与汽运, 2017, 24(3): 36-39, 48.
LIU Wei-ming, LEI Huan-yu, ZHAI Cong, et al. Expressway short-term travel time prediction based on PSO-LSSVM[J]. Highways & Automotive Applications, 2017, 24(3): 36-39, 48.

[14] 仇一颗,周翔,王家. 基于风险分析的公路工程保险费率厘定[J]. 公路工程, 2020, 45(2): 80-85.
CHOU Yi-ke, ZHOU Xiang, WANG Jia. Highway engineering insurance rate determination based on risk analysis[J]. Highway Engineering, 2020, 45(2): 80-85.