doi:10.11835/j.issn.1000-582X.2020.058

基于样本集质量的建筑能耗预测机器学习算法选择及参数设置

刘 刚[a,b],李晓倩[a,b],韩 臻[a,b]
(天津大学 a.建筑学院;b.天津市建筑物理环境与生态技术重点实验室,天津 300072)

摘要:使用机器学习算法对建筑能耗进行预测正逐渐成为建筑设计初期重要的决策辅助工具,机器学习算法的选择及其参数设置一直是机器学习领域研究的热点和难点。但现有研究大多从算法原理角度进行预测模型的选择及参数设置,训练样本集的特征信息未得到充分利用。为此,提出一种以样本量及样本分布特征为出发点的样本集质量分类方法,针对不同质量样本集测试不同机器学习算法的学习性能,制定不同质量样本集的算法选择及参数设置策略。分析样本特征与算法性能之间的关系,为建筑设计提供有效指导。

关键词:建筑能耗预测;机器学习算法;样本分布特征
中图分类号:TU17 **文献标志码**:A **文章编号**:1000-582X(2022)05-079-17

Selection of building energy consumption prediction machine learning algorithms and parameter setting based on quality of samples

LIU Gang[a,b], *LI Xiaoqian*[a,b], *HAN Zhen*[a,b]
(a. School of Architecture; b. Tianjin Key Laboratory of Architectural Physics and Environmental Technology, Tianjin University, Tianjin 300072, P. R. China)

Abstract: Machine learning algorithms are playing a more important role in building energy consumption prediction during the conceptual design. The selection of the machine learning algorithms and parameter setting have become a focus in the field of building performance design. However, the algorithms and their parameters are usually determined by the principle of algorithms rather than the features of the training samples which also have an effect on the performance of algorithms. Therefore, a classification method based on the quality of training samples which is evaluated by sample size and sample distribution characteristics is proposed. The performance of different machine learning algorithms for different quality sample sets is tested, and algorithm selection and parameter setting strategies for different quality sample sets are formulated. The relationship between sample quality and algorithm performance is investigated to provide effective guidance for architects.

Keywords: building energy consumption prediction; machine learning algorithm; characteristic of the sample distribution

收稿日期:2020-03-11 **网络出版日期**:2020-06-19
基金项目:国家重点研发计划(2016YFC0700200);国家自然科学基金(51628803)。
Supported by National Key R&D Program of China (2016YFC0700200) and the National Natural Science Foundation of China(51628803).
作者简介:刘刚(1977—),男,博士生导师,主要从事绿色建筑与性能优化设计,(E-mail)lglgmike@163.com。

可持续建筑节能效果很大程度上取决于建筑初期设计[1]。近年来,结合建筑能耗预测方法和优化算法在建筑设计初期辅助节能优化决策成为研究热点[2]。优化过程中通常会生成大量的备选方案,能否快速进行建筑能耗预测成为影响优化效率的关键因素。随着人工智能技术的不断发展,基于机器学习的能耗预测方法越来越多的应用于建筑能耗优化中。实践证明,通过机器学习预测建筑能耗大大提高了建筑节能优化设计的效率,正逐步为建筑师所接纳。在实际应用中,建筑节能优化问题多为在已知可行空间内寻找最优方案[3],但在优化过程中,个体方案多为随机生成或有引导性的随机生成[4-6],使得用于学习的训练样本集分布情况未知。现有研究中,关于样本集质量(即样本分布不均衡问题)的研究多集中于分类问题中[7-9],在关于回归问题中机器学习算法的选择及其参数设置的研究中,关于样本集质量尚无统一定义[10],多集中于样本集大小对学习效果的影响或样本个体质量对学习效果的影响,较少关注样本集样本分布情况即样本集所包含信息完整性对学习效果的影响[11-14]。但在样本分布情况未知的前提下,随意选择的算法或不合理的参数设置可能会导致算法性能不理想,从而影响建筑节能优化效果。同时,对于建筑师而言,由于机器学习原理及应用的复杂性,尚未有较统一且明确的学习方法选择及参数设置依据对其进行指导。

文中提出了基于样本量及样本覆盖性的样本集质量评价方法,通过比较几种常用的机器学习方法及参数设置在不同质量样本集情况下的学习效果,分析样本集质量与机器学习算法性能之间的关系,针对不同质量样本集提出学习方法选择及参数设置建议,为建筑师使用提供理论指导。

1 理论与方法

1.1 传统机器学习算法

支持向量回归(Support Vector Regression,SVR)[15]是支持向量机的重要分支,广泛应用于非线性回归问题[16]。该算法基于核函数的小样本统计理论,其核心是VC维理论及结构风险最小化原则,可以有效避免陷入局部最优而达到全局最优,并通过核函数将低维空间问题映射至高维空间,将其转化为线性回归关系[17]。SVR算法具有结构简单、稳定性强、泛化能力强的优点,可以有效解决模型选择与欠学习、过学习、小样本、非线性和局部最优等问题,是建筑能耗预测中常用算法[18-21]。

BP神经网络(Back-Propagation Network,BP)是一种典型的多层前向型神经网络,利用误差反向传播算法对网络进行训练,理论上通过选择适当的网络层次及神经元个数可以任意逼近非线性函数[22]。该方法具有一定的自适应与自组织能力以及非线性映射能力,在建筑能耗预测问题中显示出明显优势[16]。但性能受样本数据及神经网络拓扑结构影响较大,且随着样本量的增多训练时间会大大加长,因此,选择适当的拓扑结构对该算法尤其重要。

1.2 集成机器学习算法

集成学习(Ensemble Learning)是机器学习领域重要的研究方向之一,通过多个学习算法对同一个问题进行学习,得到多个具有差异性的学习器,并通过一定组合方法对其学习结果进行组合得到最终结果,核心思想是充分利用误差较大的个体学习器所获得的局部信息来增强集成学习器的整体准确度和可靠性,而不是直接将其舍弃。集成学习具有准确度高,稳定性高,对参数设置敏感性相对较小以及学习效率高等优点,在建筑能耗预测中应用日趋广泛[23-25]。其中,应用最多且范围最广的为Bagging算法与Boosting算法。

Bagging算法通过自主采样法(Bootstrap)产生新的训练子集训练基学习器,结合策略组合各基学习器预测结果进行输出,基学习算法对训练数据越敏感,基学习器差异性越大,集成效果越好。

算法1 Bagging算法

输入:训练集 D,个体学习器 L,迭代次数 T;

for $t=1, 2, 3, \cdots, T$:

1）对样本集进行自主采样得到训练子集 D_t；

2）使用训练子集训练得到个体学习器 h_t；

end

输出：$H(x)=\arg\max\sum_{t=1}^{T}l(h_t(x)=y)$

Boosting 算法的基本思想是将多个预测精度较低的弱学习器提升至精度较高的强学习器。其中，最具代表性的为 AdaBoost 算法，核心思想是通过将自身的学习结果反馈到问题空间来进行交互，根据自身对环境的拟合程度来改变样本的采样概率[26]，从而加强对精度较低个体的学习。

算法 2　AdaBoost 算法

输入：训练集 D，个体学习器 L，迭代次数 T；

1）样本权重初始化为 $\omega_i=1/N$，$i=1,2,\cdots,N$，其中 N 为样本总数；

2）通过迭代获得强学习器：

for $t=1,2,\cdots,T$。

①在训练集上根据权重 ω_i 进行学习获得弱学习器 h_t；

②计算当前弱学习器中每个样本的相对误差并根据误差更新权重。

相对误差：$e_{ti}=\dfrac{(y_i-h_t(x_i))^2}{(\max|y_i-h_t(x_i)|)^2}$

回归误差率：$\varepsilon_t=\sum_{i=1}^{N}\omega_{ti}e_{ti}$

弱学习器 h_t 的权重系数：$\alpha_t=\dfrac{\varepsilon_t}{1-\varepsilon_t}$

更新样本权重系数：$\omega_{t+1,i}=\dfrac{\omega_{ti}\alpha 1-e_{tit}}{\sum_{i=1}^{N}\omega_{ti}\alpha 1-e_{tit}}$ end

输出：$H(x)=\arg\max\sum_{t=1}^{T}l(h_t(x)=y)$

Bagging 及 AdaBoost 算法均为使用较广泛的集成学习算法，Bagging 主要通过减小方差来提高学习性能，而 AdaBoost 在减小方差的同时还可以减小偏差，但 Bagging 对方差的减小程度大于 AdaBoost。且 Bagging 与 AdaBoost 相比稳定性和鲁棒性更强，但 AdaBoost 在降低错误率的程度上强于 Bagging[27]。

2　样本集质量划分方法

在使用机器学习算法时，样本集质量对绝大多数机器学习算法的学习效果影响较大，学习算法选择及其参数设置一直是机器学习研究中的热点问题，目前尚未有准确的结论可供参考，多通过参数寻优或经验验证进行设置，存在较大的主观性和局限性。在回归问题中，较少考虑样本数据分布特征，未充分利用隐含在数据集中的信息[28]。在建筑节能优化实践中，其数据集存在以下特点：1）解集空间已知，属于已知范围内的寻优问题；2）训练集为无噪声仿真数据，但训练集通过性能模拟得到，耗时较长；3）样本在解集空间中分布情况未知，可能会出现样本聚集，影响学习效果。

基于以上特征，文中提出一种基于样本量及样本覆盖性的样本集质量评价方法，以此为基础，测试不同样本集质量下机器学习算法的学习效果。首先，根据“3σ”准则，将样本集样本量分为小、中、大 3 个等级。其次，引入优化算法中解集质量评价指标——覆盖性（Coverage）评价样本在可行空间内的分布情况，如图 1 所示。

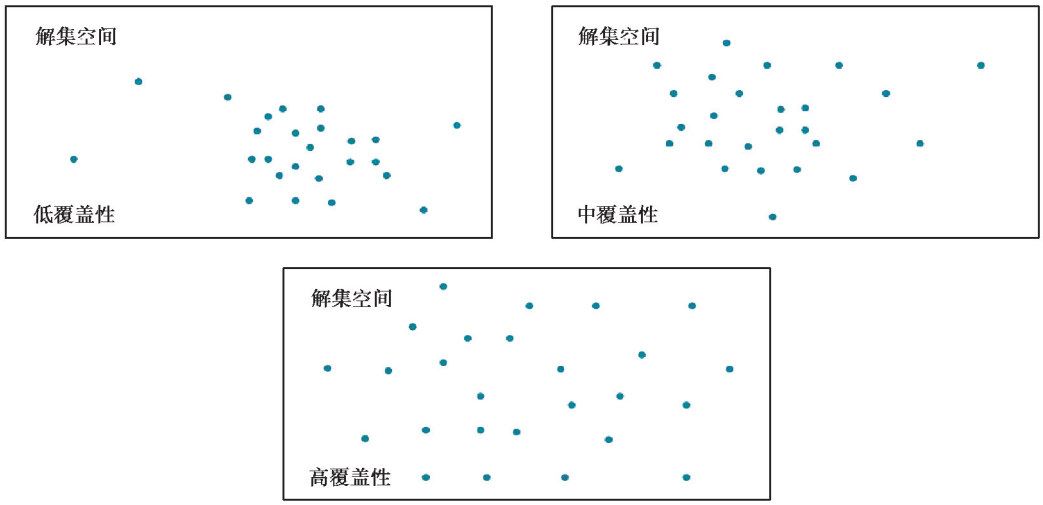


图1 样本集样本覆盖性示意图

Fig. 1 The coverage of sample set

覆盖性常用于优化算法中评价解集在解集空间中分布广泛性的指标，反映了样本点在可行求解空间中的分布情况，以表现在解集空间内的搜索程度，用以衡量是否陷入局部最优。其计算方法为

$$\mathrm{COV}=\prod_{k=1}^{m} SD_k, \tag{1}$$

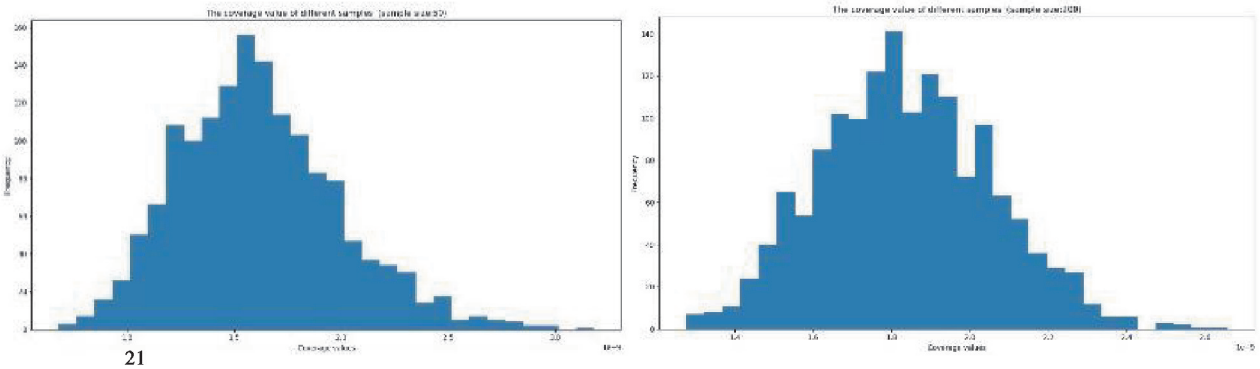
其中，

$$\mathrm{SD}_k=\sqrt{\frac{\sum_{j=1}^{t}(M_k-h_{kj})^2}{t}},$$

$$M_k=\frac{\sum_{j=1}^{t}h_{kj}}{t},$$

式中，COV为覆盖性；SD_k为第k个变量的标准差($k=1,2,\cdots,m$)，m为变量个数；h_{kj}为第j个个体第k个变量的值($j=1,2,\cdots,t$)，t为个体数量；M_k为第k个变量的平均值。由式(1)可知，覆盖性由各样本点各变量方差乘积求得，反应样本中各变量在空间中的不均衡性，即空间覆盖程度。在样本量相同的情况下，样本的覆盖性越高，说明样本在可行空间内的分布越均匀，样本集在各变量维度上的信息完整度越高，越有利于算法进行学习。

文中通过对样本量分别为50、200、500的样本集(均为随机生成)进行重复测试并计算其覆盖性。结果表明，样本集的覆盖性大致遵循正态分布，如图2所示，故样本集覆盖性等级划分采用“3σ”准则。



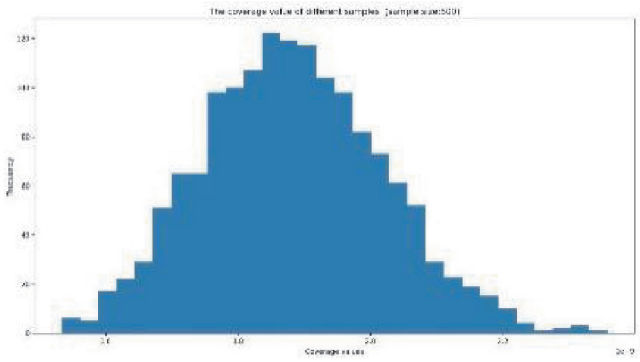


图 2　样本量为 50、200、500 的样本集覆盖性分布情况

Fig. 2　Coverage Probability Histogram for sample sets in different sizes (sample size: 50, 200, 500)

3　实验设置

设计 2 组实验测试不同样本集质量对其学习算法性能的影响。首先，测试传统机器学习算法在不同质量样本集下的表现，从中选出对于每类样本集表现较好的学习方法及参数设置，将其作为第 2 组实验的基学习器；其次，以第一阶段的实验结果为基础，测试集成学习算法对不同质量样本集的预测效果。最后，得出较好预测效果所需样本量，以及对应的机器学习方法及其参数设置，为建筑节能优化设计提供帮助。

3.1　实验环境及样本集设置

实验的运行环境为：Interi7 8 核 2.81 GHz 处理器，8G RAM 内存，64 位 Windows 10 操作系统。实验样本集来自天津一虚拟办公建筑的全年平均能耗模拟数据。该建筑共包含 4 大功能分区，分别为办公区、多媒体会议区、餐饮区及中庭交通区。因研究重点在测试机器学习算法性能，在合理的范围内简化模型，如图 3 所示。各样本集中的所有样本均为可行空间内随机生成的个体样本，其能耗通过 Grasshopper 中能耗模拟插件 Honeybee 仿真模拟得出。变量及取值范围，如表 1 所示，能耗相关参数设置及运行时间设置等均依照相关办公建筑设计规范设定。

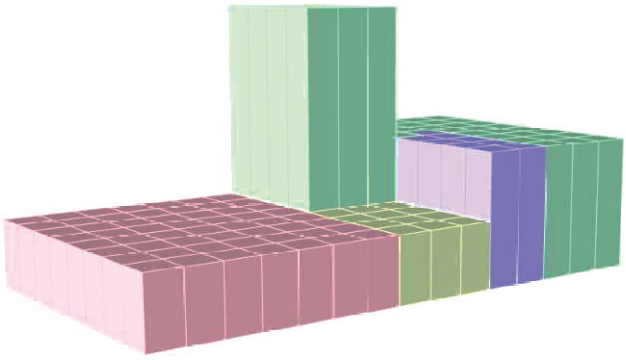


图 3　天津一虚拟办公建筑模型示意图

Fig. 3　The model of a virtual office building in Tianjin

表1 变量表

Table 1 Variables of case study

变量	类型	取值范围	变量含义
Orientation	连续	[0, π/2]	建筑朝向
WWR_n	连续	[0.1, 0.7]	北向窗墙比
WWR_w	连续	[0.1, 0.7]	西向窗墙比
WWR_e	连续	[0.1, 0.7]	东向窗墙比
WWR_s	连续	[0.2, 0.7]	南向窗墙比
Distance	连续	[0.01, 0.03]	百叶叶片宽度
w	连续	[0, 2.5]	西向百叶叶片距离
s	连续	[0, 2.222]	南向百叶叶片距离
U_value	连续	[0.8, 3]	玻璃传热系数
SHGC	连续	[0.15, 0.83]	玻璃太阳辐射得热系数
wall_R	连续	[2, 10]	外墙热阻值
roof_R	连续	[0.1, 0.45]	屋顶热阻值
floor_R	连续	[0.2, 0.45]	楼板热阻值

基于建筑节能优化实践中数据集的特征，综合考虑实际应用中的时间成本，将样本量分为50，200，500三个等级，分别代表小、中、大样本量。针对每类样本量各生成1 500个样本集并计算其覆盖性，依据“3σ”准则将其划分为低、中、高覆盖性。具体样本集分类及其特征如表2所示。

表2 样本集分类及特征

Table 2 Classification and characteristics of sample sets

样本集	样本量	覆盖性取值范围	变量数	数据来源
Sample50_low	50	[5.14E-10, 1.23E-09]	13	天津地区一虚拟办公建筑能耗模拟数据
Sample50_med	50	[1.23E-09, 2.00E-9]		
Sample50_high	50	[2.00E-09, 3.38E-09]		
Sample200_low	200	[1.22E-09, 1.62E-09]		
Sample200_med	200	[1.62 E-09, 2.05 E-09]		
Sample200_high	200	[2.05 E-09, 2.75 E-09]		
Sample500_low	500	[1.45 E-09, 1.74 E-09]		
Sample500_med	500	[1.74 E-09, 2.02 E-09]		
Sample500_high	500	[2.02 E-09, 2.32 E-09]		

3.2 实验设置

机器学习算法通过python Scikit-Learn中的SVR、MLPRegressor、BaggingRegressor及AdaBoostRegressor工具包实现。在训练学习器之前，为消除变量量级对学习性能的影响，对样本数据进行归一化处理：

$$x' = \frac{x - x_{\text{avg}}}{x_{\text{std}}} \quad , \tag{2}$$

其中，x'为归一化后的数据，x_{avg}，x_{std}分别为 x 的平均值和方差。

实验 1：传统机器学习算法性能评价。

选取 SVR 算法及 BP 神经网络算法进行训练，各类样本集中随机选择一个作为该类样本集代表。对于每一样本集，80％的样本作为训练集，剩余 20％作为测试集。对于 SVR 算法，主要超参数包括正则化参数 C，不敏感参数 ε 及核函数中的参数；对于 BP 神经网络算法，主要超参数包括隐藏层结构，激活函数以及学习率。考虑到计算时间成本，使用与待测试样本集同维度的 Scikit-Learn 自带标准数据集 Boston Housing 进行预实验，确定较优学习效果下的各参数大致范围，并选取对学习效果影响较大的超参数作为测试参数。最终选取 SVR 算法中高斯核函数的系数 γ 及 BP 神经网络中神经元结构的神经元个数作为测试超参数，其余超参数设置同样依据预实验中学习效果较优的模型参数。具体算法参数设置，如表 3 所示。

表 3　实验 1 算法超参数设置

Table 3　The hyperparameter setting of Experiment 1

算法	测试超参数设置	其余超参数设置
SVR	高斯核函数 $K(\vec{x},\vec{z})=\exp(-\gamma\|x-z\|^2)$ $\gamma=[0.01,0.02,0.03,0.04,0.05,0.06,0.07,0.08,0.09,0.1]$	核函数：高斯核函数 惩罚系数：$C=1.0$ ε 系数：$\varepsilon=0.1$
BP	神经元个数：[5，10，15，20，25，30，35，40，45，50]	隐藏层数：1 激活函数：tanh 正则化项系数：$\alpha=0.001$

实验 2：集成机器学习算法性能评价。

选取实验 1 中综合性能较好的 1 组 SVR 及 BP 设定参数，作为集成学习算法的基学习器，将 Bagging、AdaBoost 算法作为比较算法，主要分析基学习器及集成规模对集成效果的影响。训练集及测试集划分同实验 1。由于集成学习算法对于基学习器正确率的最低要求为 0.5，在集成过程中剔除正确率小于 0.5 的基学习器。算法参数设置如表 4 所示。

表 4　实验 2 算法超参数设置

Table 4　The hyperparameter setting of Experiment 2

算法	测试超参数设置	其余超参数设置
Bagging	集成规模：n_estimator = [10，15，20，25，30，40，50]	—
AdaBoost	集成规模：n_estimator = [10，15，20，25，30，40，50]	学习率：learning_rate=0.1

3.3　学习方法性能评价

算法性能评价包含拟合效果、有效率以及时间成本 3 方面。其中，拟合效果采用均方误差（mean squared error，MSE）及决定系数（Coefficient of determination，R^2）进行评价。在实验中，R^2 大于 0.9 视为优秀的学习算法，将有效率定义为用 R^2 大于 0.9 的概率，时间成本为算法运行一次的时间。

4　实验结果与分析

4.1　实验 1 结果与分析

所有算法训练 100 次取平均值作为最终结果进行比较，SVR 算法及 BP 算法对不同质量样本集的决定系数和均方误差如图 4～图 9 所示，运算时间如表 5～表 6 所示，有效性如表 7～表 8 所示。

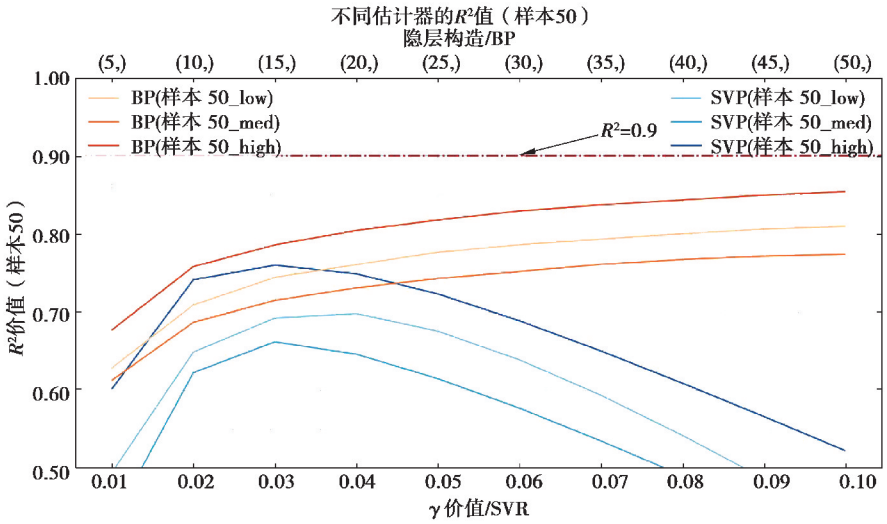


图 4 Sample50 决定系数

Fig. 4 The R^2 of Sample50

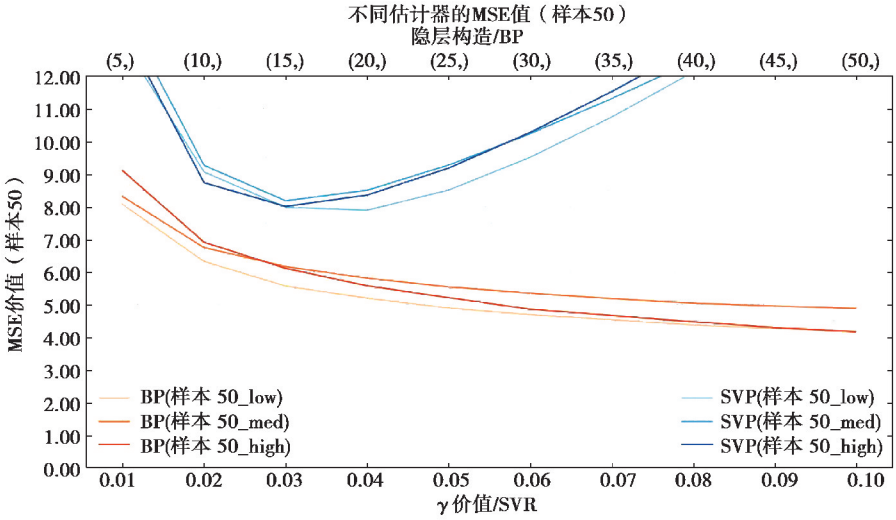


图 5 Sample50 均方误差

Fig. 5 The MSE of Sample50

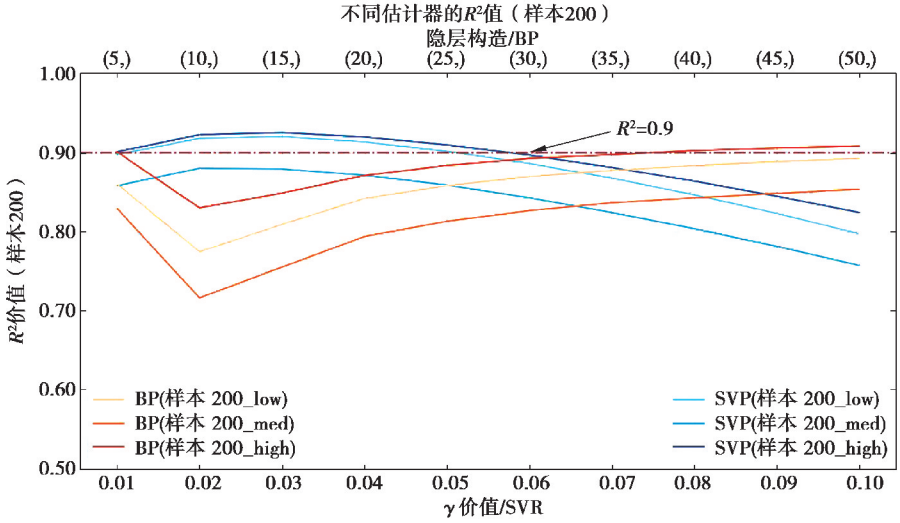


图 6 Sample200 决定系数

Fig. 6 The R^2 of Sample200

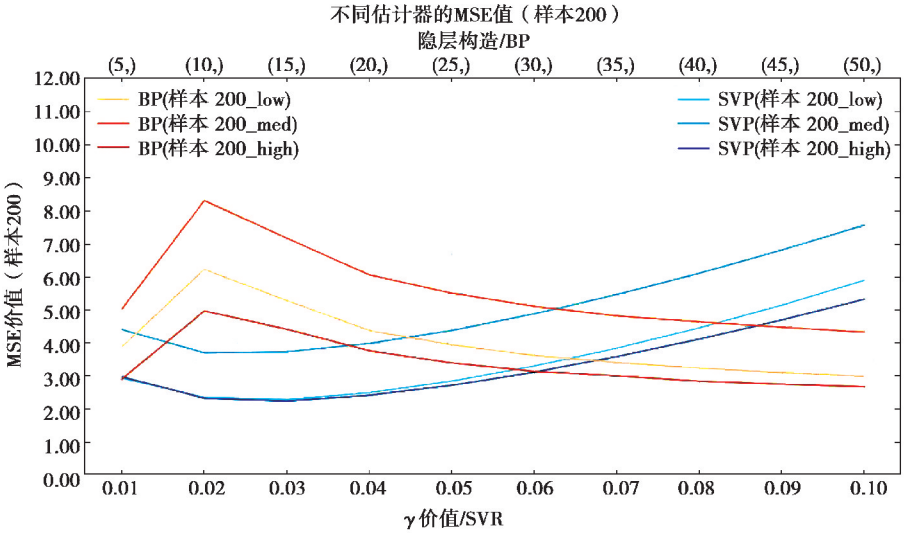


图 7　Sample200 均方误差

Fig. 7　The MSE of Sample200

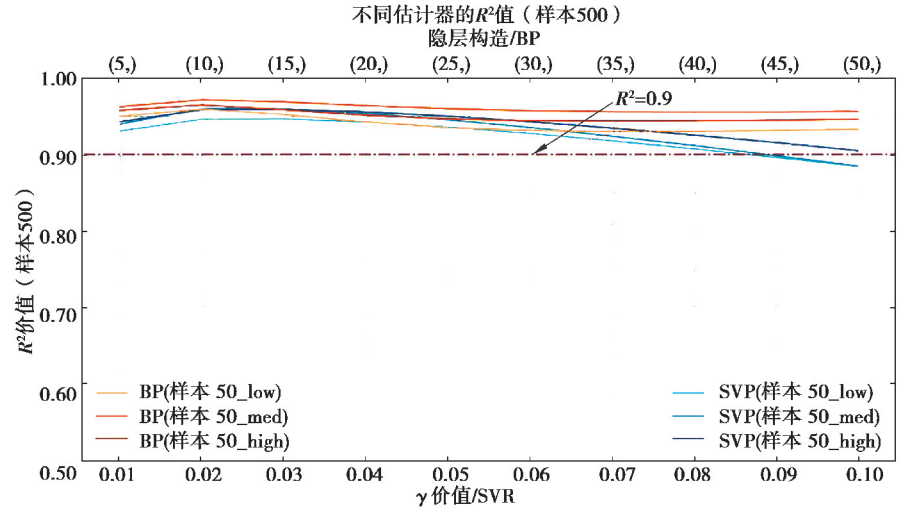


图 8　Sample500 决定系数

Fig. 8　The R^2 of Sample500

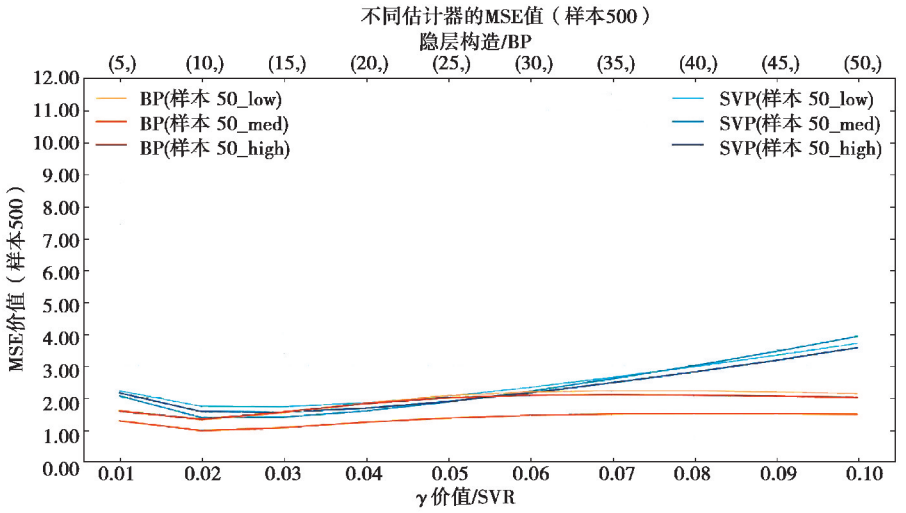


图 9　Sample500 均方误差

Fig. 9　The MSE of Sample500

表5 SVR算法计算时间

Table 5 The computation time of SVR **s**

样本集	γ									
	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
Sample50	0.001 0	0.001 0	0.001 0	0.001 0	0.001 0	0.001 0	0.001 0	0.001 0	0.001 0	0.001 0
Sample200	0.003 0	0.002 0	0.002 0	0.003 0	0.002 0	0.002 0	0.002 0	0.003 0	0.002 0	0.003 0
Sample500	0.008 0	0.009 0	0.010 0	0.008 1	0.015 6	0.015 6	0.015 6	0.015 6	0.015 6	0.015 6

表6 BP算法计算时间

Table 6 The computation time of BP **s**

样本集	隐藏层结构									
	(5,)	(10,)	(15,)	(20,)	(25)	(30,)	(35,)	(40,)	(45,)	(50,)
Sample50	0.020 1	0.015 6	0.015 6	0.015 6	0.015 6	0.015 6	0.015 6	0.015 6	0.015 6	0.031 2
Sample200	0.046 8	0.055 3	0.058 6	0.050 2	0.050 6	0.051 1	0.052 0	0.052 5	0.054 5	0.055 8
Sample500	0.060 1	0.075 1	0.088 1	0.103 7	0.091 8	0.078 1	0.082 5	0.088 1	0.089 8	0.112 4

表7 SVR算法有效率

Table 7 The available ratio of SVR

样本集	γ									
	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
Sample50_low	0	0	0	0	0	0	0	0	0	0
Sample50_med	0	0	0	0	0	0	0	0	0	0
Sample50_high	0	0.03	0.03	0.03	0.03	0.03	0	0	0	0
Sample200_low	0.50	0.85	0.86	0.77	0.51	0.35	0.18	0.05	0.01	0
Sample200_med	0.13	0.45	0.47	0.36	0.26	0.20	0.05	0.02	0.01	0
Sample200_high	0.57	0.87	0.87	0.80	0.69	0.50	0.33	0.29	0.10	0.03
Sample500_low	0.96	1.00	1.00	0.98	0.98	0.95	0.86	0.66	0.48	0.26
Sample500_med	1.00	1.00	1.00	1.00	1.00	1.00	0.92	0.79	0.53	0.21
Sample500_high	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.94	0.86	0.64

表8 BP算法有效率

Table 8 The available ratio of BP

样本集	隐藏层结构									
	(5,)	(10,)	(15,)	(20,)	(25)	(30,)	(35,)	(40,)	(45,)	(50,)
Sample50_low	0	0.01	0.01	0.06	0.08	0.10	0.12	0.17	0.17	0.19
Sample50_med	0	0	0	0	0.01	0.02	0.03	0.05	0.06	0.07
Sample50_high	0	0.04	0.10	0.15	0.21	0.24	0.29	0.31	0.32	0.33
Sample200_low	0.02	0	0	0.01	0.03	0.07	0.18	0.31	0.39	0.44
Sample200_med	0	0	0	0	0	0.01	0.01	0.02	0.04	0.06
Sample200_high	0.63	0	0.02	0.14	0.29	0.48	0.57	0.65	0.68	0.73
Sample500_low	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99	1.00	0.99
Sample500_med	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Sample500_high	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

通过对实验 1 结果进行分析，可得到如下结论：

1）拟合效果方面，Sample50 中所有覆盖性样本集拟合效果均较差，SVR 算法及 BP 神经网络算法均未达到 $R^2>0.9$ 的优秀标准，但 SVR 算法拟合效果普遍优于 BP 神经网络算法。Sample200 中，拟合效果明显提升，当神经元个数小于 35 时，SVR 算法表现优异，对于低、高覆盖性的样本集，在 γ 取 0.03 时取得最优效果并达到优秀标准，当神经元个数大于 35 时，BP 神经网络算法的拟合效果优于 SVR 算法，但计算时间较长。对于中覆盖性样本集，始终未达到优秀标准。Sample500 中，拟合效果极优，2 种算法的 r^2 均可达到 0.9 以上，BP 神经网络算法甚至可达 0.95；

2）样本量越大，计算成本越高，有效率越高，准确性越强，即训练样本中包含的可行空间内的信息越丰富，学习效果越好。同时，随着样本量的增加，拟合效果对学习算法及参数设置的敏感性下降，即各算法及参数设置之间的学习差异减小；

3）覆盖性对学习效果有一定影响，当样本量相同时，各样本集覆盖性虽然存在差异，但其学习效果的变化趋势基本相同。样本量不同时，覆盖性对学习效果的影响存异，在实验中，样本量为 50，200 时，其学习效果从优至劣依次为高、低、中覆盖性；当样本量为 500 时，SVR 算法的学习效果优劣排序为高、中、低，而 BP 神经网络算法学习效果优劣排序则为中、高、低。由此可见，覆盖性与学习器的学习效果并不始终成正相关关系，而是与样本量及学习算法有关。印证了 Zhou 等[29]在“选择性集成”概念中证明的通过选择部分个体学习器来构建集成可能要优于使用所有个体学习器构建的集成；

4）在实验中，SVR 算法在不同参数设置下算法复杂度无明显差别，BP 神经网络算法的复杂度随着隐藏层结构的复杂化而逐渐增加，结果表明，当隐藏层结构达到一定复杂度时，继续增加神经元个数，反而会降低学习效果，且神经元个数越多，时间成本越大，在选择算法及参数设定时，应选用适当复杂度算法，以防止出现过拟合现象；

5）对于不同样本量样本集，计算时间虽随着样本量增加而逐渐增大，并无数量级上的差别，测试模型较为简单，随着模型复杂度的增加，样本量带来的计算时间差异会逐渐增大。在建筑优化过程中，时间成本的增加主要来自于生成样本集时所需的模拟计算时间，故当样本量增大时，整体时间成本会大大增加。

4.2　实验 2 结果与分析

在实验 1 中，Sample50 中所有样本集均未达到优秀水平，Sample200 中覆盖性样本集未达到优秀水平，对以上样本集进行集成学习实验，以获得较好的拟合效果。基学习器综合考虑准确性、有效率及计算时间 3 方面，以实验 1 结果为参考，选取表现较好且模型复杂度适中的算法及参数设置作为基学习器。因支持向量机是一种比较稳定的学习算法，直接集成效果不佳，故基学习器均选取不同复杂度的神经网络算法。其中，Sample50 因样本量较少且 R^2 呈递增趋势，故选取 4 种隐藏层结构依次进行集成。具体学习器设置及集成学习参数如表 9 所示。

表 9　集成学习参数设置

Table 9　The parameter settings of ensemble learning

样本集	集成算法	基学习器	集成测试超参数设置	其余超参数设置
Sample50_low Sample50_med Sample50_high	Bagging	BP：[(20，)，(30，)，(40，)，(50，)]	n_estimator：[10，15，20，25，30，40，50]	—
	AdaBoost	BP：[(20，)，(30，)，(40，)，(50，)]	n_estimator：[10，15，20，25，30，40，50]	Learning_rate = 0.1
Sample200_med	Bagging	BP：[(20，)]	n_estimator：[10，15，20，25，30，40，50]	—
	AdaBoost	BP：[(20，)]	n_estimator：[10，15，20，25，30，40，50]	Learning_rate = 0.1

实验2算法R^2如图10～图13所示，计算时间如表10～表11所示，算法有效性如表12～表13所示。通过对实验2结果进行分析，可得到以下结论：

1)在拟合效果方面，对于Sample50的三类样本集，由于基学习器学习效果较差，经集成之后，绝大多数集成学习器仍未达到优秀标准，仅在高覆盖性样本集中，以隐藏层结构为(40，)的BP神经网络学习器作为基学习器时，R^2可达到0.9以上。但在各类覆盖性中，均有集成学习器R^2可达到0.85以上，达到回归学习器可使用的基本要求。在Sample200_med样本集中，当AdaBoost算法的集成规模达到40时，R^2达到0.9；

2)在基学习器方面，基学习器的拟合效果与最终集成后的拟合效果并不完全成正相关关系。因为基学习器的复杂度过高，导致其泛化能力较弱，在集成时生成的个体学习器差异度较小，从而影响其拟合效果；

3)在集成规模方面，当集成规模达到一定数值之后，继续增大集成规模并不会明显提升集成效果，甚至会减弱拟合效果(如图7中AdaBoost[(50，)])；

4)在计算时间方面，随着集成规模的扩大，时间成本逐渐增高。计算时间受样本量影响较大，当样本量增加时，时间成本明显提高。综合考虑，在设定集成规模时，应适中为宜。

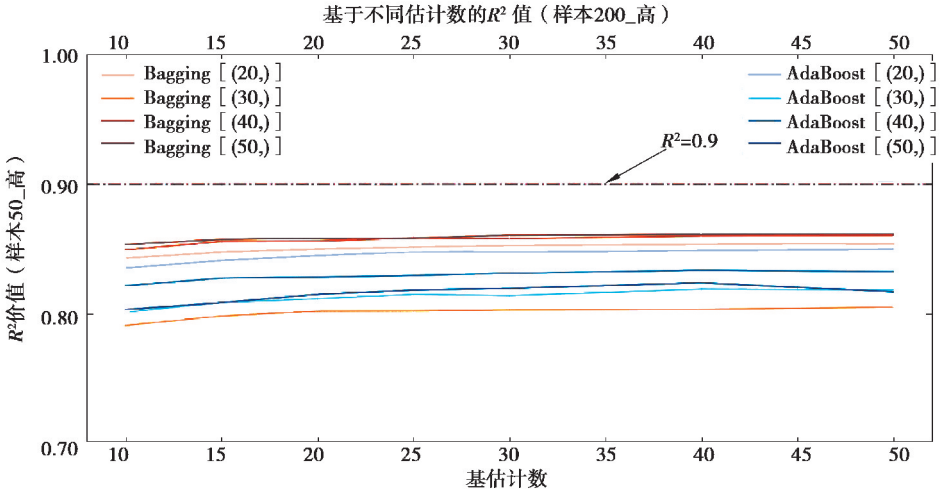


图10 Sample50_low集成学习决定系数

Fig. 10 The R^2 of Sample50_low

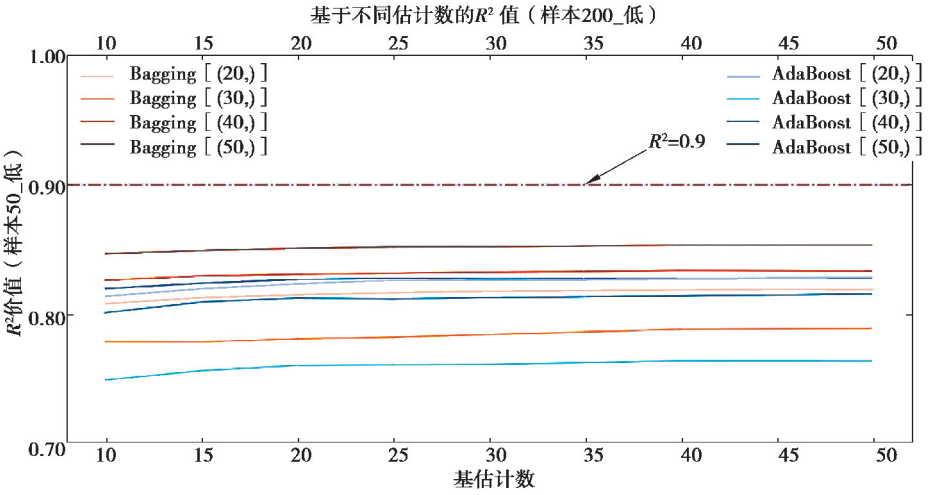


图11 Sample50_med集成学习决定系数

Fig. 11 The R^2 of Sample50_med

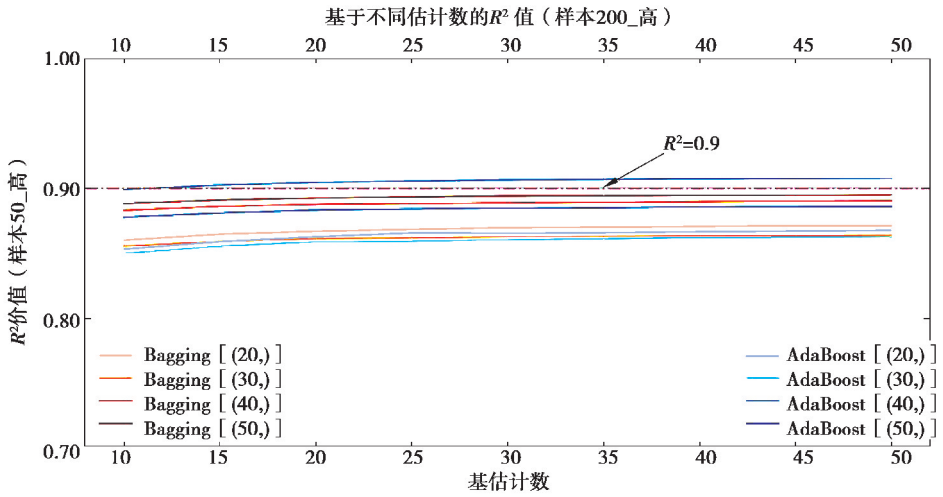


图 12　Sample50_high 集成算法决定系数

Fig. 12　The R^2 of Sample50_high

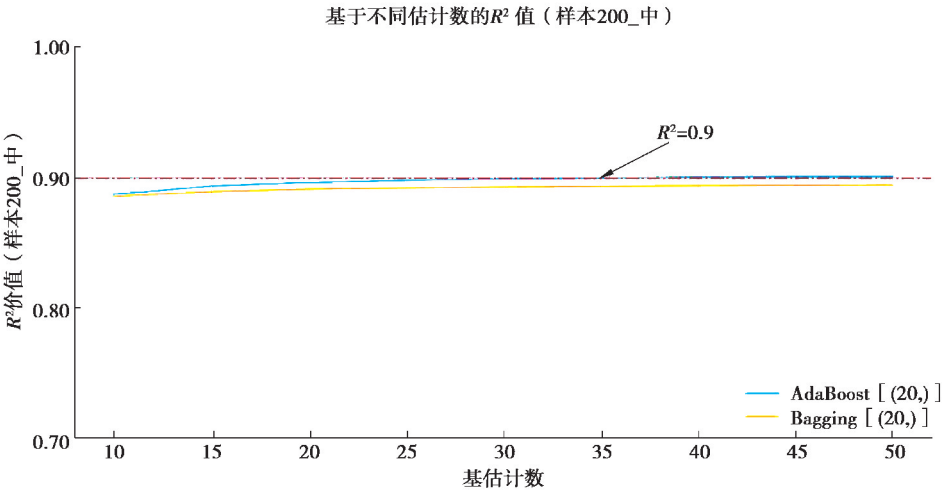


图 13　Sample200_med 集成算法决定系数

Fig. 13　The R^2 of Sample200_med

表 10　Sample50 集成算法计算时间

Table 10　The computation time of ensemble learning for Sample50

	算法	10	15	20	25	30	40	50
(20,)	Bagging	0.080 6	0.109 4	0.148 4	0.193 7	0.235 9	0.304 6	0.370 2
	AdaBoost	0.079 7	0.112 5	0.165 6	0.187 5	0.223 4	0.299 9	0.392 1
(30,)	Bagging	0.089 0	0.123 4	0.163 9	0.203 1	0.242 1	0.339 0	04 048
	AdaBoost	0.082 8	0.123 4	0.176 5	0.218 7	0.256 2	0.334 3	0.442 1
(40,)	Bagging	0.096 8	0.139 5	0.181 2	0.246 8	0.278 1	0.364 0	0.478 0
	AdaBoost	0.120 3	0.148 4	0.207 8	0.245 3	0.284 3	0.376 5	0.471 8
(50,)	Bagging	0.114 0	0.176 5	0.214 0	0.267 8	0.323 4	0.452 1	0.534 2
	AdaBoost	0.117 2	0.164 0	0.217 1	0.271 8	0.337 4	0.426 5	0.549 9

表11 Sample200集成算法计算时间

Table 11 The computation time of ensemble learning for Sample200

	算法	10	15	20	25	30	40	50
(20,)	Bagging	0.334 3	0.539 5	0.641 8	0.831 1	0.960 7	1.274 7	1.629 3
	AdaBoost	0.454 6	0.650 0	0.859 2	1.057 6	1.288 8	1.735 5	2.149 5

表12 Sample50集成算法结果有效率

Table 12 The available ratio of ensemble learning for Sample50

样本集		算法	10	15	20	25	30	40	50
Sample50_low	(20,)	Bagging	0.37	0.37	0.40	0.40	0.40	0.40	0.40
		AdaBoost	0.27	0.27	0.27	0.33	0.33	0.33	0.33
	(30,)	Bagging	0.33	0.33	0.37	0.37	0.37	0.37	0.37
		AdaBoost	0.23	0.30	0.30	0.33	0.33	0.33	0.33
	(40,)	Bagging	0.43	0.47	0.47	0.47	0.47	0.47	0.47
		AdaBoost	0.27	0.27	0.27	0.27	0.27	0.27	0.27
	(50,)	Bagging	0.37	0.40	0.40	0.40	0.40	0.40	0.40
		AdaBoost	0.37	0.37	0.37	0.37	0.37	0.37	0.37
Sample50_med	(20,)	Bagging	0.10	0.13	0.20	0.23	0.23	0.23	0.23
		AdaBoost	0.10	0.10	0.10	0.10	0.10	0.10	0.10
	(30,)	Bagging	0.17	0.17	0.20	0.20	0.20	0.20	0.20
		AdaBoost	0.03	0.07	0.07	0.07	0.10	0.10	0.10
	(40,)	Bagging	0.13	0.20	0.23	0.23	0.23	0.23	0.23
		AdaBoost	0.20	0.20	0.20	0.20	0.20	0.20	0.20
	(50,)	Bagging	0.20	0.30	0.33	0.33	0.33	0.33	0.33
		AdaBoost	0.17	0.20	0.20	0.20	0.17	0.20	0.17
Sample50_high	(20,)	Bagging	0.40	0.43	0.47	0.47	0.47	0.47	0.47
		AdaBoost	0.27	0.40	0.40	0.40	0.40	0.40	0.40
	(30,)	Bagging	0.27	0.27	0.30	0.30	0.33	0.33	0.33
		AdaBoost	0.20	0.23	0.27	0.23	0.27	0.27	0.27
	(40,)	Bagging	0.43	0.53	0.57	0.57	0.57	0.57	0.57
		AdaBoost	0.60	0.67	0.67	0.67	0.67	0.67	0.67
	(50,)	Bagging	0.53	0.53	0.53	0.53	0.53	0.57	0.57
		AdaBoost	0.43	0.50	0.50	0.50	0.50	0.50	0.50

表 13　Sample200_med 集成算法结果有效率

Table 13　The available ratio of ensemble learning for Sample200_med

样本集	算法	10	15	20	25	30	40	50
Sample200_med　(20，)	Bagging	0.37	0.40	0.40	0.40	0.40	0.40	0.40
	AdaBoost	0.37	0.43	0.47	0.50	0.57	0.57	0.57

综合实验结果，样本量为 50 时无法保证在大多数情况下达到较优学习效果，但 R^2 可以达到 0.85 以上，已达到可用标准，若建筑师无充足时间且对预测精度要求较低时，可使用样本量为 50 的样本集。样本量为 200 时，全部覆盖度可以保证 0.9 以上的 R^2，耗时在可接受范围内，为较理想的样本量。样本量为 500 时，仅使用传统机器学习算法就可以达到极好的学习效果，R^2 可达 0.95 以上，学习用时较短，但其生成样本集时间成本巨大，若建筑师有充足的时间且对预测精度有极高要求，可采用该样本集。针对不同质量样本集的学习方法及参数设置建议及其学习效果如表 14 所示。

表 14　不同质量样本集的学习方法选择与参数设置建议

Table 14　The suggestions on algorithm selection and parameter setting for different quality of sample sets

样本集	学习算法及参数设置	R^2	MSE	有效率	运算时间
Sample50_low	Bagging(基学习器：BP：(50，)，基学习器数量：40)	0.860 5	4.010 0	0.40	0.452 1
Sample50_med	Bagging (基学习器：BP：(50，)，基学习器数量：40)	0.852 6	5.397 2	0.33	0.452 1
Sample50_high	AdaBoost(基学习器：BP：(40，)，基学习器数量：40)	0.906 9	3.311 9	0.67	0.376 5
Sample200_low	SVR(核函数：tanh(γ=0.03))	0.920 5	2.283 9	0.86	0.002 0
Sample200_med	AdaBoost(基学习器：BP：(20，)，基学习器数量：40)	0.900 1	3.906 3	0.57	1.735 5
Sample200_high	SVR(核函数：tanh(γ=0.03))	0.925 6	2.235 2	0.87	0.002 0
Sample500_low	BP(隐藏层结构：(10，))	0.959 0	1.293 8	1.00	0.075 1
Sample500_med	BP(隐藏层结构：(10，))	0.971 2	0.958 1	1.00	0.075 1
Sample500_high	BP(隐藏层结构：(10，))	0.964 2	1.319 3	1.00	0.075 1

5　结　论

文中基于样本量及样本分布特征对样本集质量进行评价与分类，针对不同质量样本集构建了建筑能耗预测模型，分析样本量与样本分布特征对机器学习算法学习性能的影响，得到以下结论：

1)样本量及样本覆盖性对机器学习算法的学习性能有影响，其中，样本量的影响程度大于样本覆盖性。对于某一种机器学习算法，在相同样本量的情况下，不同覆盖性样本集的学习效果随参数变化的趋势相同。对于不同算法，在同样本量情况下，样本覆盖性对学习效果的影响有所不同。因此，样本覆盖性与算法的学习效果并不始终成正相关关系，而是与样本量及选择的学习算法有关。

2)当样本量越大时，学习效果对学习算法及参数设置的敏感性越低，各算法及参数设置之间的学习效果差异减小。虽样本量越大学习效果越佳，但时间成本亦随之增加(其主要增加量来自于生成样本所需的模拟

计算时间),针对本案例,当样本量为200时,无论覆盖性如何,均足以取得较优的学习效果。

3)集成学习对拟合效果的提升较为明显,当其集成规模达到一定程度后,继续扩大集成规模,时间成本增量较大,但其拟合效果提升较小。

在实际设计应用中,需根据特定问题选择适宜的算法进行求解。文中提出了针对各类质量样本集的适用算法及其参数设置,为建筑师实际使用提供了参考。在未来研究中,将继续研究样本集质量及其余超参数与学习效果的关系,建立自适应的机器学习算法集并将其集成至优化算法中,进一步提高建筑节能优化效率。

参考文献:

[1] 中国城市科学研究会. 中国绿色建筑2017[M]. 北京: 中国建筑工业出版社, 2017.
Chinese Society for Urban Studies. China green building2017[M]. Beijing: China Architecture & Building Press, 2017.(in Chinese)

[2] 李紫微, 林波荣, 陈洪钟. 建筑方案能耗快速预测方法研究综述[J]. 暖通空调, 2018,48(5):1-8.
Li Z W, Lin B R, Chen H Z. Review of rapid prediction method of building energy consumption[J]. Journal of HV&AC, 2018,48(5):1-8.(in Chinese)

[3] Machairas V, Tsangrassoulis A, Axarli K. Algorithms for optimization of building design: A review[J]. Renewable and Sustainable Energy Reviews, 2014,31:101-112.

[4] Wang D, Tan D, Liu L. Particle swarm optimization algorithm: an overview[J]. Soft Computing, 2018,22(2):387-408.

[5] Yang C F, Li H J, Rezgui Y, et al. High throughput computing based distributed genetic algorithm for building energy consumption optimization[J]. Energy and Buildings, 2014, 76: 92-101.

[6] Tuhus-dubrow D, Krarti M. Genetic-algorithm based approach to optimize building envelope design for residential buildings[J]. Building and Environment, 2010,45(7):1574-1581.

[7] Carvalho A M D, Prati r C. Improving kNN classification under unbalanced data. A New Geometric Oversampling Approach [J]. 2018, 1-6.

[8] Sadrawi M, Sun W Z, Ma h M, et al. Ensemble genetic fuzzy neuro model applied for the emergency medical service via unbalanced data evaluation [J]. Symmetry, 2018, 10(3): 71.

[9] 陈圣灵, 沈思淇, 李东升. 基于样本权重更新的不平衡数据集成学习方法 [J]. 计算机科学, 2018, 45(7): 31-37.
Chen S L, Shen S Q, Li D S. Ensemble learning method for imbalanced data based on sample weight updating [J]. Computer Science, 2018, 45(7): 31-37.(in Chinese)

[10] 蒋林, 陈涛, 屈梁生. 训练样本质量对人工神经网络性能的影响 [J]. 中国机械工程, 1997, (2): 123.
Jiang L, Chen T, Qu L S. The effect of training sample quality on the performance of artificial neural networks [J].China Mechanical Engineering, 1997, (2): 123.(in Chinese)

[11] Tharwat A, Hassanien A E, Elnaghi B E. A BA-based algorithm for parameter optimization of Support Vector Machine[J]. Pattern Recognition Letters, 2017,93:13-22.

[12] Wu C, Tzeng G, Lin R. A Novel hybrid genetic algorithm for kernel function and parameter optimization in support vector regression[J]. Expert Systems with Applications, 2009,36(3):4725-4735.

[13] Ding S, Su C, Yu J. An optimizing BP neural network algorithm based on genetic algorithm[J]. Artificial Intelligence Review, 2011,36(2):153-162.

[14] Ren C, An N, Wang J, et al. Optimal parameters selection for BP neural network based on particle swarm optimization: A case study of wind speed forecasting[J]. Knowledge-Based Systems, 2014,56:226-239.

[15] Vapnik V N . Support vector method for function approximation, regression estimation, and signal processing[J]. Advanced Neural Information Processing System, 1996,9(6):281-287.

[16] Zhao H, Magoulès F. A review on the prediction of building energy consumption[J]. Renewable and Sustainable Energy Reviews, 2012,16(6):3586-3592.

[17] Gu B, Sheng V S, Wang Z, et al. Incremental learning for ν-Support vector regression[J]. Neural Networks, 2015,67: 140-150.

[18] Ahmad A S, Hassan m Y, Abdullah M P, et al. A review on applications of ANN and SVM for building electrical energy

consumption forecasting[J]. Renewable and Sustainable Energy Reviews, 2014,33:102-109.

[19] Dong B, Cao C, Lee S E. Applying support vector machines to predict building energy consumption in tropical region[J]. Energy and Buildings, 2005,37(5):545-553.

[20] Chou J, Bui D. Modeling heating and cooling loads by artificial intelligence for energy-efficient building design[J]. Energy and Buildings, 2014,82:437-446.

[21] Zhong H, Wang J, Jia H, et al. Vector field-based support vector regression for building energy consumption prediction[J]. Applied Energy, 2019,242:403-414.

[22] 刘彩红. BP 神经网络学习算法的研究[D]. 重庆：重庆师范大学, 2008.
Liu C H. The study of algorithm of BP neural network[D]. Chongqing: Chongqing Normal University, 2008.(in Chinese)

[23] Chen K, Jiang J, Zheng F, et al. A novel data-driven approach for residential electricity consumption prediction based on ensemble learning[J]. Energy, 2018,150:49-60.

[24] Wang Z, Wang Y, Srinivasan R S. A novel ensemble learning approach to support building energy use prediction[J]. Energy and Buildings, 2018,159:109-122.

[25] Papadopoulos S A E W W. Evaluation of tree-based ensemble learning algorithms for building energy performance estimation[J]. Journal of Building Performance Simulation, 2017,11(3):322-332.

[26] 王永明. 集成回归问题若干关键技术研究[D]. 上海：华东师范大学, 2015.
Wang Y M. Research on some key technologies in ensemble regression problem[D]. Shanghai: East China Normal University, 2015.(in Chinese)

[27] Bauer E K R. An empirical comparison of voting classification algorithms: bagging, boosting, and variants[J]. Mach Learn, 1999,36(1-2):105-139.

[28] 梁礼明，冯新刚，陈云嫩，等. 基于样本分布特征的核函数选择方法研究[J]. 计算机仿真, 2013,30(1):323-328.
Lang L M, Feng X G, Chen Y N, et al., Method of selection kernel function based on distribution characteristics of samples[J]. Computer Simulation, 2013,30(1):323-328.(in Chinese)

[29] Zhou Z H W J X T. Ensembling neural networks: Many could be better than all[J]. Artif Intell, 2002,137(1/2):239-263.

（编辑　陈移峰）