相似性算法测试总结

一、 算法说明

由于风速的变化具有明显的季节特性，而且每天的风速变化随着日出日落的周期变化也存在一定的规律，相应的实测数据也随着风速的变化以及运维条件的变化就有一定的相似性，为了充分利用历史风速数据所蕴含的信息，对实测数据以及NWP输入数据的分析，分析预测变量和实测功率之间的相似性，相似性算法尝试从实测数据、NWP数据之间的相似性出发，用最相似的数据进行训练，选择最相似的数据进行预测。

本测试共涉及SVM、K-means聚类、KNN分类、ESN(Echo State Network，ESN)网络等四种算法，前三种算法相对比较常见，本次仅对ESN网络进行说明。

**ESN网络**

ESN网络是一种新型的递归神经网络，在模型构建和学习算法上，与传统的RNN存在较大区别，ESN网络的核心结构是一个随机生成、且保持不变的储备池，输出权重是唯一需要调整的部分，而且简单的线性回归即可完成网络的训练。

回声状态网络由输入层，中间层和输出层构成，各层间通过权值连接，结构如图所示。网络输入层和输出层间的部分就称为储备池，是一种大且稀疏的递归结构，其内部节点状态为输入信号的高维显现，本质上表征了输入信号在高维空间中的线性特性。回声状态网络的基本方程可以写作:



其中*x*(*k*)、*u*(*k*)和*y*(*k*)分别代表*k*时刻ESN的状态变量、输入变量和输出变量，*Wx*和*Win*分别为贮备池内部的连接矩阵和外部输入与储备池神经元的连接矩阵，***w***为储备池和网络输出的连接权值。sig(∙)为双曲正切Sigmoid函数。通常状态变量*X*的维数较高(一般取100~1000之间)，权矩阵*Wx*通常保持1%~5%的稀疏连接，*Wx*和*Win*一经初始化后保持不变。*Wx*的谱半径(矩阵所有特征值模的最大值)一般保持小于1，称这个高维的状态空间为“储备池空间”。

回声状态网络的训练目的是要确定输出权值***w***。使用回声状态网络进行动态系统辨识时，有两个基本步骤：首先初始化储备池，然后由给定的教师信号确定输出权值。由于储备池中的状态变量与目标输出是线性关系，因此回声状态网络的训练过程比较简单，而且解的过程不会出现传统神经网络的多个局部最小、收敛速度慢的缺点。



图1. ESN网络结构

回声状态网络输出权值确定可以通过最小化以下的指标函数来实现



其中*A=[****x****(Init),* ***x****(Init+1), …,* ***x****(Trn)]T*，是对*W*的估计，*yd=[yd(Init), yd(Init+1), …, yd(Trn)]T*，*Init*一般不从1开始，以去掉初始暂态的影响。*Trn*是训练样本最后一个时间点，所以有效的训练样本个数为*Tn*=*Trn*−*Init*+1。

储备池的选择对建模质量的影响较大，主要的待选参数包括储备池权值*Wx*谱半径、输入权值幅值、储备池大小和储备池稀疏度。这些参数的选取可以在一个较大的范围内，比如当储备池的谱半径在0.85可以得到较好的性能时，那么取其谱半径为0.8和0.9仍然可以获得较好的结果。储备池的大小可以有几十甚至几百的变化幅度而不引起辨识效果的恶化，比如大小取200到1000之间。储备池连接矩阵的稀疏度通常保持1%到10%连接。这些参数的选取可以通过交叉检验等方法获取较为精确的估计。

二、 算法流程

为估算不同样本之间的相似性度量(SimilarityMeasurement)，这时通常采用的方法就是计算样本间的“距离”(Distance)，常用的相似性指标有欧式距离、马氏距离、夹角余弦等。本次测试采用欧式距离来评价样本间的相似性。

本次测试主要分析每日实际功率、实际风速、预测变量的序列相似性。每日各变量序列共有96个点，日相似性指标定义为

*V = [Vmin,Vmax,Vmedian,Vstd]*

本次测试分别对单一SVM、单一ESN网络、基于Similarity的SVM、基于Similarity的ESN、基于K-means聚类的SVM、基于K-means聚类的ESN六种方法进行对比测试。

**1、Similarity相似性分析功率预测**

算法通过对历史实测数据进行相似性指标计算，利用K-means聚类分析，将相似的样本进行训练，按照类别分别仅网络训练，并利用对应的NWP数据进行K-NN分类训练，预测时根据NWP数据进行K-NN分类，按照相应的类别进行预测。算法流程如下：

1. 划分训练样本和测试样本集；
2. 训练样本计算实测数据相似性变量S= *[Pmin,Pmax,Pmedian,Pstd, Corr(P,Speed), Corr(P,time\_num)]，P、Speed、time\_num*分别代表实际功率、预测风速、时间变量，*Corr*表示两者之间相关性；
3. 利用K-means对日相似性变量S进行聚类分析；
4. 选择聚类后数据对应的NWP变量训练数据Snwp\_train =*[Umin,Umax,Umedian,Ustd, Vmin, Vmax, Vmedian, Vstd, Dmin, Dmax, Dmedian, Dstd]* ，*U、V、D*分别代表*U*风速、*V*风速、预测风向；
5. 测试数据计算NWP变量训练数据Snwp；对测试样本Snwp进行K-NN分类，训练数据Snwp\_train对应的类别为步骤（3）中的类别；
6. 每一类别测试数据分别利用相应的模型进行预测。

**2、K-means聚类分析功率预测**

算法通过对历史NWP数据进行K-means聚类分析，将相似的样本进行训练，按类别训练不同的预测模型，预测时首先判断待预测日的类别，然后选择相应的模型进行预测。回归网络采用SVM和ESN网络分别测试。算法流程如下：

1. 划分训练样本和测试样本集；
2. 训练样本计算NWP日相似性变量S= *[Umin,Umax,Umedian,Ustd, Vmin,Vmax,Vmedian,Vstd, Speedmin,Speedmax,Speedmedian,Speedstd, Tmin,Tmax,Tmedian,Tstd]，U、V、Speed、T*分别代表*U*风速、*V*风速、预测风速、预测温度；
3. 利用K-means对日相似性变量S仅聚类分析；
4. 分别按照各类别数据进行SVM或ESN训练，并保存相应参数；
5. 测试样本按照聚类结果通过距离测度进行类别划分；
6. 每一类别测试数据分别利用相应的模型进行预测。

三、 测试结果分析

**1、相似性结果**

通过对五凌鄯善风电场的历史数据分别进行日实测数据的相似性K-means聚类分析，并利用NWP数据仅K-NN分类，图2至图4为三个类别每日的实测功率曲线，从图中可以看出，通过对历史实测数据进行聚类，三种类别之间实测数据均有各自的相似性，即同样发电规律的数据聚为一类，图5至图7为通过K-NN分类的结果，可以看出通过K-NN同样将实测功率进行了良好的分类。



图2. K-means 类别1每日实测功率



图3. K-means 类别2每日实测功率



图4. K-means 类别3每日实测功率



图5. K-NN 类别1每日实测功率



图6. K-NN 类别2每日实测功率



图7. K-NN 类别3每日实测功率

**2、功率预测结果对比**

本次测试分别对五凌鄯善风电场、华电草湖风电场、中电投玛依塔斯三期、固原三营一期进行对比分析，对比测试结果如下：

表1.五凌鄯善测试结果对比

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 算法 | SVM | ESN | Kmeans\_ESN | Similarity\_ESN | Kmeans\_SVM | Similarity\_SVM |
| 5月 | 18.56% | 18.34% | 19.97% | 18.91% | 18.23% | 18.50% |
| 6月 | 25.63% | 23.33% | 25.18% | 23.18% | 24.61% | 25.30% |
| 平均 | 22.10% | **20.84%** | 22.57% | **21.05%** | 21.42% | 21.90% |

表2. 华电草湖风电场测试结果对比

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 算法 | SVM | ESN | Kmeans\_ESN | Similarity\_ESN | Kmeans\_SVM | Similarity\_SVM |
| 6月 | 23.89% | 22.15% | 20.27% | 19.72% | 23.19% | 22.88% |
| 7月 | 23.69% | 22.30% | 22.76% | 22.98% | 23.98% | 23.78% |
| 平均 | 23.79% | 22.23% | **21.52%** | **21.35%** | 23.58% | 23.33% |

表3. 中电投玛依塔斯三期测试结果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 算法 | SVM | ESN | Kmeans\_ESN | Similarity\_ESN | Kmeans\_SVM | Similarity\_SVM |
| 8月 | 12.98% | 11.80% | 11.32% | 12.15% | 12.94% | 13.34% |
| 9月 | 19.36% | 18.13% | 17.95% | 17.10% | 19.13% | 18.23% |
| 平均 | 16.17% | **14.96%** | **14.64%** | **14.62%** | 16.03% | 15.79% |

表4. 固原三营测试结果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 算法 | SVM | ESN | Kmeans\_ESN | Similarity\_ESN | Kmeans\_SVM | Similarity\_SVM |
| 6月 | 15.73% | 19.41% | 29.59% | 27.78% | 15.86% | 16.65% |
| 7月 | 19.55% | 19.44% | 19.21% | 21.63% | 19.55% | 20.05% |
| 8月 | 19.14% | 19.37% | 17.76% | 18.32% | 18.19% | 17.92% |
| 9月 | 13.00% | 13.16% | 12.59% | 12.89% | 13.01% | 13.71% |
| 平均 | **16.85%** | 17.85% | 19.79% | 20.15% | **16.65%** | 17.08% |

通过以上测试结果可以看出，对于四个风场的，六种不同的算法在均方根误差上表现了一定的差异性，单一算法ESN算法在五凌鄯善、中电投玛依塔斯、华电草湖三个电场表现了较好的性能，而SVM在固原表现了较好的性能。六种算法中基于ESN网络历史数据相似性的算法综合表现了相对较高的准确率优势。在凌鄯善、中电投玛依塔斯、华电草湖三个电场中相似性算法能够比现有的SVM算法准确率提升1%~2%。

四、 总结

通过对历史实测数据及NWP数据进行相似性分析，选择最相似的数据建立预测模型，采用SVM或ESN网络进行回归预测，对四个电场的测试结果表明，基于相似性算法在部分电场预测准确率提升1%~2%，相对而言基于历史实测数据相似性分析并利用KNN对预测数据分类后的预测模型准确率更高，ESN网络的引入也能比SVM网络在部分电场一定程度上提升预测准确率。

本次测试还不完善，后续需要测试选择不同的实测预测变量来构造日相似性向量后的分析结果，测试夹角余弦、马氏距离等不同相似性测度下的相似性，同时对于ESN网络的输入参量及参数设置需要进一步优化。