Florida State University

Deep Learning Applied to Automated Flagging of Meteorological and Oceanographical Observation Data

by Randy Bruno-Piverger

A Prospectus of a Thesis Submitted to the Honors College of Florida State University

in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science

In the Department of Computer Science

October 2018

Prospectus

Deep Learning techniques applied to automatic flagging of navigational, meteorological and oceanographical observations from NOAA research vessels.

1. Abstract:
   1. The motivation and research goal:
      1. The Shipboard Automated Meteorological and Oceanographic System (SAMOS) initiative improves the quality of navigational, meteorological and oceanographic observations collected on research vessels through the use of a number of quality control procedures. One such procedure is a visual quality control method which involves individual analysts manually reviewing large amounts of raw observational data. This method of quality control requires a considerable amount of time on the part of the analysts. Visual quality control is nonetheless a vital part of the quality control process as traditional automated quality evaluation methods are relatively limited in their ability to evaluate data holistically. The labor involved in this process and the potential to develop other non-traditional methods for evaluating data has inspired this research project. The goal of this project, therefore, is to develop a neural network capable of providing anomaly detection on the data with comparable accuracy to that of our human analysts.
2. The Data
   * 1. The dataset consists of “observation data”, collected via meteorological instruments installed on research vessels. Since the project’s launch in 2005, 52 research vessels have contributed to the SAMOS dataset. There are 26 labeled variable observations which are typically collected on an hourly basis. Once the raw data is submitted to a SAMOS server, an automated quality control software applies an initial set of flags on the data. This “preprocessing” step “was designed to flag data that failed to pass a series of objective evaluations” [1].
     2. These tests include, in order of application: [1]
        1. Verifying the existence of time, latitude, and longitude data for every record.
        2. Flagging non-sequential and/or duplicate times.
        3. Flagging data greater than four standard deviations from a climatology.
        4. Flagging data points that are not within a realistic range of values.
        5. Flagging platform positions and speeds that are unrealistic.
        6. Flagging positions where an oceanographic platform moves over land.
        7. Flagging inaccurate earth relative wind speed and direction data.
        8. Flagging data that fails the relationship: air temperature>=wet-bulb temperature>=dew point temperature.
     3. The preprocessor, as the name implies, was designed only as a preliminary scan through the data. The final decision to keep, reject, or add any flag falls to the DQE performing the visual inspection (section 4). The primary purpose of the preprocessor is to automate the flagging process and highlight suspect data for the DQE. [1]
     4. Over the course of the project’s 13-year history, this has culminated in a tremendous repository of raw, pre-processing, and research quality (post-processing) observation data. [1]
3. The Problem
   * 1. My goal is to reproduce the output of the visual inspection process on the SAMOS datasets. This could be approached as a classification problem (classification of observations as either reliable or unreliable), however, because this is also an anomaly detection problem, a traditional classification approach to classifying anomalies, that are by definition, rare events that would not be represented very well in the data, might be less effective than a statistical or regression method [2]. I will, instead, approach this problem with a method from regression (see the project proposal).

* 1. Relevant literature:
     1. (Neural networks used to classify)
        1. “Detection of Anomalous Drops with Limited Features and Sparse Examples in Noisy Highly Periodic Data” (Regression using LSTMs, RNNs, DNNs) (Google) [2]
        2. “Robust Principle Component Analysis” (Netflix) [8]
        3. “Seasonal Hybrid Extreme Studentized Deviate test” (S-H-ESD) (Twitter) [7]

1. The project proposal:
   1. I will apply regression-based machine learning techniques on periodic time series data with limited features and a limited set of labeled examples of anomalies.
      1. *Proposed Structural Implementations*
         1. *DNN, LSTM, RNN (Dynamic Neural Net,  Long Short-Term Memory, Recurrent Neural Network)*
2. Analytics
   1. Training accuracy
   2. Training Loss
   3. Validation accuracy
   4. Comparative Analysis of regression methods (i.e. activation function analysis, as well as structural analysis, e.g. LSTMs vs RNNs vs DNNs ect.)
   5. Skill scores, e.g. equitable threat score

REFERENCES

[1] S. R. Smith, C. Harvey, and D. M. Legler, “Handbook of Quality Control Procedures and Methods for Surface Meteorology Data,” in *Handbook of Quality Control Procedures and Methods for Surface Meteorology Data*, 1996.

[2] D. T. Shipmon, J. M. Gurevitch, P. M. Piselli, and S. Edwards, “Time Series Anomaly Detection,” *https://static.googleusercontent.com*. [Online]. Available: <https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/46283.pdf>.

[3] “The Science of Anomaly Detection,” Numenta, Redwood City, CA, 2015. [Online]. Available: https://numenta.com/assets/pdf/whitepapers/Numenta%20White%20Paper%20-%20Science%20of%20Anomaly%20Detection.pdf

[4] H. Sak, A. Senior, and F, Beaufays, “Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling”, Google Inc., Mountain View, CA, 2014.

[5] A. Krizhevsky, “Convolutional Deep Belief Networks on CIFAR-10,” University of Toronto, Toronto, ON, 2010.

[6] T. Nikolov, “Recurrent neural network based language model,” Brno University of Technology, Brno, Czech Republic, 2010.

[7] A. Kejariwal. (2015, January 6) Introducing practical and robust anomaly detection in a time series.

[8] (2015, February 10) RAD - Outlier Detection on Big Data, [Online]. Available: https://medium.com/netflix-techblog/rad-outlier-detection-on-big-data-d 6b0494371