第 43 卷 第 11 期 计 算 机 工 程 2017 年 11 月
Vol. 43 No. 11 Computer Engineering November 2017

·开发研究与工程应用· 文章编号：1000-3428(2017)11-0317-05 文献标志码：A 中图分类号：TP391

基于流形算法与 RBF 网络的超短期风速预测

张雪松[1]，朱 想[2]，赵 波[1]，魏海坤[3]，邵海见[3]

（1. 国家电网浙江省电力公司 电力科学研究院，杭州 310014；2. 中国电力科学研究院，南京 210003；
3. 东南大学 自动化学院，南京 210096）

摘 要： 传统的风速预测方法往往通过经验来确定模型结构，未考虑输入变量选取、系统的动态特性等问题，导致系统在不同时间尺度下的动态特性没有得以充分反映，降低模型的推广泛化能力。针对上述问题，提出一种基于流形算法和 RBF 网络相结合的方法，通过模型结构设计和本质特征提取等方法，增加模型预测结果的稳定性和鲁棒性，以提高模型的推广能力。以华东某风电场数据进行实验分析，结果表明，与传统风速预测方法相比，该模型结构选择方法可提高模型计算效率，降低样本复杂度，能够得到更好的预测效果。

关键词： 超短期风速预测；模型结构选择；RBF 网络；流形算法；机器学习

中文引用格式： 张雪松，朱 想，赵 波，等. 基于流形算法与 RBF 网络的超短期风速预测[J]. 计算机工程，2017，43(11)：317-321.

英文引用格式： ZHANG Xuesong，ZHU Xiang，ZHAO Bo，et al. Ultra-short Term Wind Speed Forecast Based on Manifold Algorithm and RBF Network[J]. Computer Engineering，2017，43(11)：317-321.

Ultra-short Term Wind Speed Forecast Based on Manifold Algorithm and RBF Network

ZHANG Xuesong[1]，ZHU Xiang[2]，ZHAO Bo[1]，WEI Haikun[3]，SHAO Haijian[3]

（1. Electric Power Research Institute of State Grid，Zhejiang Electric Power Company，Hangzhou 310014，China；
2. China Electric Power Research Institute，Nanjing 210003，China；3. School of Automation，Southeast University，Nanjing 210096，China）

【Abstract】 The traditional forecast design method depends on the experiences from the designer，which cannot consider the nature of the wind speed signal changes，it results in the low generality ability of the model structure. Therefore，the RBF neural network in combination with the manifold algorithm is proposed to design the model structure and extract essential features in order to increase the stability and robustness，and improve the forecast accuracy and generality ability. Experimental results using the data from a real wind farm in East China show that，compared with the traditional wind speed forecast methods，the proposed model structure selection method can improve the computing efficiency，reduce sample complexity，and has better forecast effect.

【Key words】 ultra-short term wind speed forecast；model structure selection；RBF network；manifold algorithm；machine learning

DOI： 10.3969/j.issn.1000-3428.2017.11.051

0 概述

风速的变化具有很强的间歇性和波动性，风电功率与风速的大小存在着一定的对应关系，因此风电功率也具有随机性和间歇性。由于风速是影响风电功率的重要因素，因此短期风速预测对于电网调度和协调优化具有重要作用。如果能够准确地预报风电场的风速，不但可以有效地降低风速的确定性对于电力网络的负面影响，保障电力网络的电能质量，降低系统的旋转备用容量，而且可以提高系统对风电的消纳能力[1-4]。

风速数据的数据量极大，在建模过程中由于高维数据集合中并不是所有的数据和变量都是“重要和必须的”，虽然通过大量的计算可以得到高精度和准确的模型，但是在实际应用中，对于原始数据的精简仍然有很多人感兴趣[5]，尤其是高维数据容易造成维数灾难。引入流形学习算法，对数据进行维数简约，即可以消除原有数据的冗杂信息，保持了原有

作者简介： 张雪松（1979—），男，高级工程师，主研方向为分布式新能源发电技术、微电网技术；朱 想，硕士；赵 波，高级工程师；魏海坤，教授；邵海见，博士。

收稿日期： 2016-09-01 **修回日期：** 2016-11-08 **E-mail：** haijianshao@seu.edu.cn

数据的几何结构,又有利于模型计算效率和推广能力的提高。除此之外,由于实际数据的数量很大,数据的精简有利于挖掘数据的本质特征,提高模型计算效率和推广能力。综上所述,模型结构选择技术对于模型泛化能力与推广能力的提高都有重要意义。

由于目前的风速预测方法没有考虑样本的动态特性以及变量维数太大带来的计算效率低下等问题[6-7],导致了样本在各个时间尺度下的动态特性没有得以反映,使得模型在学习样本特性时没有全面反映样本的本质特性,从而导致了模型推广能力的降低[8]。因此,本文重点考虑模型样本信息学习的动态特性,用以提高短期风速预测的精度,从而有效地提高电网对风电的接纳能力。本文首先采用流形学习算法进行方法选取、本质维数估计以及邻域结构设计等,用于提高模型的计算效率和推广能力。然后采用 RBF 神经网络建立预测模型。最后结合华东某风电场的数据进行验证,并与传统无任何结构设计的方法进行了比较。

1 数据降维方法

1.1 理论分析

数据维数降维方法在数学上可以表述为[9]:寻找恰当维数的 $k(k<p)$ 维变量 $\boldsymbol{s}=(s_1,s_2,\cdots,s_k)^{\mathrm{T}}$,在给定的判据下能够代表给定 p 维的随机变量 $\boldsymbol{x}=(x_1,x_2,\cdots,x_p)^{\mathrm{T}}$ 的特性或者特征,元素 $\boldsymbol{s}$ 有时候可以称为隐藏元素。这里的变量一般是统计术语,而在计算机科学和机器学习领域常常称为特征或者特性。

$$E(x)=\boldsymbol{\mu}=(\mu_1,\mu_2,\cdots,\mu_p)^{\mathrm{T}}$$

$$\sigma\{(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\}=\sum\nolimits_{p\times p}(\boldsymbol{x},\boldsymbol{\mu})$$

对应于 $\boldsymbol{x}$ 中元素的列向量具有 n 个观测向量,$\boldsymbol{x}=\{x_{ij},1\leqslant i\leqslant p,1\leqslant j\leqslant n\}$,而且其中每个元素的均值与方差分别是:

$$\hat{\mu}_i=\frac{1}{n}\sum_{j=1,2,\cdots,n}x_{ij},\ \hat{\sigma}_i=\frac{1}{n}\sum_{j=1,2,\cdots,n}(x_{ij}-\hat{\mu}_i)^2$$

不失一般性,针对于线性技术,其相应的表达式表述为:

$$\boldsymbol{s}=\boldsymbol{w}\boldsymbol{x},\ s_i=\sum_{i=1,2,\cdots,p;j=1,2,\cdots,n}w_{i,j}x_j$$

其中,$\boldsymbol{w}$ 是线性变换矩阵。同样地,其逆变换可以通过上式求得,那么本文通过非线性以及线性的变换矩阵的求取进行理论分析和策略设计。事实上,无必要选取所有的方法进行理论分析以及策略设计。根据之前的讨论,数据维数监督学习算法主要针对多维数变量设定目标函数最大化使得样本特征有区分度,而无监督学习主要是力求从高维空间到低维数空间映射的建立使得信息的损失最小。

1.2 常用流形算法

目前把流形引入到机器学习领域主要有 2 种用途:一种是将原来在欧氏空间中适用的算法加以改造,使得它工作在流形上,直接或间接地对流形的结构和性质加以利用;另一种是直接分析流形的结构,并试图将其映射到一个欧氏空间中,再在得到的结果上运用以前适用于欧氏空间的算法来进行学习[10-12]。常用的流行算法对比如表 1 所示[13-14]。

表 1 常用流行算法特征

算法	所保持的几何属性	全局/局部关系	计算复杂度
ISOMAP	点对测地距离	全局	非常高
LLE	局部线性重构关系	局部	低
LE	局部邻域相似度	局部	低
HLLE	局部等距性	局部	高
LTSA	局部坐标表示	全局 + 局部	低
MVU	局部距离	全局 + 局部	非常高
Logmap	测地距离与方向	局部	非常低
Diffusion Maps	Diffusion 距离	全局	中等

1.3 本质维数估计

对于时间序列的降维,其内在维数和本质维数的估计是非常重要的,因为其决定了时间序列在降维后的信息量问题。本文对于时间序列的降维主要基于信息准则进行。文献[15]指出高维数据集合的本质维数可以通过定义一个独立的标量来定性分析,比如常用的 Nystrom 方法就是一类典型的用于高维数据集合逼近的方法。类似于分形维数(fractal dimensions)的处理思想,本文将基于信息相关准则来估计内在维数。定义度量空间 $\Re$ 的一个有限集合 $\zeta_n=\{x_1,x_2,\cdots,x_n\}$,则关于高维数据集合降维前后的相关维定义如下:

$$C_n(r)=\alpha\sum_{i=1,\cdots,n;j=i+1,\cdots,n}I_r,I_r=\{\|x_i-x_j\|^2<r\}$$

其中,$\alpha=2\times(n(n-1))^{-1}$,$I_r$ 是对应的指标集合。并且根据定义的邻域大小,对于在该度量空间的可数子集 $S=\{x_1,x_2,\cdots,x_n\}\subset\Re$,存在相关极限 $C(r)=\lim_{r\to 0}C_n(r)$。如果 $C(r)$ 可以通过计算得到,那么关于数据集合 ζ_n 降维前后的相关性定义如下:

$$CD_{\mathrm{corr}}=\lim_{r\to 0}\frac{\ln C(r)}{\ln r}$$

从本质上讲,关于包含有限样本的有限集合 r 是可以获取的,CD_{corr} 容易计算得到,但是在实际上其计算结果往往不容易获取。通常的做法是通过考察对于相关维数曲线的线性逼近的梯度 $\frac{\partial\ln C(r)}{\partial\ln r}$ 进行近似得到相关维数。对应的计算公式如下:

$$\hat{CD}_{\mathrm{corr}}=\frac{\ln C(r_2)-\ln C(r_1)}{\ln r_2-\ln r_1}$$

这样基于信息准则的样本信息维数在用于降维的非线性映射建立前便可以建立,用于信息量的测量与维数的估计。

1.4 流形算法结构优化策略

由于待选的输入变量维数较大,这种高维数据

集合中并非所有变量都是必须的。为了得到高精度和准确的预测模型,同时避免维数灾难[16-17],利用流形算法优化风电场数据结构需要注意:1)风电场数据具有时间顺序同时兼具特定的几何机构。为了优化风电场数据结构,应用流形算法消除数据的冗余信息有利于降低重复信息的输入,造成模型低推广能力的可能性。同时根据相关性分析获悉,不同高度的具有相同属性的变量,比如不同高度的风速之间,抑或标准差之间的相关性极高,这说明此类数据间存在冗余信息的可能性较大。为了保持数据特定的几何结构,利用流形算法对这些数据变量进行结构优化。2)由于部分数据间具有极高的相关性,这容易导致数据间存在相互重叠的现象,如果邻域等设计不当,容易造成邻近图中出现短路现象。根据在数据低密度区域其最短路径相对较易寻找的原理,利用本文提出的邻域设计策略和相关维度估计,采用合理的邻域与本质维度估计避免数据计算出现大面积衰退的现象的同时,保证了数据固有维数为观察特性所需要的参数最小化。而且此时经过数据结构优化后的数据与原始数据存在相对应的内在关系或者映射。本文对于原有高维数据进行降维后再按照之前的结构设计进行新的预测。流形算法具体流程如图 1 所示。

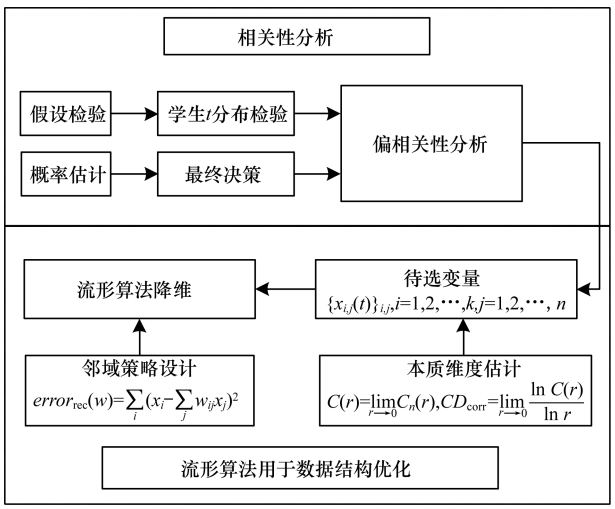

图 1　流形算法结构优化策略流程

本文主要解决了如下 2 个问题:1)模型结构设计。本文提出了一种合理的模型结构设计方法,克服了传统方法将风电场数据处理作为静态模型处理的弱点,并充分反映样本信息在不同季节与不同地区的风电输出动态特性,从而有效地提高模型的计算效率和预测精度。2)推广能力的提高。本文利用流形学习算法对样本进行特征提取,降低了数据的维数,并提出一种稀疏算法来选择合适的邻域数据以提高流形学习算法的计算效率。利用不同信息准则用于样本的本质维数估计来设计合理的模型结构。

2　实验数据与指标

2.1　风电场数据

本文将根据华东某风电场的数据进行变量分析。该风电场采用直径 77 m 的弱风型 FD-77 测风塔进行采集,采样频率是每点 5 min,数据共含有 18 个变量,即 2011 年 6 月 1 日—2013 年 4 月19 日的 183 212 × 18 个数据(部分日期没有相应数据),其中持续超过 24 个点(2 h)为 0 的组数有135 个,最长有超过 196 个点全部为 0,根据与电力科学技术研究院南京分院的工程师们的研讨,如果 FD77 停止运作2 h,那么此时数据无任何贡献率,应当予以删除。测风塔观测点分别为 10 m、50 m、70 m。变量主要有 10 m、50 m、70 m 的平均风速、风向、5 min 内数据采集的标准差、实时风速、实时风速的风向以及 10 m 的温度、湿度、气压。不失一般性,将数据按照高度给出如下的 4 个分组并将数据变量分别编号为1 ~ 18。

1)10 m 分组 1:x1 AWS;x2 AWD;x3 SSD;x4 RTWS;x5 RTWD;

2)50 m 分组 2:x6 AWS;x7 AWD;x8 SSD;x9 RTWS;x10 RTWD;

3)70 m 分组 3:x11 AWS;x12 AWD;x13 SSD;x14 RTWS;x15 RTWD;

4)其他分组 4:x16 10 m 温度;x17 10 m 相对湿度;x18 气压。其中,AWS 为平均风速;AWD 为平均风速对应的风向;SSD 为样本标准差;RTWS 为实时风速;RTWD 为实时风速对应的风向。

2.2　模型评价指标

文中采用如下的 3 种评价指标进行模型性能的评价,分别是均方根误差(Root Mean Square Error,RMSE),平均绝对误差(Mean Absolute Error,MAE)和相对绝对误差(Relative Mean Absolute Error,RMSE):

$$R_{\mathrm{RMSE}}=\frac{1}{\sqrt{n}}\times\sqrt{\sum_{t=1}^{n}(y_t-\hat{y}_t)^2}$$

$$M_{\mathrm{MAE}}=\frac{1}{n}\times\sum_{t=1}^{n}|y_t-\hat{y}_t|$$

$$R_{\mathrm{MAE}}=\frac{1}{\sum_{t=1}^{n}|y_t|}\times\sum_{t=1}^{n}|y_t-\hat{y}_t|$$

其中,RMSE 反映了预测样本与实际样本相差的离散或者偏差程度,MAE 反映了预测样本与真实样本的绝对偏差程度,RMAE 反映的是预测样本与真实样本的绝对偏差与真实样本的比例,其通常可以用于预测精度的测量。

3　实验结果与分析

3.1　推广能力分析

流形降维算法对时间序列进行降维时需要保留原始的时间结构。而且内在维数和本质维数的估计

是非常重要的,因为其决定了时间序列在降维后的信息量问题。本文对于时间序列的降维主要基于信息准则进行。相关计算结果如表 2 所示。

表 2 相关维数估计

数据长度	CorrD	NearND	MLE	GMST
1 000	1	1	1	2
6 000	1	1	1	2
9 000	1	1	1	2

其中,CorrD 是信息准则维数估计中的方法;NearND(Nearest Neighbor Dimension)是指最近邻维数;MLE(Maximum Likelihood Estimator)是最大似然估计方法;GMST(Geodesic Minimum Spanning Tree)是指测地最小生成树方法。如果按照设计的邻域策略选择方法,数据长度为 1 000 时采用 Isomap 算法的耗用时间大约为 0.740 8 s,而经典的方法大约需要 1.123 6 s,这说明策略设计是有效的。如果数据长度太大,可以采用分布式算法,用以提高计算效率。第一子集 10 d(2 880 ×4 个样点)内的部分数据降维效果如图 2、图 3 所示。

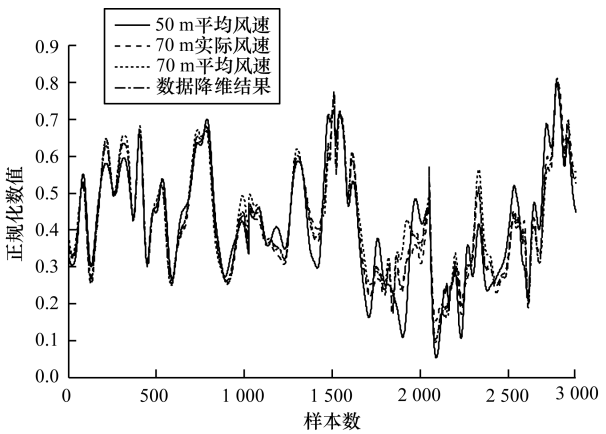

图 2 LPP 平均风速数据降维效果

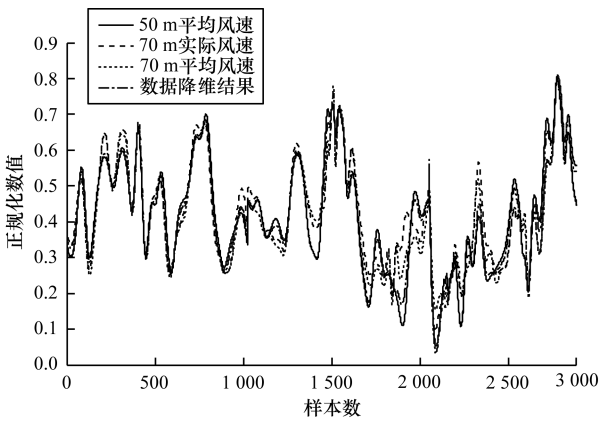

图 3 Isomap 平均风速数据降维效果

本节将对上述理论策略设计进行数据维数降低(DR)的实验验证。不失一般性地,2 种代表性的无监督学习算法 LPP(局部,线性)和 Isomap(全局,非线性)方法用于本文的高维数据降维。实际上,针对时间序列的维数降低,可以理解为在高维曲面流形上的维数映射使得维数降低的特征提取过程,不过整体数据的长度没有变化。而且合理的结构设计可以反映原始样点的走势,这说明数据降维后在邻近流形的局部或全局特征的结构上是不变的。

3.2 性能比较

本节结合上述理论分析的真实数据进行算例分析。采用 70 m 平均风速进行输出,选取样本中 60% 的数据作为训练学习集合,20% 的数据作为验证集合用于测试目前的学习性能及其相关误差,其余的 20% 作为测试集合用以测试得到的模型结构。同时为了对比性能,将对传统无任何结构设计的方法进行比较,如表 3 所示。需要说明的是,由于流形算法对于原始数据的结构有一定的假设和要求,当原始数据出现相互重叠的现象时,流形算法对于数据的计算趋向于大面积的衰退,使得数据在流形嵌入时得到了人们并不期望的放缩现象,导致降维失败。由于重新组合后的数据存在这种重叠现象,因此本文对于原有高维数据进行降维后再按照之前的结构设计进行新的预测。

表 3 各算法性能比较

季节	性能指标	TAR	MSS	MSS-LPP	MSS-ISO
春季	RMSE	1.295 3	1.584 8	1.181 5	1.132 2
	MAE	1.401 2	0.902 2	0.751 6	1.025 4
	RMAE	0.172 6	0.121 9	0.121 3	0.138 9
	ET/s	698.68	269.09	166.63	284.86
夏季	RMSE	1.730 3	1.395 1	1.711 6	1.872 6
	MAE	0.362 8	0.901 6	1.485 5	1.631 7
	RMAE	0.160 6	0.121 8	0.110 3	0.118 6
	ET/s	768.99	269.09	182.21	294.60
秋季	RMSE	1.279 4	1.034 3	1.190 3	1.191 6
	MAE	1.020 2	0.771 6	0.792 3	0.746 3
	RMAE	0.167 6	0.122 4	0.128 5	0.136 0
	ET/s	738.39	232.17	122.23	294.60
冬季	RMSE	1.125 6	1.930 6	1.682 0	1.787 0
	MAE	0.850 5	0.710 6	0.760 2	0.960 2
	RMAE	0.151 3	0.121 9	0.133 8	0.124 1
	ET/s	887.36	223.99	122.93	259.15

其中,TRA 为采用所有输入变量的传统预测方法;MSS 为采用本文设计的模型结构选择方法;MSS-LPP,MSS-ISO 分别是采用流形算法 LPP,Isomap 与模型结构设计 MSS 结合的算法;ET 为以秒计算的耗用时间。表 3 展示的是 12 步(1 h)预测结果,相关变量在输入前进行了归一化,用于消除局部数值过大等因素以免影响计算性能。由于有 9(春夏秋季子集)×8(阶次组合方式)×3(预测方法,MSS,MSS-LPP,MSS-ISO)+3(冬季子集)×4(阶次组合方式)×3(预测方法)=252 种结果需要显示,为简便起见,这里仅显示关于第一子集的图形结果,如图 4 所示。

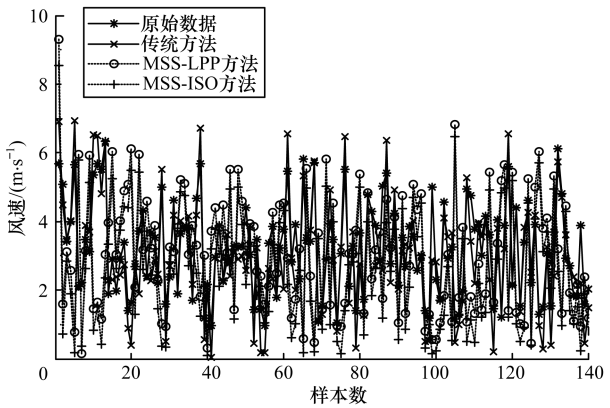

图 4　各方法结果比较

实验的对比结果分别从数据趋势、结果精度和算法时间上来进行分析:

1)数据趋势。通过对比实验结果可以看出,Isomap 和 LPP 与原始数据集合的趋势基本一致。这是由于 Isomap 算法在处理非线性的流形等在内的高维数据可以进行全局优化,而且无论输入空间是否高度折叠、扭曲,其仍能够全局优化低维的欧式距离来表示,并能保证其渐进地恢复到真实的维度,而 LPP 算法对于高维流形数据的局部线性投影能反映一定的数据结构。

2)结果精度。对比统计结果,MSS 比 TRA 的整体预测结果有所提升。LPP 在春季和夏季的提升明显,而秋季和冬季在整体表现上部分有所降低,Isomap 算法在夏季的表现较 MSS 有所提高,其他季节有所降低,精度提升不明显。这可能是由于高维流形存在重叠,LPP 的局部投影效果没有反映数据的本质结构,而 Isomap 由于数据的类间差别较小,导致在多维流形上重合部分较多,使得降维后的数据结构在整体上体现不明显。

3)算法时间。从计算效率上来说,LPP 对于速度的提升非常明显,而非线性全局算法 Isomap 的计算效率较低。这是因为从高维流形直接局部投影在选定合适参数后可以直接映射降维,其速度快而且适合在线学习。虽然在理论上类别差别小的单一数据流形,在 Isomap 算法中采用测地距离来表征整个全局的几何特性时,仍然可以用邻近图中的欧式距离来替代测地距离,能较好地体现最优的路径,但是从测试算例上,由于其对数据结构有一定的依赖性,在非线性流形的几何结构计算时,随着邻近节点增加时,要得到接近于流形上测地距离的点与点之间的距离,需要耗费较多的时间。

综上所述,结合误差性能指标和算例分析,可以得知数据分布类型的考察与处理、模型结构的设计与优化对于时间序列的建模与优化是有效的。

4　结束语

可靠准确的风速预测是风电并网稳定运行的重要条件,并且短期风速预测是电力系统操作中电力调度和系统电力分配中的必要步骤。本文主要对短期风电场时间序列的建模与预测进行了模型结构的设计、理论分析与结构验证。先后对模型输入变量选择 IVS、阶次估计、参数优化、计算效率和模型结构分析设计等进行理论分析与实验设计,并且根据华东某风电场的真实数据进行了实例验证。实验结果表明,本文提出的策略能够提高模型计算效率,降低样本复杂度和提高模型的推广能力。因此,本文的研究对于模型的理论研究和工程应用都有较好的应用价值。

参考文献

[1] 韩　璞,王东风,王富强. 基于误差补偿的风速时间序列多步预测[J]. 计算机仿真,2014,31(2):206-209.
[2] 刘　琳. 新能源风电发展预测与评价模型研究[D]. 保定:华北电力大学,2013.
[3] 罗　欢,陈民铀,程庭莉. 含风力发电的配电网自适应分时段无功优化[J]. 电网技术,2014,38(8):2207-2212.
[4] 王晓兰,李　辉. 基于 EMD 与 LS-SVM 的风电场短期风速预测[J]. 计算机工程与设计,2010,31(10):2303-2307.
[5] BREIMAN L. Random Forests[J]. Machine Learning,2001,45(1):5-32.
[6] OTHERS V J. Dimensionality Reduction for Visual Exploration of Similarity Structures[D]. Helsinki,Finland:Helsinki University of Technology,2007.
[7] ERNST B. Wind Power Prediction[M]. Chichester,UK:John Wiley and Sons,Ltd.,2012.
[8] 张振华,马　超,徐瑾辉,等. EMD 与 NARX 神经网络的风电场总功率组合预测[J]. 计算机工程与应用,2016,52(12):265-270.
[9] 王桂芝,宋迎曦,来　鹏,等. 超高维数据降维与 Logistic 广义线性拟合分析[J]. 统计与决策,2016(7):38-41.
[10] 沈　亮,许青松,曹东升,等. 基于 Markov 性的半监督流行学习算法研究[J]. 中国科学:数学,2015,45(5):703-712.
[11] 戴良斌. 基于统计流行降维的通用性隐藏检测分析[D]. 北京:北京工业大学,2014.
[12] 李　昕,钱　旭,王自强. 用于文档聚类的间隔流形学习算法研究[J]. 计算机工程,2010,36(15):40-42.
[13] WEI H,SHAO H,DENG X. Using a Model Structure Selection Technique to Forecast Short-term Wind Speed for a Wind Power Plant in North China[J]. Journal of Energy Engineering,2016,142(1):5005-5016.
[14] SHI L,GU J. A Fast Manifold Learning Algorithm[J]. Information Technology Journal,2012,11(3):380-383.
[15] BURGES C. Dimension Reduction:A Guided Tour[J]. Machine Learning,2009,2(4):275-365.
[16] 邵　超,万春红,赵静玉. 流形学习算法中邻域大小参数的递增式选取[J]. 计算机工程,2014,40(8):194-200.
[17] CHANG H, YEUNG D. Robust Locally Linear Embedding[J]. Pattern Recognition. 2006, 39(6):1053-1065.

编辑　索书志