基于机器学习的住宅项目工程造价预测

Housing Project Cost Prediction Based on Machine Learning

唐安彬

Anbin Tang

西南科技大学　中国·四川 绵阳 621010

Southwest University of Science and Technology, Mianyang, Sichuan, 621010, China

摘　要：工程造价行业数字化转型满足建筑行业精细化管理、高质量发展的要求，所以基于机器学习的工程造价预测对工程造价数字化转型具有重要的意义。论文通过对中国成都市住宅项目的工程造价数据进行采集、清洗、降维，并将数据集划分为训练集和验证集。此后建立支持向量机回归、随机森林回归、套索回归和 BP 神经网络模型，利用训练集的数据对机器学习模型训练，最后训练后的模型对验证集的数据进行预测，得到的预测造价与实际造价相差小于 5%，满足工程造价估概算的误差要求。

Abstract: The digital transformation of engineering cost industry meets the requirements of fine management and high-quality development of construction industry, so the project cost prediction based on machine learning is of great significance to the digital transformation of engineering cost. In this paper, the cost data of residential projects in Chengdu, China are collected, cleaned and reduced in dimension, and the data set is divided into training set and verification set. Thereafter based support vector machine (SVM) regression, random forest regression, the lasso regression and BP neural network model, using the training set of training data for machine learning model, after the last training model to forecast the validation set of data, the forecast cost and actual cost far less than 5%, satisfies the requirement of engineering cost estimation of error.

关键词：工程造价预测；随机森林；支持向量机；神经网络；套索回归

Keywords: project cost forecast; random forest; support vector machine; neural network; lasso regression

DOI: 10.12346/etr.v4i3.5797

1 引言

随着中国国名经济和科技水平的不断发展与创新，对工程建筑行业的要求也从粗放型的建筑管理向精细化数字化管理转变。2020 年，住建部针对建设项目工程造价行业今后的发展方向发布了《工程造价改革工作方案》（建办标〔2020〕38 号），文中在主要任务中明确提出要加强工程造价数据积累。通过建立工程造价数据库和造价指标指数库，利用大数据、人工智能等信息化技术实现工程造价的数字化转型，保证行业在新时期的高质量发展。

项目的工程造价预估是项目在决策阶段的重要工作，对后期整体项目的工程项目的成本管控起着至关重要的影响。然而传统的工程造价概算的方法都是基于类似项目进行计算，无法全面准确的对项目造价进行预测。综上，论文将利用机器学习对住宅项目的工程造价进行预测，并对其准确性以及可行性进行分析。

2 理论基础

2.1 支持向量机回归模型

支持向量机 (Support Vector Machine，SVM) 是基于结构风险最小化的一类机器学习建模方法 [1]，其结构风险最小化原则使得所构建的模型拥有良好的泛化能力。当 SVM 应用于预测回归建模时，其中引入 ε 不敏感损失函数从而形成新的建模方法－支持向量回归 (SupportVector Ｒegression，SVR) 模型，SVR 模型是一种非参数回归模型，具有较好的非线性预测能力，其中回归超平面是通过优化与附近数据点 (支持向量) 的距离来确定的。通过考虑考虑预

【作者简介】唐安彬（1994-），男，中国四川成都人，硕士，从事工程造价信息化与应用研究。

测器相互作用的核函数，得到了非线性 SVR 公式。其多项式核函数表示为：

$$g_{i,j} = G(X_i, X_j) = (1 - X_i{}^T X_j)^q, q \in 2,3,\cdots$$

其中，G 是多项式核函数；X_i, X_j是模型的预测变量；q 是多项式函数的阶。通过最小化模型的特征系数 α 。

2.2 随机森林回归模型

随机森林 (Random Forest, RF) 是 CART 方法的扩展，由大量的决策树组成。它既可用于回归问题，也可用于分类问题。在回归的情况下，每颗决策树都是一个回归器，在训练阶段，随机森林使用 bootstrap 采样从输入训练数据集中采集多个不同的子训练集依次来训练不同的二叉决策树，即每输入一个样本，n 棵树就会有 n 个回归结果。这样的做法增加了最终模型预测结果的鲁棒性和稳定性；在预测阶段，随机森林将内部 n 棵决策树的预测结果取平均得到最终的输出，实现了从弱到强的过程。随机森林模型可以表示为：

$$y(x) = \frac{1}{k}\sum_{n=1}^{k} y_n(x)$$

其中，$y_n(x)$表示第 n 棵回归树的结果。随机森林中抽样的随机性与特征子集选取的随机性，使得该模型不容易陷入过拟合，并且具有很好的抗噪能力，在当前的很多数据集上，相对其他有些算法展现出了很大的优势，表现良好。

2.3 套索回归模型

套索回归算法 (Lasso Regression，LR) 是在最小二乘法的基础上，通过在拟合模型中加入 *L*1 范式作为正则化项，对模型的特征系数进行约束，使得模型训练过程的计算量和模型的复杂度得到简化 [2]。对于一般的线性回归模型：

$$y(x) = \sum_{i=1}^{n}(a_i x_i) + b$$

其中， $y(x)$ 为输出变量；x_i是模型的输入值；a_i是模型的回归系数；b 为截距。而套索回归算法的损失函数可以表示为：

$$L = (y(x) - Y)^T(Y - y(x)) + \alpha\|a_i\|_1$$

其中，Y 为实测值，$\alpha\|a_i\|_1$ 为 *L*1 正则化的收缩惩罚。为了确定期望的预测结果，需要通过调整 α 值使 L 值最小化。当 α =0 时，套索回归不在收缩惩罚，拟合模型与最小二乘法相同。然而，当 α 逐渐增大时，收缩惩罚项的影响变得更大，一些不太重要的输入变量的估计系数 a_i 缩小到零，当 α 接近无穷时，有些甚至可能被排除在特征池之外。因此，套索回归除了开发具有显式表达式的预测模型外，如果特征之间存在多重共线性问题，套索回归还可以自动进行特征选择或变量消除，使得套索回归更有吸引力和优势。

2.4 BP 神经网络模型

BP 神经网络模型是一种基于误差反向传播算法的多层前向网络模型，可以模拟人脑神经元的信息传递方式，对复杂的信息变量进行非线性变换和回归处理，得到拟合度高的运行结果。就像大脑神经元之间的信息传递一样，输入变量在输入层被输入后，会根据初始设定的权值得到输入变量的线性组合。当不断修改权值使线性组合值超过阈值时，信息就会传递到输出层。

BP 神经网络的结构包括输入层、隐含层和输出层，如图 1 所示。大量测试表明 [3]，三层 BP 神经网络非常适合对

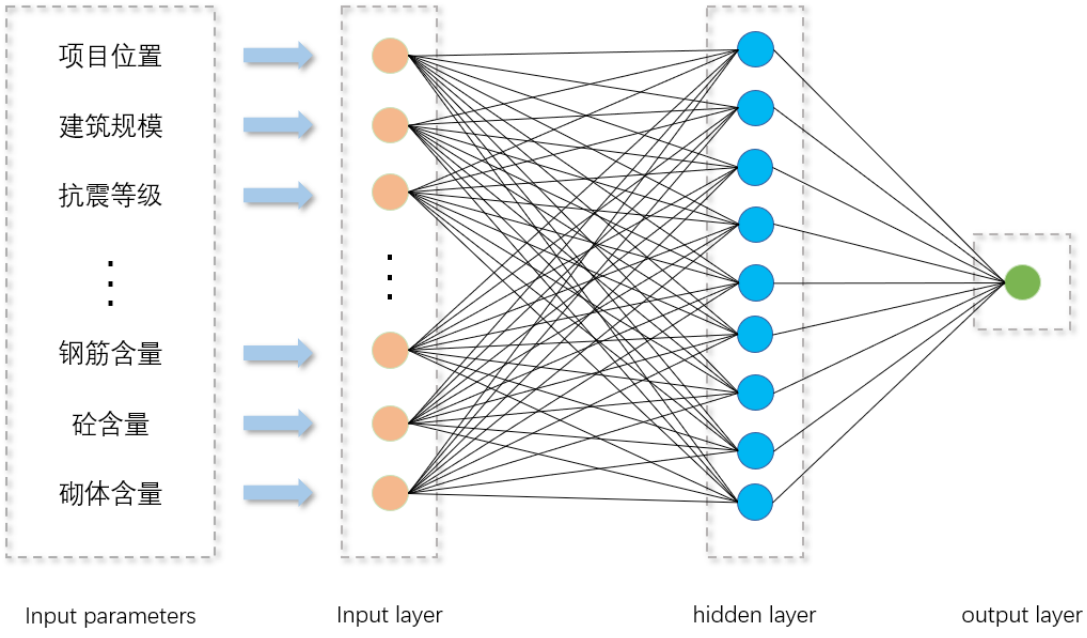

图 1 BP 神经网络结构

预测模型进行回归分析，因此，论文将建立三层 BP 神经网络进行项目工程造价预测分析。

3 数据预处理与特征提取

基于机器学习的房建项目造价预测模型需要一个由多个项目造价数据组成的数据集，本文通过对四川省成都市的 43 个已完工的住宅项目工程造价清单数据进行采集，按照国标清单的层级划分对项目的造价指标数据进行梳理，再结合实际情况以及行业内专家的相关意见，筛选出如钢筋单方含量、混凝土单方含量、模板单方含量、砌体单方含量以及外立面系数等相关特征。此外，房建项目的抗震等级、结构形式、项目位置、项目规模、建筑高度以及建筑面积等都会影响项目的工程造价，再综合考虑后，本文最终确定 15 个特征作为房建项目工程造价预测模型的输入特征，如表 1 所示。

表 1 特征指标

指标编号	指标名称	单位	指标类别
X_1	项目位置	无	输入类 - 定性指标
X_2	建筑规模	㎡	输入类 - 定量指标
X_3	抗震等级	无	输入类 - 定性指标
X_4	结构形式	无	输入类 - 定性指标
X_5	户型类型	无	输入类 - 定性指标
X_6	建筑高度	m	输入类 - 定量指标
X_7	标准层层高	m	输入类 - 定量指标
X_8	钢筋含量指标	kg/ ㎡	输入类 - 定量指标
X_9	砼含量指标	m³/ ㎡	输入类 - 定量指标
X_{10}	装配式构件指标	m³/ ㎡	输入类 - 定量指标
X_{11}	模板含量指标	㎡ / ㎡	输入类 - 定量指标
X_{12}	砌体含量指标	m³/ ㎡	输入类 - 定量指标
X_{13}	轻质隔墙含量指标	㎡ / ㎡	输入类 - 定量指标
X_{14}	内墙抹灰含量指标	㎡ / ㎡	输入类 - 定量指标
X_{15}	外墙抹灰含量指标	㎡ / ㎡	输入类 - 定量指标
Y	项目土建单方造价	元 / ㎡	输出类 - 定量指标

由于收集的 43 个项目造价数据存在部分项目的数据不完整或因为施工管理等原因造成成本过高，为防止模型训练后存在较大误差，需要对不完整的数据以及数据集中的异常值剔除。本文通过箱型图和散点图对项目的部分异常指标数据进行分析，对数据中存在离群点以及异常值的项目进行剔除，最终得到 35 个完整的项目造价数据。

4 预测与对比分析

4.1 模型建立

本文的数据集中输入特征有数值型和文本型两种类型，其中文本型数据需要转化为数字型，如户型类型有 T2、T4 以及 T6 等，分别用 0，1，2 进行表示。此外，由于不同特征之间的数据级别相差较大，对类特征的数据归一化和标准化处理，消除量纲对模型的影响。

利用 scikit-learn 和 keras 分别搭建支持向量机回归、随机森林回归、套索回归和神经网络模型。为防止模型的过拟合，将数据集划分为训练集和验证集，其中验证集由不同的结构形式抽取的 5 条数据组成。不同的模型在搭建时，需要设置许多的参数，如随机森林回归模型需要设置决策树的个数、评估准则、特征数等，支持向量机回归模型需要设置内核类型、惩罚参数和 epsilon 值等，神经网络需要设置激活函数、学习率、隐藏层神经元个数等。面对这些繁多的参数设置，本文采用网格搜索 - 交叉验证法寻找模型的最优参数，最终训练后的最优模型性能如图 2 所示。

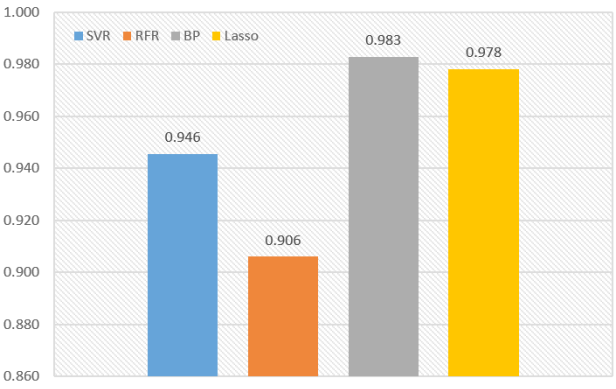

图 2 机器学习模型性能

4.2 模型分析评价

本研究为验证不同预测模型的拟合效果，将采用五项指标，在验证集上对模型的性能进行评估。

①均方误差（MSE）

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \quad (1)$$

②均方根误差（RMSE）

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \quad (2)$$

③平均绝对误差（MAE）

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \quad (3)$$

④平均绝对百分比误差（MAPE）

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{\hat{y}_i}\right| \quad (4)$$

⑤对称平均绝对百分比误差（SMAPE）

$$SMAPE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|\hat{y}_i - y_i|}{(|\hat{y}_i| + |y_i|)/2} \quad (5)$$

式（1）~（5）中：$\hat{y}_i$为实际值；y_i为预测值。

表 2 机器学习模型性能评估

回归模型	MSE	RMSE	MAE	MAPE	SMAPE
SVR	709.518	26.637	2.652	1.22%	1.22%
RFR	5668.622	75.290	38.644	2.97%	3.03%
BP	1086.953	32.969	24.418	1.26%	1.27%
Lasso	335.398	18.314	11.184	0.63%	0.64%

由表 2 可知，四个模型的平均绝对百分比误差（MAPE）

和对称平均绝对百分比误差（SMAPE）的值均小于 5%，满足工程造价估算的精度要求。

BP 神经网络模型在训练集上的 R^2=0.983，高于其余三个模型，但在验证集上的性能表现却差于支持向量机回归模型和套索回归模型，可见 BP 神经网络模型对训练集的数据有着较高的拟合度，但鲁棒性低于支持向量机回归模型和套索回归模型。

套索回归模型在训练集上的 R^2=0.978，与 BP 神经网络模型的指标仅差 0.005，但在验证集上的 MSE、RMSE、MAPE 和 SMAPE 四项指标上的表现均优于其他三个模型，可见套索回归模型更适合用于住宅项目工程造价预测。

5 结语

建筑项目的工程造价估概算对项目的成本管控存在决定性的影响，提高项目工程造价估概算的准确度能够有效的控制施工图设计和施工成本的管控。随着信息技术的发展和建筑行业的数字化进程加快，行业内将会存储越来越多的工程项目的数据，这些数据将会是构建工程造价数字化的基础。通过机器学习算法可以对多为异构数据之间的关联进行有效的挖掘，发现工程造价与建筑项目数据之间的隐藏关系，准确高效的完成项目工程造价的估概算工作，提高企业对项目的造价管控水平。本文通过对采集的住宅项目数据信息进行处理，并建立机器学习模型，并将训练好的模型对验证集的数据进行预测，得到的预测造价与实际造价之间的误差小于 5%，满足项目估概算的误差要求，说明机器学习在住宅项目工程造价预测的可行性，为未来工程工程造价工作有一定的指导意义。

参考文献

[1] Mangasarian O L, Musicant D R. Robust linear and support vector regression[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2002,22(9):950-955.

[2] Townsend W. Elasticregress: Stata module to perform elastic net regression, lasso regression, ridge regression[J]. Statistical Software Components, 2018(3):23-40.

[3] Sun W, Xu Y. Financial security evaluation of the electric power industry in China based on a back propagation neural network optimized by genetic algorithm[J]. Energy,2016,101(15):366-379.