

智能制造大数据技术实践 风电功率预测建模方法

一、实验目的

1. 掌握风电场数据清洗和预处理方法；
2. 掌握多维特征的相关性分析方法；
3. 掌握 K-means 聚类方法；
4. 掌握最小二乘法、支持向量回归和 BP 神经网络预测建模方法。

二、实验仪器设备

1. PC
2. Matlab 或 Python

三、实验原理

风电功率预测技术：对未来一段时间内风电场所能输出的功率大小进行预测，以便安排调度计划。

$$P = \frac{1}{2} \rho A v^3 C_p \quad (1)$$

其中， P 是风能转化成的机械能的功率； ρ 是空气密度，受温度、湿度和压强的影响； A 是风轮叶片扫过的面积； v 是风速； C_p 是功率系数，取决于风车的设计和效率。从式(1)中可知，风速与功率之间存在 3 次方的关系。

数据清洗和预处理：由于传感器精度、电磁干扰、信息处理、错误存储或通信故障等，以及目前电力系统对风电的消纳能力有限，风电场强制弃风已成为常态，这就使得原始记录数据中会存在大量的异常数据，用风速-功率 (v - p) 散点图表示时，这些异常数据严重破坏了风速和功率所应有的整体分布规律和对应关系，所以对风电场数据进行数据清洗和预处理是必要过程。

皮尔逊相关性分析：一种广泛使用的统计方法，用于衡量两个变量之间的线性关系强度和方向。它通过计算两个变量的协方差与它们各自标准差的乘积之比来确定相关系数。这个相关系数（通常表示为 r ）的值范围从 -1 到 1，其中 -1 表示完全负相关，1 表示完全正相关，而 0 表示没有线性关系。

Density-Based Spatial Clustering of Applications with Noise (DBSCAN)：是一个比较有代表性的基于密度的聚类算法。它将簇定义为密度相连的点的最大集合，能够把具有足够高密度的区域划分为簇，并可在噪声的空间数据库中发现任意形状的聚类。

BP(Back Propagation)神经网络：BP 神经网络是一种按误差反向传播(简称误差反传)训练的多层前馈网络，其算法称为 BP 算法，它的基本思想是梯度下降法，利用梯度搜索技术，以期使网络的实际输出值和期望输出值的误差均方差为最小。

四、实验任务与步骤

根据如图 4-1 所示框图进行风电功率预测。

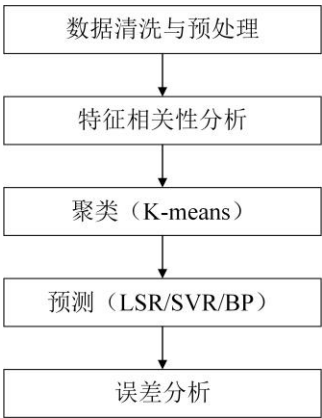


图 4-1 风电功率预测流程框图

4.1 数据清洗与预处理

1. 缺失值清洗：对风电场数据中为 **NAN**（空值）的数据值进行删除，并进行数据填充。
2. 异常值清洗：对数据进行可视化，删除明显的离群点；进行简单的数据分析，判断属性的取值是否超过合理范围。
3. 格式内容清洗：对风电场数据中不该存在的字符，属性与内容不符的格式问题进行清洗。
4. 逻辑错误清洗：首先若数据中的重复数据与实际情况不符，对数据进行去重处理；之后对风电数据中的不合理值进行去除，例如低风速高功率点。
5. 非需求数据清洗：若风电场数据中含有实际操作中不需要的数据，则进行清洗。
6. 数据归一化：消除所有变量量纲和取值范围差异。

4.2 特征相关性分析

1. 计算相关系数：使用皮尔逊相关系数公式计算每对特征变量之间的相关系数，该系数范围从-1 到 1，表示变量间的线性关系强度和方向。
2. 构建相关性矩阵：将所有特征变量两两之间的相关系数整合成一个相关性矩阵，便于直观地查看和分析特征间的相关性。
3. 可视化：通过绘制散点图矩阵或热力图等可视化工具，直观展示特征间的相关性。
4. 特征筛选：筛选与功率相关性较高的特征作为聚类与预测模型的输入。

4.3 数据聚类

K-means 算法流程：

- （1）从集合 **D** 中随机取 k 个元素，作为 k 个簇的各自的中心。
- （2）分别计算剩下的元素到 k 个簇中心的相异度，将这些元素分别划归到相异度最低的簇。
- （3）根据聚类结果，重新计算 k 个簇各自的中心，计算方法是取簇中所有元素各自维度的算术平均数。
- （4）将集合 **D** 中全部元素按照新的中心重新聚类。
- （5）重复第 4 步，直到聚类结果不再变化。

4.4 预测建模（BP 神经网络）

BP 神经网络设计问题：

(1) 网络层数：具有至少一个 S 型隐含层加上一个线性输入层的网络，能够逼近任何有理函数。增加层数可以进一步的降低误差，提高精度，但同时也使网络复杂化。

(2) 隐含层神经元：网络训练精度的提高，可以通过采用一个隐含层，而增加其神经元数的方法来获得。

(3) 初始权值选取：BP 神经网络是非线性优化算法，初始值设置不当，可能陷入局部极小。

(4) 学习速率：决定每一次循环训练中所产生的权值变化量。为了减少寻找学习速率的训练次数以及训练时间，比较合适的方法是采用变化的自适应学习速率，使网络的训练在不同的阶段设置不同大小的学习速率。

4.5 误差分析

分析实测功率与预测功率之间的误差特性，主要为统计关系。采用平均绝对误差（MAE）、均方根误差（RMSE）和判定系数（ R^2 ）来表示模型预测效果的评价指标。表达式如下：

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{\sum_{i=1}^n (\bar{y}_i - y_i)^2} \quad (4)$$

其中 \bar{y}_i , y_i 和 \hat{y}_i 分别是真实值的平均值，真实值和预测值。

五、实验注意事项

1. 完成风电功率预测的每个步骤时，注意重新保存数据；
2. 保证代码美观并及时注释，保持良好的代码编写习惯。

六、实验报告

1. 根据图 4-1 的步骤来进行风电功率预测，根据风电功率数据特征，选取合适的数据预处理方法，完成预测建模，实现风电功率预测，对预测结果进行误差分析。
2. 绘制每个模块的流程图。
3. 完成各个步骤时，对数据进行可视化显示。

七、代码参考

Matlab

DBSCAN 聚类核心代码：

```
[IDX, isnoise]=DBSCAN(samples, epsilon, MinPts);
```

K-means 聚类核心代码：

```
[cluster_label, sse]=kmeans(samples, K);
```

BP 神经网络核心代码:

```
[error, BPoutput]=BP(train_input, train_output, test_input, test_output);
```

Python

DBSCAN 聚类核心代码:

```
from sklearn.cluster import DBSCAN  
y_pred = DBSCAN(eps=3, min_samples=2).fit_predict(cluster_data)
```

K-means 聚类核心代码:

```
from sklearn.cluster import Kmeans  
y_pred = Kmeans(n_clusters=2).fit_predict(cluster_data)
```

BP 神经网络核心代码:

```
from sklearn.neural_network import MLPClassifier  
BP = MLPClassifier(solver='sgd', activation='relu', max_iter=500, alpha=1e-3, hidden_layer_sizes =  
(32,32), random_state=1)  
BP.fit(train_samples, train_labels)  
predict_labels = BP.predict(train_samples)
```