文章编号:1674-2974(2019)02-0123-08 DOI:10.16339/j.cnki.hdxbzkb.2019.02.017

基于贝叶斯模型组合的随机森林预测方法

董娜†,常建芳,吴爱国
(天津大学 电气自动化与信息工程学院,天津 300072)

摘 要:为了能够精准可靠地估计太阳能辐照度,本文提出一种基于贝叶斯模型组合的随机森林算法用于太阳能辐照度预测.首先,引入K-means聚类和K折交叉验证将气象数据训练集生成多个训练子集,以增加训练子集的多样性并保证均匀采样.其次,将随机森林作为基学习器建立集成学习预测模型,导入训练子集并训练各个随机森林.之后,依据各个随机森林在验证集上的预测性能,采用贝叶斯模型组合算法制定组合策略.个体随机森林在测试集上的预测值经过模型组合策略得到最终输出.最后,基于气象实测数据建立仿真实验,并引入其他四种预测方法进行对比仿真研究,通过实验结果验证了文中所提出预测方法在太阳能辐照度预测问题中的准确性和可靠性.

关键词:K均值聚类;交叉验证;随机森林;贝叶斯模型组合;太阳能辐照度
中图分类号:TP181 **文献标志码**:A

Random Forest Prediction Method Based on Bayesian Model Combination

DONG Na†, CHANG Jianfang, WU Aiguo
(School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China)

Abstract: To accurately and reliably estimate the solar irradiance, a random forest algorithm was proposed based on the Bayesian model combination for solar irradiance prediction. Firstly, the K-means clustering and K-fold cross validation were introduced to generate multiple training subsets so as to increase the diversity of training subsets and to ensure uniform sampling. Secondly, the random forests were defined as base learners to establish an ensemble learning prediction model, with each training subset being used to train the corresponding individual random forest. Then, according to the prediction performance of each individual random forest on the verification set, the Bayesian model combination algorithm was applied to formulate the combination strategy. The prediction values of individual random forest on the test set were fused to the final output through the model combination strategy. Finally, the proposed method was applied to solve the solar irradiance prediction problem. Simulation experiments were carried out by measured meteorological data. Other four kinds of prediction methods were also introduced to establish the contrast experiments, and the accuracy and reliability of the proposed method in the solar irradiance prediction were verified by comparison results.

Key words: K-means clustering; cross validation; random forest; Bayesian model combination; solar irradiance

---

* **收稿日期**:2018-04-25
基金项目:国家自然科学基金资助项目(61403274, 61773282), National Natural Science Foundation of China(61403274, 61773282)
作者简介:董娜(1983—),女,天津人,天津大学副教授,博士
† 通讯联系人,E-mail:dongna@tju.edu.cn

太阳能在光热领域和光电领域被广泛应用并被视为最佳代替能源. 季节、气候、云层密度等气候因素引起太阳能辐射量的不确定性制约了其应用领域的发展. 高精度的预测方法一直是太阳能预测研究的热点[1-2].

当前,太阳能辐照度的预测研究主要是使用支持向量机[3-6](SVM)和人工神经网络[7-9](ANN)算法. 这类学习算法难于平衡训练集的训练误差和测试集的泛化误差之间的关系[10],在训练过程中容易出现过拟合或欠拟合的现象. 然而, 在太阳能供热系统的热水供应量估计研究中,保证预测精度的同时预测结果的可靠性显得更为重要[11]. 集成学习为提高预测结果的可靠性提供了思路. 集成学习(ensemble learning,EL)[12]将多个基学习器组合在一起,常可获得比单个基学习器更显著的泛化性能和可靠性.

本文提出一种基于贝叶斯模型组合的随机森林预测方法 (Bayesian model combination-ensemble learning,BMC-EL)用于太阳能辐照度预测,使用随机森林作为基学习器建立集成学习模型. 首先引入 K-means 聚类[13]和 K 折交叉验证[14]将气象数据训练集划分为多组训练子集,以增加基学习器输入样本的多样性. 其次导入训练子集并训练各个随机森林. 之后, 依据个体随机森林在验证集上的预测精度,采用贝叶斯模型组合[15]算法制定个体随机森林的组合策略. 最后将各个随机森林在测试集上的预测输出依据模型组合策略得到最终太阳能辐照度预测值.

使用美国气象协会 2013-2014 年太阳能预测竞赛数据[16]作为数据集,使用经典 ANN、SVM 、Multikernel_SVM、K-means_RBF 算法建立基于气象数据的太阳能辐照度预测对照实验. 实验结果验证了提出的算法在太阳能辐照度预测研究中的准确性和可靠性.

1 训练子集多样性处理

待组合基学习器之间的差异性比较显著时,集成学习模型会拥有更好的性能. 故增加训练子集的多样性以提高基学习器输入样本的差异性. 基于气象数据的太阳能辐照度预测研究中,不同天气状况下气象数据呈现差异性,然而传统随机采样过程会导致训练子集中不同天气状况样本分布不均匀. 针对上述问题,提出 K-means 聚类和 K 折交叉验证方法增加训练子集的多样性, 如图 1.(为了区别 K-means 聚类和 K 交叉验证的下标,后文中将 K 折交叉验证改为 M 折交叉验证)

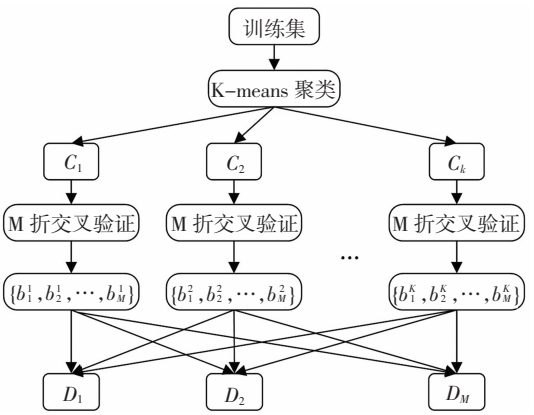


图 1 训练子集采样

Fig.1 Sampling of training subsets

假设需要生成 M 个训练子集$\{D_1, D_2, \cdots, D_M\}$. 对 K-means 聚类生成的簇 C_1 进行 M 折交叉验证并随机生成 M 个包$\{b_1, b_2, \cdots, b_M\}$. 将$\{b_2, b_3, \cdots, b_M\}$导入训练子集 D_1,将$\{b_1, b_3, \cdots, b_M\}$导入训练子集 D_2,依次将不同的 M-1 个包导入对应的训练子集, 直至将$\{b_1, b_2, \cdots, b_{M-1}\}$导入训练子集 D_M. 类似地,对簇$\{C_1, C_2, \cdots, C_k\}$ 都进行 M 折交叉验证, 并分别将其不同的 M-1 个包导入训练子集$\{D_1, D_2, \cdots, D_M\}$.

先聚类再交叉验证,可以使每个训练子集中都包含不同类型天气对应的气象数据,这保证了均匀采样. 交叉验证方法划分训练子集增加了训练子集的多样性.

2 随机森林基学习器

集成学习可以通过组合策略提高预测方法的可靠性. 随机森林中回归树的剪枝操作可以有效降低过拟合的风险,它简单高效,容易实现,计算开销小, 在很多分类回归问题中展现出强大的性能. 故本文采用随机森林算法作为基学习器.

本文采用 CART 回归树建立随机森林的基学习器. 训练集 D=$\{(x_1, y_1), (x_2, y_2), (x_3, y_3), \cdots, (x_n, y_n)\}$,输入样本 $x_i = (x_i^1, x_i^2, \cdots, x_i^Z)$包含 Z 个属性变量,输出 $Y = (y_1, y_2, \cdots, y_n)$为连续值. 回归树的节点对样本 $x_i(1 < i < n)$的属性变量 j 设置切分点 s,该输入变量大于 s 划分为一个区域, 否则划分到另一个区域. 对划分得到的区域使用不同的属性变量进

一步划分,依据节点的切分点将输入划分为 m 个区域,分别记为 $R_1,R_2,\cdots,R_m$. 定义每个区域的输出值分别为:$c_1,c_2,\cdots,c_m$. 则 CART 的模型为公式(1):

$$f(x)=\sum_{m=1}^{m}c_mI(x\in R_m) \tag{1}$$

其中 $I(x\in R_m)=\begin{cases}1 & (x\in R_m)\\ 0 & (x\notin R_m)\end{cases}$

回归树模型的平方误差为式(2):

$$E=\sum_{x_i\in R_m}(y_i-f(x_i))^2=\sum_{x_i\in R_m}(y_i-\sum_{m=1}^{m}c_mI(x\in R_m))^2 \tag{2}$$

由上式可得, 优化区域 R_m 的输出值 c_m 可以使得平方误差最小化. 易得当 c_m 为属于 R_m 区域的输入样本对应真实输出值的均值时,平方误差 E 最优,即 $\hat{c}m=\mathrm{ave}(y_i|x_i\in R_m)$.

假设选择样本中的变量 $x^{(j)}$为切分变量,节点取值 s 为切分点. 输入样本中变量 j 与切分点 s 比较就可以得到区域 $R_1(j,s)=\{x|x^{(j)}\leqslant s\}$和区域 $R_2(j,s)=\{x|x^{(j)}>s\}$. 当 j 和 s 设为确定值时,区域 $R_1(j,s)$和 $R_2(j,s)$包含的样本也确定. 故需要确定每个区域的输出值 c_1 和 c_2 使各自区间上的平方差最小,如式(3):

$$\min_{j,s}\left[\min_{c_1}\sum_{x_i\in R_1(j,s)}(y_i-c_1)^2+\min_{c_2}\sum_{x_i\in R_2(j,s)}(y_i-c_2)^2\right] \tag{3}$$

则 $\hat{c}_1=\mathrm{ave}(y_i|x_i\in R_1(j,s))$,$\hat{c}_2=\mathrm{ave}(y_i|x_i\in R_2(j,s))$.然后遍历样本中所有的变量,不同切分变量的最优切分点 s 得到的平方误差最小时记为最优切分变量 j. 类似的,对切分好的区域进一步划分,求取最优的切分变量和切分点,最终得到回归树 $f(x)=\sum_{m=1}^{m}c_mI(x\in R_m)$.

3　贝叶斯模型组合

3.1　贝叶斯模型平均

贝叶斯模型平均（Bayesian model averaging, BMA）是为解决模型的不确定性而提出的. 它是通过模型在验证集上预测精度的后验概率作为模型的权重, 对多个随机森林模型赋以合理的权重,解决单个模型的不确定性和单一性,将多个模型组合到一起的降低风险的方法.

给定数据集 D,样本 d_i 是由基学习器随机森林的输出值 x_i 和太阳能辐照度真实值 y_i 组成. 模型空间 H 是由有限个个体假设近似,h 作为模型空间的个体假设. 在模型空间和数据集 D 条件下 y_i 的后验分布为:

$$p(y_i|x_i,D,H)=\sum_{h\in H}p(y_i|x_i,h)p(h|D) \tag{4}$$

式中:$p(y_i|x_i,D,H)$为所有个体假设估计 y_i 的后验分布加权平均值, 其中,$p(y_i|x_i,h)=\int p(y_i|\theta_k,h,D)\times p(\theta_k|h,D)\mathrm{d}\theta_k$ 为假设空间 h 对 y_i 的预测分布, θ_k 是个体假设 h 对应的参数向量.

通过 BMA,数据集 D 下个体假设 h 的后验概率(h 假设作为数据生成模型的后验概率)$p(h|D)$可以由式(5)计算:

$$p(h|D)=\frac{p(D|h)p(h)}{\sum_{h\in H}p(D|h)p(h)} \tag{5}$$

式中:$\sum_{h\in H}p(D|h)p(h)$为常数,故 $p(h|D)\propto p(D|h)\times p(h)$.$p(D|h)=\int p(D|\theta_k,h)p(\theta_k|h)\mathrm{d}\theta_k$ 是个体假设 h 的积分似然估计,$p(\theta_k|h)$是 h 对应的向量参数 θ_k 的先验分布,$p(D|\theta_k,h)$是似然估计. $p(h)$是个体假设 h 的先验概率. 虽然集成学习方法中引入训练集采样扰动和属性扰动增加基学习器的差异性, 但是为保证所有基学习器都有较高的预测性能, 基学习器的初始参数设置并无差异,故先验概率 $p(h)$无需“偏袒”某一个个体假设,本文中 $p(h)=\frac{1}{k}$(k 为假设空间中个体假设的数量).

3.2　贝叶斯模型组合

贝叶斯方法在理论上是最优的,并且在许多任务中具有很好的性能. 贝叶斯模型平均也被视为集成学习中结合基学习器的一种标准方法. 然而在贝叶斯模型平均中,积分似然估计的计算方式容易使轻微精度提升的假设获得极高的权重[16],贝叶斯模型平均比 stacking 更容易过拟合, 对模型的近似误差非常敏感,且表现性能更差[17].

为了在太阳能辐照度预测试验中更加高效地获得集成学习的固有优势,组合策略应该侧重地反映各个假设空间的优势互补,而不仅仅是通过贝叶斯模型平均找出最优的假设.

针对上述问题,为贝叶斯模型平均增加假设空间 E 建立贝叶斯模型组合,将公式(4)修改为式(6):

$$p(y_i|x_i,D,H,E)=\sum_{h\in E}p(y_i|x_i,H,e)p(e|D) \quad (6)$$

式中:e 是组合模型空间 E 中的个体假设模型. 贝叶斯模型平均和贝叶斯模型组合示意图如图 2 和图 3 所示.

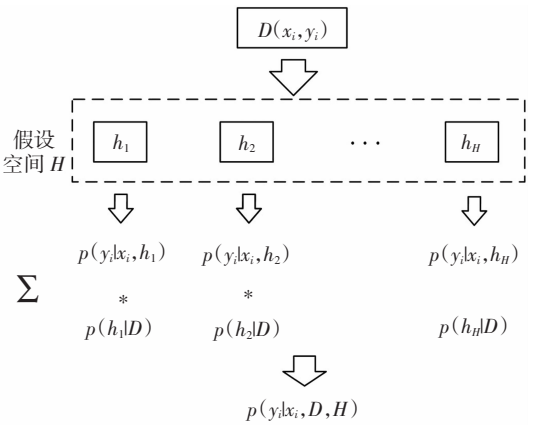


图 2 贝叶斯模型平均

Fig.2 Bayesian model averaging

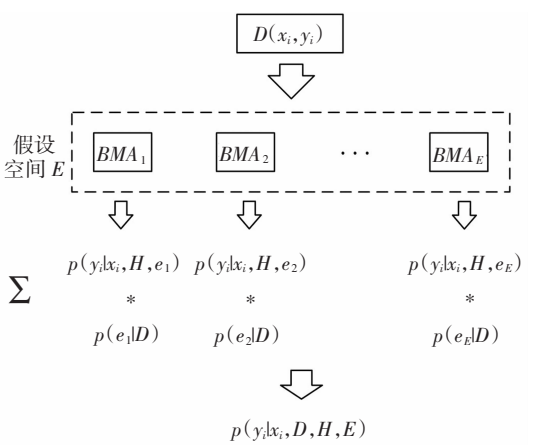


图 3 贝叶斯模型组合

Fig.3 Bayesian model combination

这样的修正克服了贝叶斯模型平均给个体假设 h 所有权重的倾向.

4 基于贝叶斯模型组合的随机森林预测方法

4.1 预测方法流程

基于贝叶斯模型组合的随机森林预测方法流程图如图 4 所示.

基于贝叶斯模型组合的随机森林预测方法在太阳能辐照度预测实验中的具体实施步骤如下.

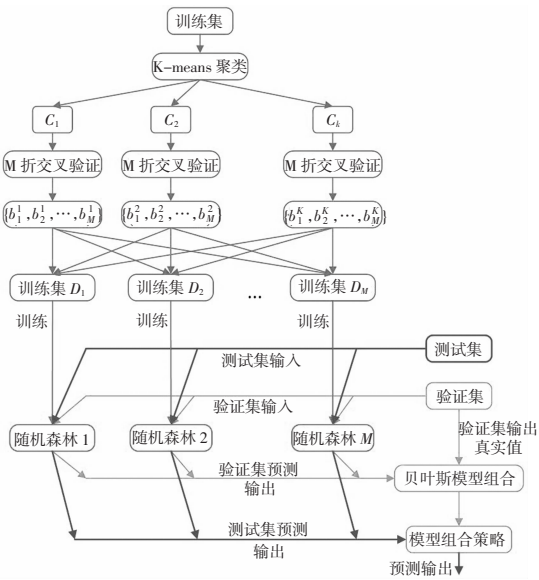


图 4 BMC-EL 预测方法流程图

Fig.4 The structure chart of BMC-EL

1) 首先采用公式 $x^*=\dfrac{x-\min(x)}{\max(x)-\min(x)}$ 将原始数据归一化处理,对训练集进行 K-means 聚类操作生成簇划分$\{c_1,c_2,\cdots,c_k\}$. 对每个簇 c_k 进行 M 折交叉验证,并依次生成 k 个训练子集$\{D_1,D_2,\cdots,D_k\}$. 采用 K-means 聚类和 M 折交叉验证划分训练子集,同时保证了训练子集的多样性和均匀采样.

2)使用多个 CART 回归树构建随机森林,将 k 个训练子集训练 k 个随机森林算法.

3)向训练好的 k 组随机森林导入验证集,输出 k 组预测输出值$(y_1,y_2,\cdots,y_k)$. 假设验证集的真实输出值为 y,构建矩阵$(y,y_1,y_2,\cdots,y_k)$并导入贝叶斯模型组合方法,根据 k 组随机森林在验证集的预测性能输出模型组合策略.

4)向训练好的 k 组随机森林导入测试集,个体随机森林输出各自模型预测值$(\bar{y}_1,\bar{y}_2,\cdots,\bar{y}_k)$,则集成学习方法的预测输出为 $p(Y|\bar{y}_k,D,H,E)=\sum_{h\in E}p(y_i|x_i,H,e)p(e|D)$.

4.2 预测方法的复杂度分析

CART 回归树在寻找切分节点时需要遍历当前特征的所有可能取值. 设数据样本具有 F 个特征,每个特征有 N 个切分点,CART 回归树共生成 S 个内部节点,则 CART 回归树的时间复杂度为 $O(F*N*S)$.

设每个随机森林基学习器中包含 M 个 CART 回归树,集成学习预测方法中共包含了 K 个基学习

器,则集成学习的时间复杂度为 $O(F*N*S*K*M)$.

在训练子集采样部分,设 k 个聚类中心,每个样本包含 F 个特征, 聚类中心的迭代次数为 t,则 K-means 聚类的时间复杂度为 $O(k*F*t)$. 设每个簇包含 m 个样本,则训练子集采样过程的时间复杂度为 $O(k*F*t+m*M)$.

贝叶斯模型平均包含 h 个体假设,验证集包含 c 个样本,则贝叶斯模型平均的算法复杂度为 $O(c*h)$. 贝叶斯模型组合的假设空间 E 包含 e 个体假设,则贝叶斯模型组合方法的时间复杂度为 $O(c*h*e)$.

由于切分点个数 N 较大,故集成学习的时间复杂度远大于训练子集采样和贝叶斯模型组合. 所以基于贝叶斯模型组合的随机森林预测方法的时间复杂度大于集成学习方法,训练子集采样和贝叶斯模型组合的时间复杂度相对较小.

5　太阳能辐照度预测实验

5.1　性能指标

均方误差(MSE)和绝对平均误差(MAE)作为太阳能辐照度预测的误差评价指标,本文额外定义了平均误差率(Average Error Rate,AER)和误差率小于 0.1 的预测成功率(Rate of success,RS)两个评价指标,如公式(7)~(9):

$$E_r = \frac{|Y_{pre} - Y_{real}|}{Y_{real}} \tag{7}$$

$$AER = \frac{\sum_{i=1}^{Num} E_r(i)}{Num} \tag{8}$$

$$RS = \frac{Num(error\_rate < 0.1)}{Num} \tag{9}$$

式中:Y_{pre} 是预测输出;Y_{real} 是真实值;E_r 是每个预测样本的误差率;AER 为平均误差率;Num 为预测结果的样本数;RS 表示精确预测样本的百分比, 它反映了预测结果的可靠性.

5.2　训练子集多样性

将原始气象数据归一化处理,然后对训练集进行 K-means 聚类操作. 太阳能辐照度预测实验中将训练集分为 10 个簇$\{C_1, C_2, \cdots, C_{10}\}$. 取输入样本的 dswrf_sfc,dswrf_sfc,tmp_sfc 三个属性建立三维坐标系, 簇中的样本在坐标系中分布如图 5 所示. 属性参数范围经过归一化处理到[-1,1]区间.

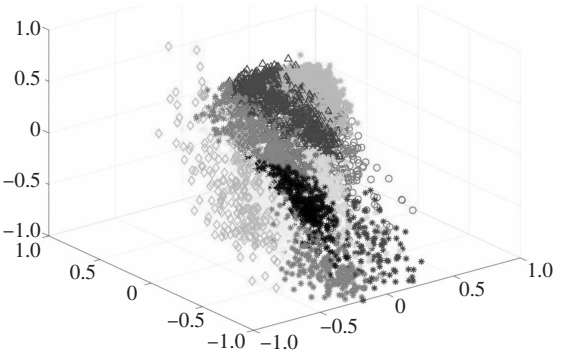
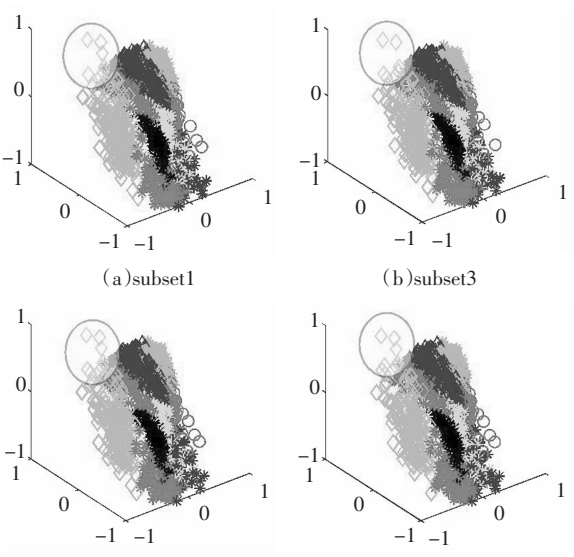


图 5　训练集中不同天气状况的气象样本分布

Fig.5　Distribution of different meteorological samples in training set

分别对上述划分的簇进行 10 折交叉验证,将其中不同的 9 个包分别导入各个训练子集. 4 个训练子集的气象样本分布如图 6 所示. 由于聚类后又采用 10 折交叉验证, 训练子集的样本量为训练集的 90%. 由 4 个训练子集的样本分布图可得, 采样过程并未影响不同天气状况的样本分布. K-means 聚类和 M 折交叉验证结合的采样过程不仅保证了均匀采样,同时增加了训练子集的差异性.



(a)subset1　(b)subset3

(c)subset5　(c)subset7

图 6　训练子集中不同天气状况的气象样本分布

Fig.6　Distribution of different meteorological samples in training subsets

5.3　模型误差估计及参数设置

随机森林回归模型是通过袋外数据(OOB)来估计模型误差的. 随机森林回归模型中 Bagging 采样理想状态下会有 37%的数据未被抽取,则将这些样本进行模型的误差估计. 由于随机森林算法本身的

属性扰动,只有当 CART 回归树的数量达到一定量级时，随机森林才会收敛到更低的泛化误差. 将所有气象预报样本导入单个随机森林算法,记录 OOB 误差与回归树数目之间的关系如图 7 所示.

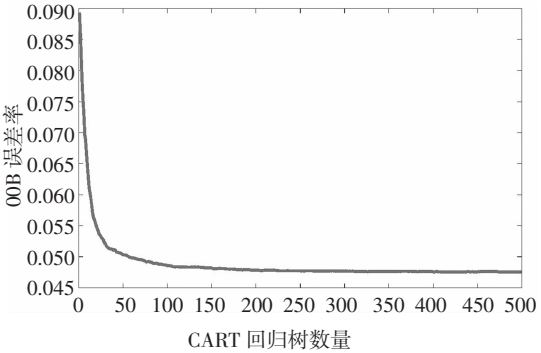


图 7 随机森林回归模型的误差估计图

Fig.7 Error estimation of random forest regression model

由图 7 可知，当 CART 回归树的数量到达 200 时，随机森林的 OOB 误差趋于收敛. 将太阳能辐照度预测实验中随机森林模型 CART 回归树数目设置为 200. 在太阳能辐照度仿真试验中,集成学习模型设置 10 组随机森林，对应上述 10 组训练子集，对照实验参数设置如表 1 所示. 对照实验中 EL 预测方法训练子集使用完整训练集,模型组合策略采用贝叶斯模型平均,其他设置与贝叶斯模型组合的随机森林预测方法一致.

表 1 太阳能辐照度预测对照实验的参数设置

Tab.1 The parameter settings of solar irradiance prediction experiments

对照实验	参数设置
ANN[7]	trainParam epoches = 1 trainParam goal=0.001 learning rate α=0.1
K-means_RBF[9]	density coefficient ψ=0 overlap coefficient ε=1 cluster radius a=1 trainParam.epoch=1 trainParam.goal=0.001 learning rate α=0.1
SVM[3]	cost=1 gama=1 model=epsilon-SVR epsilon=0.01 kernel=RBF
MultiKernel_SVM[5]	cost=1 gama=1 model=epsilon-SVR epsilon=0.01 $k(x,y)=0.15*[x^{T}y+c]+0.15*[(ax^{T}y+c)^{d}]+0.5*e^{-g\|x-y\|^{2}}+0.2*e^{-\frac{\|x-y\|^{2}}{2s^{2}}}$

5.4 实验结果

本章将 HOBA 中尺度站的太阳能辐照度和周围 GEFS 站点的气象数据作为数据集(2008 年之后的太阳能辐照度未公开),1994 年 1 月 1 日~2004 年 12 月 31 日的样本作为训练集. 使用随机函数(randvector,MATLAB）为聚类和交叉验证处理过的训练子集重新排序,缓解相似的气象预测样本在训练时接连出现. 将 2005 年 1 月 1 日~2006 年 12 月 31 日的样本作为验证集. BMC-EL 基于验证集制定模型组合策略,对照实验中利用验证集优化预测模型的超参数. 将 2007 年全年的气象预报样本作为测试集，用于太阳能辐照度预测实验. 对照实验中各类预测方法的太阳能辐照度输出值如图 8 所示.

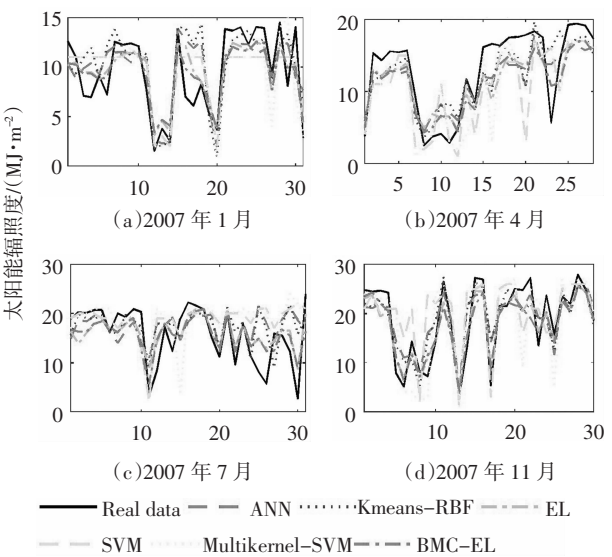


图 8 太阳能辐照度预测输出对比曲线

Fig.8 The comparison of solar irradiance prediction curve

图 8 中,对照实验中各类预测方法的预测曲线和真实曲线有相近的变化趋势. 然而各类预测方法在太阳能辐照度连续波动较大的样本出现较大的误差. SVM 和 Multikernel-SVM 算法在太阳能辐照度预测中分别出现了过拟合或欠拟合. 其中,BMC-EL 虽然也出现了较明显的偏差,但由于贝叶斯模型组合策略,其预测曲线更接近真实曲线.

为了直观地展示各类算法的预测误差,对照实验中太阳能辐照度预测输出的误差曲线如图 9 所示，其中 BMC-EL 方法预测误差曲线波动最小. 各类预测方法在 2007 年 7 月 30 日的太阳能辐照度预测时出现极大的偏差,最大的预测偏差近 20 MJ·m^{-2},然而 BMC-EL 此处的偏差略小于 10 MJ·m^{-2}. 由于贝叶斯模型组合是从不同的模型空间中选择最好的模型组合策略,故 BMC-EL 算法在一些较难预测的样本上依然保证了稳定的预测精度. 贝叶斯模型组合策略极大地提高了预测方法的可靠性.

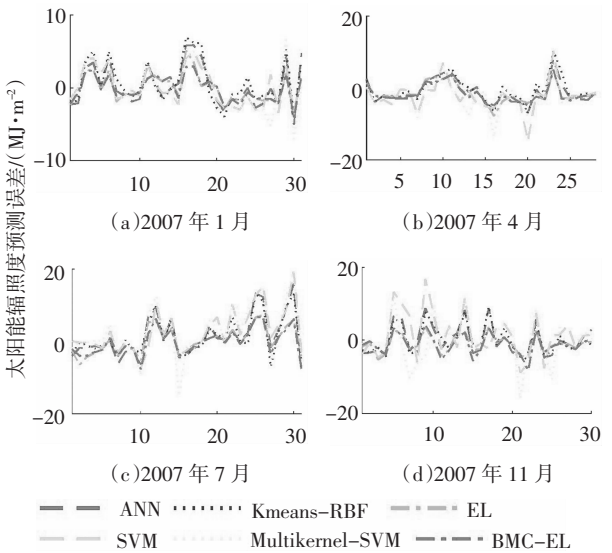


图 9　太阳能辐照度预测输出误差对比曲线

Fig.9　The comparison of solar irradiance prediction error

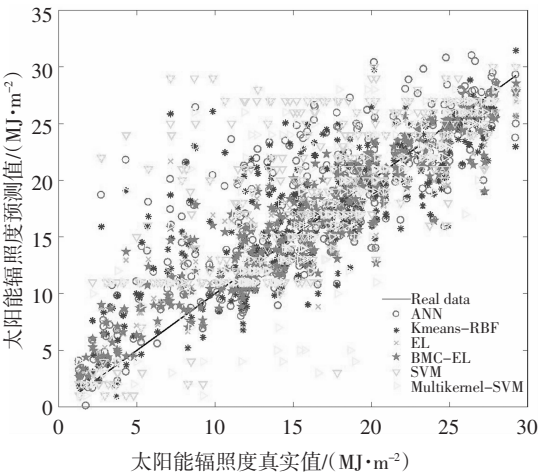


图 10　太阳能辐照度预测输出散点图

Fig.10　Scatter plot of solar irradiance prediction

在坐标系中绘制 $y = x$ 的直线表示预测辐照度和真实辐照度完全相同. 在散点图中样本点距离 $y = x$ 直线的距离越远则误差越大. 将各类预测方法在测试集上的预测输出和真实输出绘制到坐标系中,如图 10 所示.

图 10 中 BMC–EL 预测方法对应的散点更加集中,并且更加贴近于 $y=x$ 直线. 在 BMC–EL 的预测样本中, 在太阳能较为丰富的气象条件下 (晴,无云)散点最为集中,预测精度最高;太阳能辐照度匮乏的天气(阴,雨,多云,仪器无读数)包含了复杂的非线性,各类预测方法在气象条件糟糕的样本中都出现较大的偏差,而 BMC–EL 在这类样本点的预测偏差最小,它通过贝叶斯模型组合策略有效地提高了预测方法的可靠性.

各类预测方法在测试集在不同月份的平均性能指标如表 2 所示.

表 2　各类预测方法的性能指标对比

Tab.2　Performance index of the different methods

	BMC–EL	EL	ANN	Kmeans–RBF	Multikernel–SVM	SVM
MSE	6.79E+12	7.53E+12	1.95E+13	1.78E+13	2.95E+13	3.27E+13
MAE	1.862 8	1.900 2	3.244 8	3.014 84	3.629 57	3.882 53
RS	53.87%	51.42%	36.10%	40.08%	44.30%	40.18%
AER	0.208 5	0.210	0.317 96	0.296 6	0.346 8	0.373 8

综述太阳能辐照度预测实验结果,基于贝叶斯模型组合的随机森林预测方法在太阳能辐照度预测研究中具有非常好的预测性能,可靠性强,对不同的天气状态下的太阳能辐照度都能实现精确可靠的预测.

6　总　结

本文提出一种基于贝叶斯模型组合的随机森林方法用于太阳能辐照度预测, 首先引入 K–means 聚类和 M 折交叉验证将气象数据训练集生成多组彼此相交且不相同的训练子集,以增加随机森林输入样本的多样性. 其次将多组训练子集导入并训练集成学习模型中的个体随机森林. 之后, 将多组随机森林在验证集上的输出结果输入贝叶斯模型组合算法,依据验证集上预测性能的后验分布制定随机森林模型的组合策略. 最后各个随机森林在测试集上的预测输出经模型组合策略输出太阳能辐照度预测值. 在太阳能仿真实验中,BMC–EL 方法通过增加贝叶斯模型组合方法显著减少了单个随机森

林算法的不确定性,增加了太阳能预测输出的可靠性. 多组辐照度预测实验结果证明了所提出的预测方法预测精度高,可靠性强,可以精确可靠地预测不同气象环境中的太阳能辐照度.

参考文献

[1] 田翠霞,黄敏,朱启兵. 基于 EMD-LMD-LSSVM 联合模型的逐时太阳辐照度预测[J]. 太阳能学报,2018,39(2):504—512.
TIAN C X,HUANG M,ZHU Q B. Hourly solar irradiance forecast based on EMD-LMD-LSSVM joint model [J]. Acta Energiae Solaris Sinica,2018,39(2):504—512.(In Chinese)

[2] 路志英,任一墨,葛路琨. 基于样条估计分位数回归的光伏功率回归模型[J]. 湖南大学学报(自然科学版),2017,44(10):91—98.
LU Z Y,REN Y M,GE L K. Photovoltaic power regression model based on spline estimation and quantile regression [J]. Journal of Hunan University (Natural Sciences),2017,44 (10):91—98.(In Chinese)

[3] YANG X,JIANG F,LIU H. Short-term solar radiation prediction based on SVM with similar data[C]// Renewable Power Generation Conference. IET,2014:1.11—1.11.

[4] GUO W,MINGJIA L I,TAO L I,*et al*. Parameter identification of Hammerstein ARMAX model based on APSO-WLSSVM algorithm [J]. China Science Paper,2018,13(2):136—142.

[5] ALAM S,KANG M,PYUN J Y,*et al*. Performance of classification based on PCA,linear SVM,and multi-kernel SVM [C]//Eighth International Conference on Ubiquitous and Future Networks. IEEE,2016:987—989.

[6] ZHOU Y,CUI X,HU Q,*et al*. Improved multi-kernel SVM for multi-modal and imbalanced dialogue act classification [C]// International Joint Conference on Neural Networks. IEEE,2015: 1—8.

[7] RABBI K M,NANDI I,SALEH A S,*et al*. Prediction of solar irradiation in Bangladesh using artificial neural network (ANN) and data mapping using GIS technology[C]//2016 4th International Conference on the Development in the in Renewable Energy Technology(ICDRET). IEEE,2016:1—6.

[8] ANAMIKA,KUMAR N,AKELLA A K. Prediction and efficiency evaluation of solar energy resources by using mixed ANN and DEA approaches [C]// Pes General Meeting | Conference & Exposition. IEEE,2014:1—5.

[9] YADAV A K,MALIK H,CHANDEL S S. ANN based prediction of daily global solar radiation for photovoltaics applications[C]//India Conference. IEEE,2016:1—5.

[10] CHEN L G,CHIANG H D,DONG N,*et al*. Group-based chaos genetic algorithm and non-linear ensemble of neural networks for short-term load forecasting [J]. Iet Generation Transmission & Distribution,2016,10(6):1440—1447.

[11] BAILI H,LI Y F. Online reliability prediction of energy systems with wind generation [C]//International Midwest Symposium on Circuits and Systems. IEEE,2016:1—4.

[12] KROGH A,VEDELSBY J. Neural network ensembles,cross validation and active learning [C]//International Conference on Neural Information Processing Systems. MIT Press,1994:231—238.

[13] ALOISE D,DESHPANDE A,HANSEN P,*et al*. NP-hardness of Euclidean sum-of-squares clustering [J]. Machine Learning, 2009,75(2):245—248.

[14] 刘芳,夏洪山,艾军,等. 基于氧化动态模型的沥青热氧老化性能预测 [J]. 湖南大学学报 (自然科学版),2018,45(1):136—141.
LIU F,XIA H S,AI J,*et al*. Prediction of asphalt thermal oxidative aging performance based on oxidation dynamic model [J]. Journal of Hunan University (Natural Sciences),2018,45(1): 136—141.(In Chinese)

[15] MONTEIH K,CARROLL J L,SEPPI K,*et al*. Turning Bayesian model averaging into Bayesian model combination [C]// International Joint Conference on Neural Networks. IEEE,2011: 2657—2663.

[16] AMS 2013-2014 Solar energy prediction contest,forecast daily solar energy with an ensemble of weather models [EB/OL]. https://www.kaggle.com/c/ams-2014-solar-energy-prediction-contest.

[17] CLARKE B. Comparing Bayes model averaging and stacking when model approximation error cannot be ignored [J]. Journal of Machine Learning Research,2003,4(4):683—712.