第 39 卷　第 12 期　2015 年 6 月 25 日
DOI：10.7500/AEPS20140916005
电力系统自动化
Automation of Electric Power Systems
Vol.39 No.12 June 25，2015

基于灰色投影改进随机森林算法的电力系统短期负荷预测

吴潇雨，和敬涵，张　沛，胡　骏
（北京交通大学国家能源主动配电网技术研发中心，北京市 100044）

摘要：针对短期负荷预测领域传统的机器学习算法（如人工神经网络、支持向量机等）存在的诸如泛化性能不强、参数和模型结构确定困难等问题，将随机森林回归算法引入短期负荷预测领域。同时应用投影原理改进了传统的灰色关联相似日选取算法，提出了一种基于灰色投影改进随机森林算法的电力系统短期负荷预测组合方法。基于灰色投影的相似日选取方法，采用灰色关联度判断矩阵表征历史样本与待预测日影响因素间的关联关系，并用熵权法确立影响因素的权重对判断矩阵加权，最后利用各个样本关联度投影值排序得到相似日集合。采用随机森林算法建立预测模型，利用灰色投影筛选出的相似日样本集合训练模型，最后输入预测日特征向量（天气预报数值、日类型等）完成预测。以浙江电网某县级市的负荷数据作为实际算例，并将上述方法与支持向量机方法以及未作灰色投影改进的随机森林算法进行对比。实验结果表明，新方法具有较高的预测精度和鲁棒性。

关键词：短期负荷预测；相似日；灰色投影法；随机森林；Bagging 抽样方法；袋外估计

0　引言

电力系统负荷预测是指从电力负荷历史数据及其影响因素数据出发，运用某种数学方法去推测未来某段时间电力负荷需求情况[1]。长期以来，国内外学者对短期负荷预测的理论和方法做了大量的研究。其中传统的方法是以时间序列预测原理为基础建立起来的预测方法，以自回归（AR）方法、自回归滑动平均（ARMA）方法、累积式自回归滑动平均（ARIMA）方法等为代表[2-3]。该类方法具有所需数据少、模型简单且应用广泛的特点，但其对原始时间序列的平稳性要求较高，预测误差较大。

近年来，另一类以机器学习为理论基础的智能算法开始崭露头角。其中，人工神经网络（ANN）和支持向量机（SVM）是该类方法的典型代表。ANN 理论用于短期负荷预测的研究很多，其突出优点[4-5]是对大量非结构性、非精确性规律具有自适应功能，具有信息记忆、自主学习、知识推理和优化计算的特点。ANN 具有很强的自学习和复杂的非线性函数拟合能力，很适合于电力负荷预测问题，但研究过程中也表明 ANN 方法具有局部最优、泛化误差较大、隐单元数目难以确定等问题[6]。与 ANN 不同的是，SVM 方法在结构风险最小化准则（SRM）的基础上同时最小化经验风险和 VC 维（Vapnik-Chervonenkis dimension）的界，在预测对象上取得了较好的泛化性能。同时，其解决回归问题时，最后的问题等价为一个凸优化问题，又保证了其全局最优的特点[7-8]，这些都是在 ANN 预测法上取得的进步。但是，SVM 方法也存在许多缺陷，例如：核函数完全凭借经验选取，对于核参数和惩罚参数的确定，虽然有很多学者提出了粒子群寻优、遗传算法寻优等多种寻优手段[9]，但仍然存在着优化过程复杂、收敛速度慢等问题。总之，SVM 方法在模型构造上存在太多人为决定的因素，不利于预测精度和速度的进一步提高。

随机森林回归（random forest regression，RFR）算法是随机森林（RF）理论[10]的重要应用之一，是 Breiman L 于 2001 年提出的一种统计学习方法。RFR 算法是利用 Bootstrap 重抽样方法从原始样本中抽取多个样本，对每个 Bootstrap 样本集进行决策树建模，然后组合多棵决策树进行预测，并通过取平均值得出最终预测结果。其本质是利用组合多棵决策树做出预测的多决策树模型，该算法具有预测精度高、泛化误差可控、收敛速度快以及调节参数少等优点，可有效避免“过拟合”现象发生，适用于各种数据集的运算，尤其适用于超高维特征向量空间[11]。本文将 RFR 算法引入负荷预测领域，并在训练样本的选取上采用了灰色关联投影法选取相似

收稿日期：2014-09-16；修回日期：2014-11-10。
国家自然科学基金资助项目（51277009）。

日，最后的实际算例表明该组合算法在预测精度和鲁棒性方面均具有突出优势。

1　加权灰色关联投影法选择相似日

加权灰色关联投影法是构建在灰色系统理论和矢量投影原理上的一种综合评价方法[12]。该方法克服了仅利用灰色关联系数评价样本关联度的劣势，引入了加权和投影的概念，先利用适当的加权方法突出关键负荷影响因素，再利用历史样本在待预测日样本上的投影值来综合评价历史样本与待预测日样本的关联度，得出与预测日相似的历史日数据集。

1）选取影响电力负荷变化的若干关键因素，如湿度、气温、降水、风速、日类型等 m 个影响因素，则第 i 天样本的特征向量可以表示为：

$$\boldsymbol{Y}_i=[y_{i1}\quad y_{i2}\quad\cdots\quad y_{im}]\qquad i=1,2,\cdots,n \tag{1}$$

式中：n 为历史样本总数；y_{im} 为第 i 个样本的第 m 个影响因素值。

含有天气预报信息的待预测日特征向量为：

$$\boldsymbol{Y}_0=[y_{01}\quad y_{02}\quad\cdots\quad y_{0m}] \tag{2}$$

式中：y_{0m} 为待预测日特征向量的第 m 个影响因素值。

2）构建灰色关联判断矩阵，以 $\boldsymbol{Y}_0$ 为母序列（作为矩阵第 1 行元素），$\boldsymbol{Y}_i$ 为子序列，计算子序列与母序列间的关联系数，得到如下关联度判断矩阵：

$$\boldsymbol{F}=\begin{bmatrix}F_{01}&\cdots&F_{0m}\\ \vdots&&\vdots\\ F_{n1}&\cdots&F_{nm}\end{bmatrix} \tag{3}$$

式中：F_{nm} 为第 n 个样本的第 m 个因素对应的灰色关联度值。很显然，该矩阵的第 1 行（即母序列所在行）元素全为 1。

3）采用熵权法[13]确立各影响因素的权重，得到权向量如下：

$$\boldsymbol{W}=[w_1\quad w_2\quad\cdots\quad w_m] \tag{4}$$

式中：w_m 为第 m 个影响因素的权重值。

4）用上述权向量对灰色关联判断矩阵加权，得到加权灰色关联决策矩阵 $\boldsymbol{F}'$ 如下：

$$\boldsymbol{F}'=\boldsymbol{F}\boldsymbol{W}^{\mathrm{T}}=\begin{bmatrix}w_1&\cdots&w_m\\ \vdots&&\vdots\\ w_1F_{n1}&\cdots&w_mF_{nm}\end{bmatrix} \tag{5}$$

5）将矩阵 $\boldsymbol{F}'$ 中的每一行视为一个行向量，则第 1 行为待预测日行向量，记为 $\boldsymbol{A}_0$，其他每个历史样本行向量记为 $\boldsymbol{A}_i$。每个样本 $\boldsymbol{A}_i$ 与 $\boldsymbol{A}_0$ 向量间的夹角即是该样本的灰色投影角。因此，各个历史日行向量与待预测日行向量的灰色关联投影值为[12]：

$$D_i=\frac{\sum_{j=1}^{m}w_jF_{ij}w_j}{\sqrt{\sum_{j=1}^{m}(w_jF_{ij})^2}\sqrt{\sum_{j=1}^{m}w_j^2}} \tag{6}$$

式中：D_i 为第 i 个样本向量在待预测日向量上的投影值；$i=1,2,\cdots,n$。

6）根据各个历史日向量的灰色投影值按从大到小排序，设置投影值阈值，选择投影值较大的样本作为相似日样本集。

2　随机森林理论

随机森林是一种有监督的集成学习算法，其核心思想是将性能较弱的多个分类回归树（classification and regression tree，CART）经过一定规则组合成一片森林，结果由森林中所有的决策树投票得出。

2.1　CART 决策树

CART 决策树[14]是 Breiman L 等人于 1984 年提出的一种二分递归分割技术，在每个节点（除叶节点外）将当前样本集分割为两个子集。CART 算法所采用的属性选择量度是基尼指数（Gini index）。假设数据集 D 包含 m 个类别，那么其基尼指数 G_D 的计算公式为：

$$G_D=1-\sum_{j=1}^{m}p_j^2 \tag{7}$$

式中：p_j 为 j 类元素出现的频率。

基尼指数需要考虑每个属性的二元划分，假定属性 A 的二元划分将数据集 D 划分成 D_1 和 D_2，则此次在子节点以某属性 A 划分样本集 D 的基尼指数为：

$$G_{D,A}=\frac{|D_1|}{D}G_{D_1}(D_1)+\frac{|D_2|}{D}G_{D_2}(D_2) \tag{8}$$

对于每个属性，考虑每种可能的二元划分，最终选择该属性产生的最小基尼指数的子集作为其分裂子集。因此，在属性 A 上的基尼指数 $G_{D,A}$ 越小，则表示在属性 A 上的划分效果越好。在此规则下，由上至下不断分裂，直到整棵决策树生长完成[15-16]。

2.2　Bagging 方法和随机属性子空间抽样法

为改善 CART 决策树预测精度不高的劣势，Breiman L 于 1994 年引入了 Bagging（Bootstrap aggregating）算法。该方法利用 Bootstrap 可重复抽样从原始训练集中为每棵 CART 树抽取等规模的子训练集，研究表明[17]该方法能够有效提高不稳定基分类器的泛化能力。同时，Breiman 还在随机森林理论[10]中提出：CART 树在每个节点分裂时，采用随机抽取若干属性组成属性子空间进行选择分

裂。Bagging 方法增强了森林中单棵决策树的性能，而属性子空间抽样法则降低了每棵树间的相关性。结合 2.3 节中的定理 2 可知，这正是降低随机森林泛化误差的关键。

2.3 RFR 算法[10]

定义 1 随机森林 f 是决策树$\{h(\boldsymbol{X},\boldsymbol{\theta}_k),k=1,2,\cdots,N_{\text{tree}}\}$的集合，元分类器 $h(\boldsymbol{X},\boldsymbol{\theta}_k)$是用 CART 算法构建的未剪枝的 CART；$\boldsymbol{\theta}_k$ 是与第 k 棵决策树独立同分布的随机向量，表示该棵树的生长过程；采用多数投票法(针对分类)或求算术平均值(针对回归)得到随机森林的最终预测值。

定义 2 对输入向量 $\boldsymbol{X}$，最大包含 J 种不同类别，设 Y 为正确的分类类别，对于输入向量 $\boldsymbol{X}$ 和输出 Y，定义边缘函数为：

$$K(\boldsymbol{X},Y)=a_k I(h(\boldsymbol{X},\boldsymbol{\theta}_k)=Y)-\max_{j\neq Y} a_k I(h(\boldsymbol{X},\boldsymbol{\theta}_k)=j) \tag{9}$$

式中：j 为 J 种类别中的某一类；$I(\cdot)$为指示函数；a_k 为取平均函数；$k=1,2,\cdots,n$。

从式(9)可看出，函数 K 描述了对向量 $\boldsymbol{X}$ 正确分类 Y 的平均得票数超过其他任何分类的平均得票数的最大值。因此，边缘函数越大，正确分类的置信度就越高。由此定义随机森林的泛化误差为：

$$E^{*}=P_{\boldsymbol{X},Y}(K(\boldsymbol{X},Y)<0) \tag{10}$$

式中：$P_{\boldsymbol{X},Y}$为对给定输入向量 $\boldsymbol{X}$ 的分类错误率函数。当森林中决策树数目较大时，利用大数定律得到如下定理。

定理 1 当树的数目增加时，对于所有序列 $\boldsymbol{\theta}_k$，E^{*} 几乎处处收敛于：

$$P_{\boldsymbol{X},Y}(P_{\boldsymbol{\theta}}(h(\boldsymbol{X},\boldsymbol{\theta})=Y)-\max_{j\neq Y} P_{\boldsymbol{\theta}}(h(\boldsymbol{X},\boldsymbol{\theta})=j)<0) \tag{11}$$

式中：$P_{\boldsymbol{\theta}}$ 为对于给定序列 $\boldsymbol{\theta}$ 的分类错误率。

该定理表明随机森林的泛化误差随着树的数目增加不会造成过拟合，而会趋于某一上界。

定理 2 随机森林泛化误差的上界为：

$$E^{*}\leqslant\frac{\bar{\rho}(1-s^{2})}{s^{2}} \tag{12}$$

式中：$\bar{\rho}$ 和 s 分别为树的平均相关系数和平均强度。

由定理 2 可知，随着树的相关性的降低和单棵树强度的提高，随机森林的泛化误差上界将会减小，其泛化误差将会得到有效的控制。因此，提高随机森林预测精度主要有 2 条途径，即降低树相关性以及提高单分类器(即单棵决策树)性能，具体的 RFR 算法流程参见附录 A。

2.4 随机森林的统计学优点

1)随机森林仅需调整 2 个参数，即森林中树的数量 N_{tree}和每棵树选取的分裂特征数 M_{try}。

2)在大数定律的保证下，随机森林具有很高的分类准确率，且不会出现过拟合。

3)随机森林还有一个特点是袋外(out-of-bag，OOB)估计[10]，当通过 Bagging 生成训练子集时，对于每一棵 CART 树，原始样本集 S 中接近 37%的样本不会出现在该树的训练子集中，这些样本被称为 OOB 样本。OOB 样本可以用来估计随机森林的泛化误差，也可以计算每一特征的重要性。

3 加权灰色投影改进随机森林算法流程

综上所述，本文提出的基于加权灰色关联投影改进随机森林算法的步骤如下。

1)对历史样本集进行相似日选取。采用第 1 节中的加权灰色关联投影法，形成具有高度相似性的相似日样本集，样本含有 11 个输入特征维，格式参见附录 B。

2)对相似日样本集进行 Bootstrap 重抽样，生成 k 个子训练集。

3)根据 2.3 节中的算法生成对应的 k 棵 CART 决策树，在此过程中，随机选取的特征数目取$M_{\text{try}}=\log_2(M+1)$($M$ 为样本输入特征的维数)，而随机森林的规模需根据预测结果调整 N_{tree}的大小。

4)将待预测日特征向量$\boldsymbol{Y}_0=[y_{01},y_{02},\cdots,y_{0m}]$输入上述随机森林模型，求取各棵树输出的平均值，得到负荷预测结果，整体算法流程如图 1 所示。

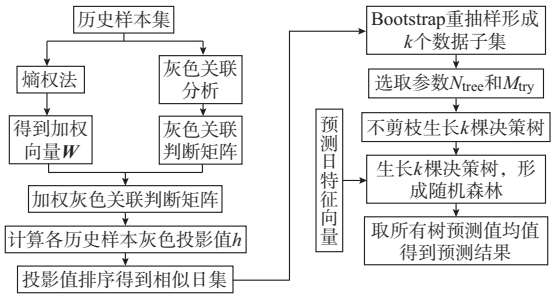

图 1 灰色投影改进随机森林算法流程

Fig.1 Flow chart of random forest algorithm improved by grey relation projection method

4 预测实例及结果分析

4.1 样本数据

选取浙江电网某市 2012 年 1 月至 4 月期间的负荷数据作为训练样本，2012 年 4 月 21 至 5 月1 日的负荷数据作为测试样本。需要指出的是，2012 年 4 月 21 日至 22 日为双休日，23 日至 28 日为工作日，4 月 29 日至 5 月 1 日为五一节假日。为突出本算法的优势，本文选取 SVM 方法、未做改进的随机森林算法以及利用灰色投影选择相似日改进的随机

森林(IRF)算法3种方法对2012年4月21至5月1日连续11 d的预测结果进行预测，并比较三者的预测精度。

4.2　误差分析标准

本文结合负荷预测实际应用需求，选取平均相对误差 e 作为预测方法的效果判断依据：

$$e=\frac{1}{n'}\left|\frac{R(i)-F(i)}{R(i)}\right|\times 100\% \tag{13}$$

式中：$R(i)$和$F(i)$分别为1 d中某时刻实际的负荷值和预测的负荷值；n'为1 d中的预测点总数；$i=1,2,\cdots,n'$。

本文以4%为判断标准，若某点的$e>4\%$，则判定该点的预测结果不合格。则定义 r 为该工作日负荷预测结果的不合格率：

$$r=\frac{N(e>4\%)}{n'}\times 100\% \tag{14}$$

式中：$N(e>4\%)$为某工作日预测结果相对误差超过4%的点数。

4.3　预测结果分析

按照图1所示的算法流程，首先通过灰色关联投影法选取待预测日的相似日训练集。如待预测日为2012年4月27日，将2012年1月1日至2012年4月26日的历史数据作为样本数据，每个采样时刻的数据作为一个样本集分析，如1月1日至4月26日10:00:00样本(共117个样本)作为一个样本集。通过第1节所述加权灰色关联投影值的计算，以10:00:00的样本数据为例。计算结果如图2所示。

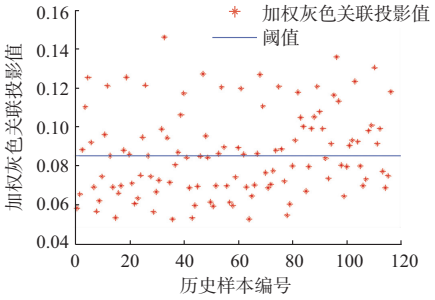

图2　历史样本的加权灰色投影值
Fig.2　Weighted grey projection value of historical samples

图2中的红色数据点即是117个样本的灰色投影值，设置投影值大于0.85(如图中蓝色实线所示)的55个样本作为4月27日10:00:00的相似日样本集，用来训练随机森林模型。由2.3节中定理2可知，加权相关系数 $\bar{\rho}$ 和 s 是影响随机森林模型预测精度的关键因子，而实际应用中，在输入向量确定的条件下，随机森林中决策树棵数 N_{tree} 及分裂特征集中的特征个数 M_{try} 对预测精度及泛化能力有较大影响。经调试，设定 N_{tree} 为500，M_{try} 为3时模型具有较好的预测效果。其他2种方法分别选取待预测日前3个月的同类型日作为训练集，3种方法的预测精度见表1。

表1　2012年4月21日至5月1日预测精度比较
Table 1　Comparison of three methods of prediction accuracy from April 21 to May 1, 2012

日期	SVM		RF		IRF	
	e/%	r/%	e/%	r/%	e/%	r/%
2012-04-21	2.67	5.21	1.17	0	1.00	0
2012-04-22	2.30	3.13	2.14	3.13	1.83	3.13
2012-04-23	2.72	6.25	3.09	5.21	2.22	3.13
2012-04-24	2.06	2.08	2.63	7.29	1.76	2.08
2012-04-25	2.42	4.17	2.51	2.08	2.23	3.13
2012-04-26	2.79	6.25	3.30	8.33	2.50	6.25
2012-04-27	1.99	2.08	2.57	5.21	1.33	1.04
2012-04-28	2.67	6.25	2.54	3.13	1.73	0
2012-04-29	2.54	6.25	3.03	6.25	2.36	5.21
2012-04-30	4.45	10.42	2.97	7.29	1.59	0
2012-05-01	4.89	12.50	4.11	9.38	2.87	1.35
总平均值	2.86	5.87	2.73	5.21	1.95	2.30

表2和图3展示了2014年4月27日(正常工作日)内的预测误差。

表2　2012年4月27日3种算法预测误差比较
Table 2　Comparison of three methods of prediction accuracy on April 27, 2012

预测时刻	相对误差/%		
	SVM	RF	IRF
00:00:00	0.55	0.06	0.14
00:01:00	3.54	4.21	3.54
00:02:00	4.13	3.16	1.36
00:03:00	3.61	2.43	1.36
00:04:00	5.06	3.68	2.54
00:05:00	4.64	4.53	1.92
00:06:00	2.83	2.26	0.29
00:07:00	3.04	3.19	3.07
00:08:00	2.52	0.65	0.56
00:09:00	2.26	1.26	0.30
00:10:00	5.50	0.67	0.51
00:11:00	0.25	1.07	5.04
00:12:00	2.73	1.56	0.74
00:13:00	1.46	0.26	1.78
00:14:00	5.23	3.17	0.25
00:15:00	0.31	0.58	1.25
00:16:00	1.67	1.29	0.83
00:17:00	1.57	1.35	2.29
00:18:00	0.89	1.65	0.13
00:19:00	2.93	2.46	1.19
00:20:00	0.02	0.46	0.22
00:21:00	0.64	1.93	0.99
00:22:00	2.85	2.55	0.70
00:23:00	3.49	3.19	0.82
平均值	1.99	2.57	1.33

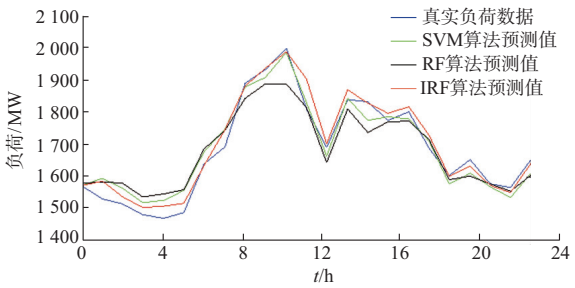

图 3　2012 年 4 月 27 日 3 种方法预测的负荷曲线
Fig.3　Comparison of three methods of forecasting load curve

由于篇幅所限，4 月 22 日（双休日）、4 月 30 日（五一节假日）的 24 h 预测结果展示在附录 C 中。结合表 1 和表 2 可以看出，随机森林算法的预测误差略小于 SVM 方法，而经过改进的随机森林方法误差要明显小于 SVM 方法和未经改进的随机森林算法。在鲁棒性方面，4 月 29 日至 5 月 1 日为重大节假日，在其他两种算法误差均较普通日预测出现较大误差时，本文算法仍保持较低的误差，显示了较强的鲁棒性。

5　结语

本文将随机森林算法应用到负荷预测领域，同时采用灰色关联投影法选取待预测日相似日，达到简化模型训练，提高预测精度的目的。该算法在文中数学证明和实例分析中体现出了以下优势。

1）充分考虑待预测日与历史日特征向量间的耦合关系，采用投影值这一综合考虑向量模值与夹角值的综合评价指标。对训练样本进行了有效的约简，减少了计算量，降低了预测误差。

2）预测模型采用随机森林算法，文中数学证明了该算法泛化误差可控的特点。经过加权灰色关联投影筛选出的样本作为该算法的学习集，缩小了训练集规模，提高了预测精度。

3）经过与 SVM 方法以及未经灰色关联加权选取相似日训练集训练的随机森林方法进行比较，证明该方法有效提高了短期负荷预测系统的精度。

附录见本刊网络版（http://www.aeps-info.com/aeps/ch/index.aspx）。

参考文献

[1] 康重庆，夏清，张伯明.电力系统负荷预测研究综述与发展方向的探讨[J].电力系统自动化，2004，28(17)：1-11.
KANG Chongqing，XIA Qing，ZHANG Boming. Review of power system load forecasting and its development [J]. Automation of Electric Power Systems，2004，28(17)：1-11.

[2] 刘晨晖.电力系统负荷预报理论与方法[M].哈尔滨：哈尔滨工业大学出版社，1987.

[3] 牛东晓，魏亚楠.基于 FHNN 相似日聚类自适应权重的短期电力负荷组合预测[J].电力系统自动化，2013，37(3)：54-57.
NIU Dongxiao，WEI Yanan. Short-term power load combinatorial forecast adaptively weighted by FHNN similar-day clustering [J]. Automation of Electric Power Systems，2013，37(3)：54-57.

[4] PARK D C，EL-SHARKAWI M A. Electric load forecasting using an artificial neural network [J]. IEEE Trans on Power Systems，1991，6(2)：442-448.

[5] KHOTANZAD A，AFKHAMI-ROHANI R，LU T L，et al. ANNSTLF-a neural network based electric load forecasting system[J]. IEEE Trans on Neural Networks，1997，8(4)：835-846.

[6] 赵登福，张涛，杨增辉，等.基于 GN-BFGS 算法的 RBF 神经网络短期负荷预测[J].电力系统自动化，2003，27(4)：1-4.
ZHAO Dengfu，ZHANG Tao，YANG Zenghui，et al. Short-term load forecasting using radial basis function (RBF) neural networks based on GN-BFGS algorithm [J]. Automation of Electric Power Systems，2003，27(4)：1-4.

[7] 畅广辉，刘涤尘，熊浩.基于多分辨率 SVM 回归估计的短期负荷预测[J].电力系统自动化，2007，31(9)：37-41.
CHANG Guanghui，LIU Dichen，XIONG Hao. Short term load forecasting based on multi-resolution SVM regression [J]. Automation of Electric Power Systems，2007，31(9)：37-41.

[8] 黄帅栋，卫志农，高宗和，等.基于非负矩阵分解的相关向量机短期负荷预测模型[J].电力系统自动化，2012，36(11)：62-65.
HUANG Shuaidong，WEI Zhinong，GAO Zonghe，et al. A short-term load forecasting model based on relevance vector machine with nonnegative matrix factorization[J]. Automation of Electric Power Systems，2012，36(11)：62-65.

[9] 陆宁，武本令，刘颖.基于自适应粒子群优化的 SVM 模型在负荷预测中的应用[J].电力系统保护与控制，2011，39(15)：44-51.
LU Ning，WU Benling，LIU Ying. Application of support vector machine model in load forecasting based on adaptive particle swarm optimization[J]. Power System Protection and Control，2011，39(15)：44-51.

[10] BREIMAN L. Random forests[J]. Machine Learning，2001，45(1)：5-32.

[11] 许保勋.面向高维数据的随机森林算法优化研究[D].哈尔滨：哈尔滨工业大学，2013.

[12] 吕锋，崔晓辉.多目标决策灰色关联投影法及其应用[J].系统工程理论与实践，2002，22(1)：103-107.
LÜ Feng，CUI Xiaohui. Multi-criteria decision grey relation projection method and its application[J]. Systems Engineering-Theory & Practice，2002，22(1)：103-107.

[13] 欧阳森，石怡理.改进熵权法及其在电能质量评估中的应用[J].电力系统自动化，2013，37(21)：100-106.
OUYANG Sen，SHI Yili. A new improved entropy method and its application in power quality evaluation [J]. Journal of Hydraulic Engineering，2013，37(21)：100-106.

[14] BREIMAN L，FREIDMAN J H，OLSHEN R A，et al. Classification and regression trees [M]. Chapman & Hau / CRC，1984.

[15] 张松林.CART-分类与回归树方法介绍[J].火山地质与矿产，1997，18(1)：63-73.
ZHANG Songlin. An introduction to the methodology of CART-classification and regression trees[J]. Volcanology & Mineral Resources，1997，18(1)：63-73.

[16] 朱六璋.短期负荷预测的组合数据挖掘算法[J].电力系统自动

化，2006，30(14)：82-86.
ZHU Liuzhang. Short-term electric load forecasting with combined data mining algorithm[J]. Automation of Electric Power Systems, 2006, 30(14): 82-86.
[17] BREIMAN L. Bagging predictors[J]. Machine Learning, 1996, 24(2): 123-140.

吴潇雨(1991—)，男，通信作者，博士研究生，主要研究方向：电力系统负荷预测、电力系统保护与控制。E-mail：13117375@bjtu.edu.cn

和敬涵(1964—)，女，教授，博士生导师，主要研究方向：智能电网、电力系统保护与控制、电动汽车与V2G等。E-mail：jhhe@bjtu.edu.cn

张　沛(1972—)，男，教授级高级工程师，主要研究方向：电力大数据应用开发。E-mail：2512692577@qq.com

(编辑　孔丽蓓)

Power System Short-term Load Forecasting Based on Improved Random Forest with Grey Relation Projection

WU Xiaoyu, HE Jinghan, ZHANG Pei, HU Jun

(National Active Distribution Network Technology Research Center, Beijing Jiaotong University, Beijing 100044, China)

Abstract: In view of the problems with typical machine learning algorithms (for example artificial neural network (ANN) and support vector machine (SVM)), such as the difficulty in determining the number of hidden units and parameter optimization, a random forest regression method is first introduced to power system load forecast. A new combinatorial algorithm involving two steps is proposed. Firstly, a grey relational judgment matrix is built to characterize the relationship between historical samples and forecasting sample. Secondly, the entropy method is used to determine the weights of all load influencing factors and the weighting matrix is got. Thirdly, the historical samples with bigger grey relation projection values are used to form the training set. After getting the training set, this data set is used to train random forest models. Then, the eigenvectors of the forecasting day are input to the trained model to finish the forecasting process. The real load data of one city in Zhejiang Province is used to test the proposed algorithm, and the results are compared with SVM and random forest method with no improvement made on grey relation projection. The results show that the new combinatorial method has higher precision and robustness than the other two methods.

This work is supported by National Natural Science Foundation of China (No. 51277009).

Key words: short-term load forecasting; similar day; grey relation projection method; random forest; Bagging sampling method; out-of-bag estimation