**FDA 期末專題(B)報告 主題:Wind Power Forecast**

**組別:25 成員:蔡順先、余俊輝**

1. Introduction

近年來可再生能源中，成長幅度最高的能源之一為風力發電，但由於風力的多變性，導致發電量的不穩定。在電力系統中，由於需要維持發電量和耗電量的平衡，風力發電的不穩定意味著需要從其他電力來源補充。因此提前得知風力發電的變化就能先規劃與啟用其他電力來源，做到更有效率的能源調配。

1. Problem Definition

透過SCADA系統每十分鐘記錄一次從風力發電農場收集到的發電機資訊以及天氣資訊，並附上了各發電機的相對位置。藉由過去收集到的資訊，預測出未來48小時各發電機的發電量。

1. Dataset Description

資料集中收集了134台發電機在245天中的相關資料，下表為資料集欄位說明。

|  |  |
| --- | --- |
| Column Name | Specification |
| TurbID | Wind turbine ID |
| Day | Day of the record |
| Tmstamp | Created time of the record |
| Wspd(m/s) | The angle between the wind direction and the position of turbine nacelle |
| Etmp(℃) | Temperature of the surounding environment |
| Itmp (℃) | Temperature inside the turbine nacelle |
| Ndir (°) | Nacelle direction |
| Pab1 (°) | Pitch angle of blade 1 |
| Pab2 (°) | Pitch angle of blade 2 |
| Pab3 (°) | Pitch angle of blade 3 |
| Prtv (kW) | Reactive power |
| Patv (kW) | Active power (target variable) |

1. Data Analyze

圖一為發電機的缺失值數量，X軸為發電機ID，Y軸為缺失值資料數量，從圖可以看出每個發電機的平均缺失值是在300左右。

圖二為每天的缺失值數量，X軸為天數，Y軸為缺失值資料數量。發現到每天缺失值的數量是很少的，只有在某一天會有很誇張的缺失值數量，因為數據跟時間有高度的關係，所以可採用該時間附近的非缺失值數據去填補。

Shape

Description automatically generatedHistogram

Description automatically generated with medium confidence

圖一、發電機缺失值 圖二、每天缺失值

A screenshot of a computer

Description automatically generated with medium confidence分析各特征與發電量的關係，發現說跟發電量有高度正相關和負相關的特征有，風速(Wspd)、葉片角度(Pab1~3)、發電機的朝向角度(Ndir)。

圖三、各特徵關係圖

觀察各數據的分佈圖，結合各特征關係圖，其中高度相關性的特征有Wspd， Pab 1 – 3， Ndir的數據分佈量相對其他特征來說分佈比較廣。因此將嘗試著重針對這些特徵進行處理。

Chart, box and whisker chart

Description automatically generated

圖四、各特徵分布圖

1. Data Preprocess

**NaN值處理**

資料集中含NaN值的資料數約有5萬筆，且分布廣泛，較少連續出現的情況。透過呼叫pandas.Dataframe.fillna(method=’bfill’)，以下一筆非NaN值的資料填補。

**異常值處理**

SCADA系統在收集資料的過程中可能會出現錯誤，導致異常值的產生。當資料滿足某些條件時，此時的Patv值是不可確定的。為了訓練資料的完整性與連續性，需針對這些資料進行處理。

**異常值處理 – Pab**

Pab>89時代表錯誤發生。由於Pab的合理範圍應在0~89，因此以比89多出的值作為取代。

If Pab>89: Pab -= 89。

**異常值處理 – Ndir**

Ndir>720 或 <-720時代表錯誤發生。由於Ndir的合理範圍應在-720~720，因此超出範圍的值應從另一個邊界開始計算。Ex: 750 -> -690。

If abs(Ndir)>720:

Ndir = Ndir – math.copysign(1, Ndir) \* 720 \* 2

**異常值處理 – Wdir**

Wdir>180 或 <-180時代表錯誤發生。由於Wdir的合理範圍應在-180~180，因此超出範圍的值應從另一個邊界開始計算。Ex: 200 -> -160。

If abs(Wdir)>180:

Wdir = Wdir – math.copysign(1, Wdir) \* 180 \* 2

**異常值處理 – Patv**

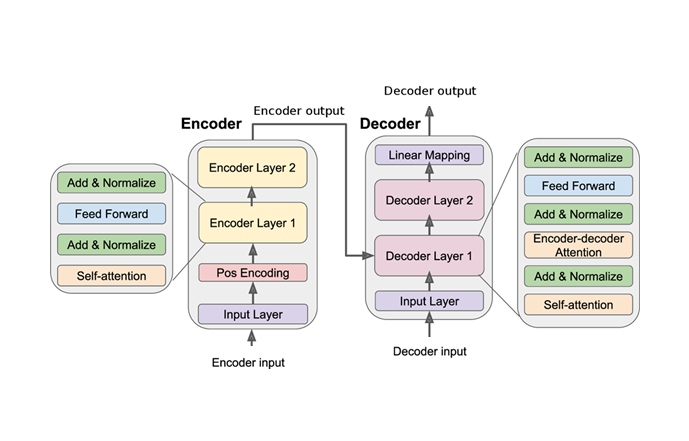
當Patv<=0且Wspd>2.5時代表錯誤發生，此時有一定強度的風力，但發電量卻不是正值，因此需針對Patv值做處理。另外在上述關於滿足Pab, Ndir, Wdir異常條件的資料中，其Patv值亦須處理。透過收集這些資料的index，找出兩端最近的正常值做內插法去取代原本的異常值。

**Feature Selection**

捨棄TurbID, Day, Tmstamp欄位，加入Normalize過的發電機空間位置loc\_x, loc\_y作為兩個新欄位。

**Normalize**

依不同台發電機分割資料，針對除空間位置之外的feature做normalize。

1. Model Architecture

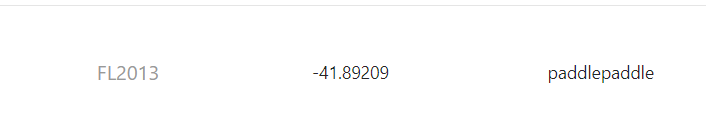
模型是以Transformer為主架構。Encoder Input為Day1~4的資料，經過Input Layer (Linear)增加feature數。接著做Positional encoding，將資料相對位置加入feature。再經過兩層Encoder Layer得到Encoder Output。

Decoder Input 為Day4~6的資料，一樣經過Input Layer(Linear)增加feature數，接著和Encoder Output一起輸入到兩層Decoder Layer，最後再做一層Linear Mapping減少feature數到一項作為Patv。截取最後兩天資料的長度作為對Day7~8的發電量預測。

1. Train

從各發電機的資料中取150天的資料作為訓練資料集，15天的資料作為驗證資料集，並訓練10個epoch。Loss Function選擇MSE Loss，Optimizer選擇Adam，learning rate設1e-4。

1. Result



訓練出來的模型透過官方的評分系統得分為-41.89，而目前Leaderboard的前幾名約為-40.79。

1. Conclusion

原本的資料集中因為資訊系統有時會出錯，導致收集到的資訊有誤，需要針對異常值進行處理。針對與角度相關的異常值，將其換算成正常範圍內的數值。而發電量的異常值則以內插法進行取代。模型部份以Transformer為架構，以前六天的資料作為模型輸入，預測出後兩天的發電量。最後在官方評測系統中得分為-41.89。若是要再增加得分，可往資料前處理的方向進行優化，透過資料分析後對其進行一些資料轉換。

[Code連結](https://github.com/FL2013/KDD2022-WPF)