August 1st, 2017

Udacity Machine Learning Nanodegree

Igor Oliveira

Capstone Proposal

Machine Learning for Fog Forecasting

Domain Background

This project will attempt to use meteorological observations data to forecast the occurrence of fog, an extremely important low visibility event that have profound impact in air transport operations. This application of Machine Learning has been used before, specially the use of Artificial Neural Networks (ANN) (Refs: [1]-[5]). The existing research in this field provides a good guidance about the type of data to be considering in the training process. It also provides background about the good performance of ANN methods when compared to traditional techniques (Ref: [5]).

As a meteorology researcher, this seemed like a good choice for capstone project in the Machine Learning Nanodegree Course.

Problem Statement

The problem to be solved is the forecast of Fog occurrence with a few hours of lead time (from 3 to 12). Airports and other sites report the occurrence of fog, as well as corresponding atmospheric variables, allowing for the application of machine learning techniques. Since we have the reference use of successful ANN techniques in the prediction of this phenomena, this project will try to reproduce this method and try Deep Neural architectures in search of possible better performance. The binary characteristic of the problem (occurrence/ no occurrence) allow for metrics of verification derived

from confusion matrices. Regression results, for the case of forecasting actual values of ceiling and visibility, can also be verified via metrics like R_squared, Mean Absolute Error or Mean Squared Error. Since the data is available in many sites worldwide, replication is possible. However, since the strong dependence with local information it is possible that model development could vary depending on the site where forecast is being performed.

Datasets and Inputs

There is a public dataset of weather observations provided by several aerodromes (METAR) and other reporting stations (SYNOP) which consist in coded data, captured manually or automatically and containing observed values for several meteorological characteristics such as air temperature, relative humidity, wind speed and direction, atmospheric pressure, and many more. These reports are gathered at hourly time plus at significant weather events (METAR) or 3-hourly (SYNOP)

These datasets can extent for many decades, depending on the history of the aerodrome/station. Many sources can be found online such as the Iowa State University and weather.cod.edu, sometimes already in decoded format. These datasets contain the necessary features necessary for the training, validation and testing according to the references, such as:

- Air Temperature
- Dewpoint Temperature
- Atmospheric Pressure
- Wind Speed & Direction

Classification approaches will be tested using present weather categorical observation (yes/no for the occurrence of fog) which are reported in METAR/SYNOP messages as

labels in the classificators. Regression techniques will be used by using ML to predict the value of visibility or ceiling reported as targets for the algorithms.

Solution Statement

The solution to the fog forecasting problem will be achieved through at least one classification method and one regression method. For classification, Random Forest will be tested using atmospheric parameters observed at 3 to 12 hours of lead time prior to the occurrence of fog and using the occurrence (or absence) of Fog as labels for the classificators. For regression, difference architectures of Neural Networks will be tested, using the value of visibility as targets for training and testing.

Benchmark Model

The references cited in this proposal (Refs [1] to [5]) provide different benchmarks results for fog forecasting in different aerodrome. Costa et al (2006), for example, provides reasonable values of Root Mean Square Error (RMSE) for prediction of visibility values and Marzban et al 2006 provides reference values for confusion-matrix based verification metrics such as probability of detection and false alarm rate. These works use the same type of data considered in this proposal and allow for a reasonable comparison.

Evaluation Metrics

The evaluation metrics that will be considered here can be divided in 2 types. The classification metrics are derived from confusion matrices, a commonly use method for binary classification models. In this work, the matrix will consist of a 2x2 table where

rows represent the observation (or lack thereof) of fog while columns represent the prediction of fog (Table 1)

Table 1 - Confusion Matrix Example

	Predicted: Yes	**Predicted: No**
Observed: Yes	**TP**	**FP**
Observed: No	**FP**	**TN**

Where,

- **TP** = True Positives (Correct Prediction of Fog)
- **TN** = True Negatives (Correct Prediction of non-occurrence of Fog)
- **FP** = False Positive (False prediction of Fog, when no Fog was observed)
- **FN** = False Negative (Prediction of non-occurrence of Fog, but Fog was observed)

This matrix allows for several metrics, such as:

- **Accuracy**: (TP+TN)/(TP+TN+FP+FN)
- **Missclassification Rate**: (FP+FN)/(TP+TN+FP+FN)
- **False Positive Rate**: FP/(TN+FP)

For the regression methods, several metric are also available, such as the RMSE

$$RMSE = \sqrt{\frac{\sum_{k=1}^{N} \left| \frac{a_k - y_k}{a_k} \right|^2}{N}}$$

Where,

a_k = Wanted exit or observation

y_k = Calculated exit

N = Number of forecasts

August 1st, 2017

Project Design

An initial step necessary in this work is to analyze aerodromes amongst the total available globally and find some with a reasonable number of fog observations, in order to derive suitable training sets. The sites where rare fog or low visibility events are observed might not provide a consistent dataset. This will result in the selection of a few sites for forecasting.

After that, we should organize the data in order to match the fog observations and the visibility values with the corresponding meteorological features. After some data cleaning, which includes checks for missing and unreasonable data points, this will generate the proper dataset for training, validation and testing.

Next, it will be time to divide the workflow in a classification and regression efforts. The classification front will use random forests to classify the dates against the binary observation of Fog while the regression front will apply Artificial Neural Networks for prediction of visibility values.

The performance of the machine-learning algorithms applied in each of these fronts will be measured according to the appropriate verification metrics.

References:

1. Pasini, Antonello, Vinicio Pelino, and Sergio Potestà. "A neural network model for visibility nowcasting from surface observations: Results and sensitivity to physical input variables." Journal of Geophysical Research: Atmospheres 106.D14 (2001): 14951-14959.

2. Costa, Saulo B., et al. "Fog forecast for the international airport of Maceió, Brazil using artificial neural network." Proc. 8th ICSHMO, Foz do Iguac u, Brazil (2006): 24-28.

3. Gultepe, I., et al. "Fog Research: A Review of Past Achievements and Future Perspectives." Pure and Applied Geophysics 6.164 (2007): 1121-1159.

4. Bremnes, John Bjørnar, and Silas Chr Michaelides. "Probabilistic visibility forecasting using neural networks." Pure and Applied Geophysics 164.6-7 (2007): 1365-1381.

5. Marzban, Caren, Stephen Leyton, and Brad Colman. "Ceiling and visibility forecasts via neural networks." Weather and forecasting 22.3 (2007): 466-479.