**Context**

Optima. Inc is a leading weather forecasting company. They acquire the weather data from sensors installed in the supported regions and store sensor data in the repository for analysis. The company needs to analyze the anomalies in the acquired weather for further forecasting related activities.

**Technical requirements**

* Show case at least 2 approaches (Conventional ML/Statistical and Deep learning) to detect anomalies in the weather data.
* Model should identify different patterns of the anomalies found in the weather data.

**Response expectation**

* Python code or Jupyter notebook containing following sections with appropriate comments & assumption added.
  + Exploratory Data Analysis.
  + Algo-1: Conventional ML/Statistical based approach for anomalies detection.
  + Algo-2: Deep learning-based approach for anomalies detection.
  + Graphs plots with clear showcase of anomalies in data.

**Note**:

Please do pip freeze of your environment and share it as text file. The same will help us to reproduce the development environment.

**Data Set**

The data set is separately provided in the csv file as data\_sample.csv.

**Anamoly Detection from Sensor Data**

Aniket Gurav

This report explains how anomolies can be detected from sensor input data. The sensor measures wind speed and wind direction at every 10 seconds.

**Data-set description**

The data-set contains 106749 readings of sensors out of which only 3 values are none. Considering very small proportion of missing values, rows containing those values are removed.

Assumption: It is assumed that wind direction is actually a angle in degree which can vary from 0 to 360.

As these are sensor readings taken at interval 10 seconds I am also treating these observations as a time-series data.

In this study I am assuming, **anomaly detection** is referred to the identification of items or events that do not conform to an expected pattern or to other items present in a dataset. The data points which have anomalies are outliers so I have mainly done outlier detection on given data.

This report is divided into following parts

1) Data exploration and checking for visual patterns

2) Anomalies detection using ML/ Statistical methods and Using Deep Learning DL Based methods

3) Analysis of Different patterns of Anomalies found in Data

1) Data exploration and checking for visual patterns:

After removing none values from data it is divided into train and test set for using ML and DL based algorithms but in 1st section exploration is done on the entire data-set.

Following is statistical properties of the entire data-set

stat properties of data :

wind\_speed wind\_direction

count 106746.000000 106746.000000

mean 4.659675 198.913124

std 2.164393 88.798419

min 0.000000 0.000000

25% 3.220000 143.700000

50% 4.680000 204.900000

75% 6.050000 261.800000

max 26.090000 359.900000

**Table 1**

For related code details please check statAndPreprocessing.ipnb cell 8

The table on shows wind speed vary from 0 to 26 units and wind direction from 0 to 360 degree.

TimeUTC value present shows that data is present from 2014/12/19 to

2016/12/29 it is for 742 unique days.

Monthly average of wind speed is as follow:

**Table 2 a Table 2 b**



For detail please refer code in 1 .ipnb cell 29

Monthly brake up of wind direction



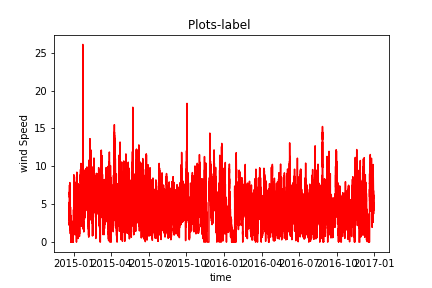
**Table 3a Table 3b**

For detail please refer code in 1.ipnb cell 30

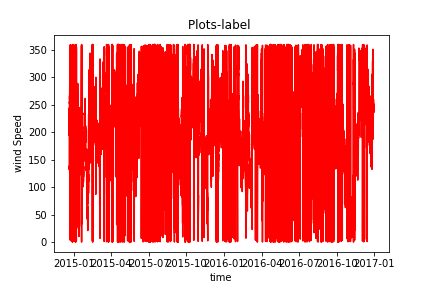
This gives insight that most of time wind direction is between angle 160 to 240.

Now observe the wind speed vs time

**figure 1 Wind Speed VS Time**



From figure 1 it is clear that wind speed data is highly noisy and there are multiple outliers present.

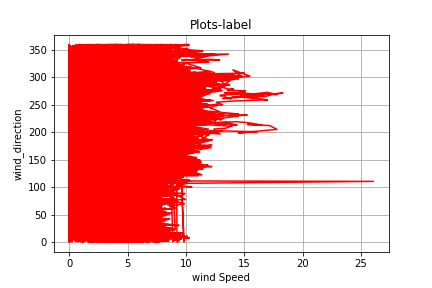
****

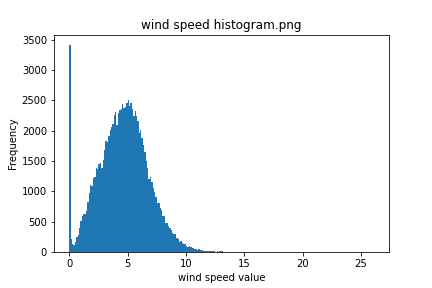
**Figure 2 Wind Direction VS time**

From figure 2 it is clear that wind direction data is highly noisy and there are multiple outliers present.

Figure 3 compares wind speed vs wind direction.

**Figure 3 Wind direction VS wind Velocity**



****

**Figure 4 Wind Speed Histogram**

Figure 4 shows histogram distribution of wind speed, from it is clear that wind speed is around 5 units.

**Figure 5 Wind Direction Histogram**

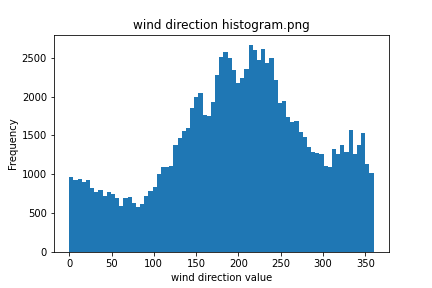


Table 2 and Table 3 gives us insight that wind speed is around 4.6 and angle is around 200.

2) **Anomalies detection using ML/ Statistical methods**

a) Isolation Forest

b) KNN

c) clustering

d) abod

e) Histogram

f) lof

g) pca

e) svm

As already described in section 1, the provided data has 2 features wind speed and wind direction and no labels for anomalies are provided. It makes this problem unsupervised learning problem. Considering this all the classifiers mentioned above are unsupervised.

The given data is divided into two parts train (95% in size) and test (5% in size).

Train data contains 101411 rows and test data contains 5338 samples.

Following table shows method detected outliers and name and location of result file.

|  |  |  |  |
| --- | --- | --- | --- |
| classifier | Train outliers | Test outliers | Result files in folder results |
| Isolation Forest | 5072 | 2955 | **outlier\_results\_iforest\_train.csv**  **outlier\_results\_iforest\_test\_predictions.csv** |
| KNN | 5071 | 371 | **outlier\_results\_knn\_train.csv**  **outlier\_results\_knn\_test\_predictions.csv** |
| clustering | 5072 | 367 | **outlier\_results\_cluster\_test\_predictions.csv**  **outlier\_results\_cluster\_train.csv** |
| abod | 0 | 0 | **outlier\_results\_abod\_train.csv**  **outlier\_results\_abod\_test\_predictions.csv** |
| Histogram | 5047 | 168 | **outlier\_results\_histogram\_test\_predictions.csv**  **outlier\_results\_histogram\_train.csv** |
| lof | 5071 | 325 | **outlier\_results\_lof\_train.csv**  **outlier\_results\_lof\_test\_predictions.csv** |
| pca | 5072 | 627 | **outlier\_results\_pca\_train.csv**  **outlier\_results\_pca\_test\_predictions.csv** |
| Svm | 5072 | 573 | **outlier\_results\_svm\_train.csv**  **outlier\_results\_svm\_test\_predictions.csv** |

Above table shows outlier detected through multiple unsupervised techniques.

The required code is present in file mlModels.ipynb

The outlier detected through this method are present in folder “results”. Each csv file contains column Anamoly, the 1 value shows presence of outlier.

The outlier detected by these unsupervised methods are not common, So as a next step I have ensemble results of all these methods.

Ensemble file are present at location “results” folder, it is present in **outlierVoting.csv**

In above outlierVoting.csv, column vote shows from above 8 methods how many of them think given row is outlier.

Results from file outlierVoting.csv

This file contains combined results of above mentioned all 8 methods. The vote count is present in column vote. Where vote ranges from 0 to 7. 7 indicates that out of 8 methods 7 methods mark given sample as a outlier.

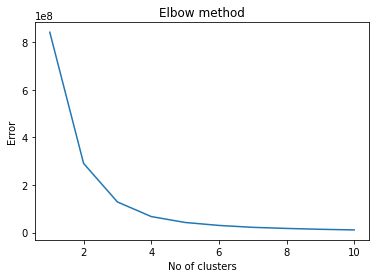
|  |  |
| --- | --- |
| Method count | count |
| 1 | 8262 |
| 2 | 2381 |
| 3 | 1602 |
| 4 | 1838 |
| 5 | 1296 |
| 6 | 503 |
| 7 | 117 |

2) **Anomalies detection using K-means clustering (Statistical and ML mix)**

K-means clustering is unsupervised method of clustering. The code for this method is present in file **anamolyModels\_knn.ipynb**. The result is present at **results/kmeans.csv** file.

This method has detected 1115 outliers from 101409 samples.

Foe kmeans clustering 4 clusters are used. The optimal number of clusters are determined using elbow method.



The diagram shows error vs number of clusters.

Using k-means cluster center for each row is determined. From cluster center, distance between each sample and its cluster is found. Those samples which are 2 standard deviations away from the respective cluster center are marked as a outlier.

3) Anomalies detection using Deep Learning

To detect the anomalies using Deep Learning I have used LSTM based autoencoder.

The autoencoder is unsupervised learning method where it tries to regenerate samples. Here encoder-decoder architecture learns mostly the average distribution of data, Outliers is something which deviates from the distribution of the data. So when autoencoder tries to regenerate that the corresponding loss is high. In this method I have decided threshold on loss and all the samples which has more loss than specific value are marked as a outlier.

Below is a architecture of network

Layer (type) Output Shape Param #

=================================================================

input\_1 (InputLayer) [(None, 1, 2)] 0

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

lstm (LSTM) (None, 1, 16) 1216

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

lstm\_1 (LSTM) (None, 4) 336

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

repeat\_vector (RepeatVector) (None, 1, 4) 0

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

lstm\_2 (LSTM) (None, 1, 4) 144

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

lstm\_3 (LSTM) (None, 1, 16) 1344

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

time\_distributed (TimeDistri (None, 1, 2) 34

The above mentioned architecture is trained for 500 epochs.