Machine Learning with Feature Importance Analysis for Tornado Prediction from Environmental Sounding Data

Brice Coffer[1], **Michaela Kubacki**[2], **Yixin Wen**[3,4], **Ting Zhang**[5], **Carlos A. Barajas**[6,*], and **Matthias K. Gobbert**[6]

[1] Department of Marine, Earth, and Atmospheric Science, North Carolina State University, USA
[2] Department of Mathematics, Middlebury College, USA
[3] Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, USA
[4] NOAA/National Severe Storms Laboratory, Norman, Oklahoma, USA
[5] Department of Mathematics and Computer Science, McDaniel College, USA
[6] Department of Mathematics and Statistics, University of Maryland, Baltimore County, USA

Tornadoes pose a forecast challenge to National Weather Service forecasters because of their quick development and potential for life-threatening damage. The use of machine learning in severe weather forecasting has recently garnered interest, with current efforts mainly utilizing ground weather radar observations. In this study, we investigate machine learning techniques to discriminate between nontornadic and tornadic storms solely relying on the Rapid Update Cycle (RUC) sounding data that represent the pre-storm atmospheric conditions. This approach aims to provide for early warnings of tornadic storms, before they form and are detectable by weather radar observations. Feature analysis of a Random Forest machine learning model uncovers that the pressure variable has little impact on the classification process, which is consistent with known key physical attributes of tornado formation, demonstrating the ability of machine learning techniques to provide insight solely based on the data.

1 Introduction

Tornadoes can develop quickly, cause severe damage across a large spatial area, and create life-threatening conditions, thus posing a forecast challenge. Approximately 1,200 tornadoes are reported in the U.S. each year, resulting in roughly 80 deaths and millions of dollars in damage. Timely and accurate predictions of these severe weather events are key to mitigating casualties. However, tornadoes are very difficult to predict, as most severe weather will not lead to tornadogenesis.

The majority of studies of severe weather exclusively rely on commonly used parameters that describe the environment (e.g., how much instability or wind shear is in the atmosphere). The goal of this programme is to explore the use of machine learning techniques to predict significant tornadoes using only Rapid Update Cycle (RUC) sounding data that represent the pre-storm atmospheric conditions. This approach aims to provide for early warnings of tornadic storms, before they form and are detectable by ground weather radar observations. In this particular study, we investigate whether novel structures in the modeled sounding data, that discriminate between nontornadic and tornadic storms, can be found by the power of machine learning without preconceived computed environmental parameters.

The severe weather event database used in this study is that of [1] and [2], expanded to include the years 2005–2017 for tornadic thunderstorms and 2005–2015 for nontornadic thunderstorms. All tornado, significant hail (sighail), and significant wind (sigwind) reports are filtered for the largest magnitude report per hour on a 40 km spacing Rapid Update Cycle (RUC) model analysis grid and then assigned to the closest analysis hour. Since roughly 80% of tornadic events originate from supercell storms, we focus on identifying if a supercell will become tornadic. The 20,194 supercells in the dataset consist of 10,839 significantly severe nontornadic cases, 7,743 weakly tornadic (F0–F1) tornadic damage cases, and 1,612 significantly tornadic (F2–F5) tornado damage cases.

2 Results

After pre-processing the data, we build and test a Random Forest model [3, 4] using the machine learning package `sklearn` (https://scikit-learn.org/stable/) in Python. See [5] for details, pre-processing, and additional results. For our base model, hyperparameter settings are as follows: class_weight = balanced, max_depth = 200, n_estimators = 200 (all other parameters use their default settings). We use the 'balanced' setting for class weights because we have imbalanced data: 54% nontornadic, 38% weakly tornadic, and 8% significantly tornadic. The balanced setting weighs the classes inversely proportional to their frequency of occurrence. Using these settings, and a training-testing split of 80-20, the Random Forest Model predictions on the testing data results in an overall accuracy score of 71.6%.

* Corresponding author: e-mail barajasc@umbc.edu

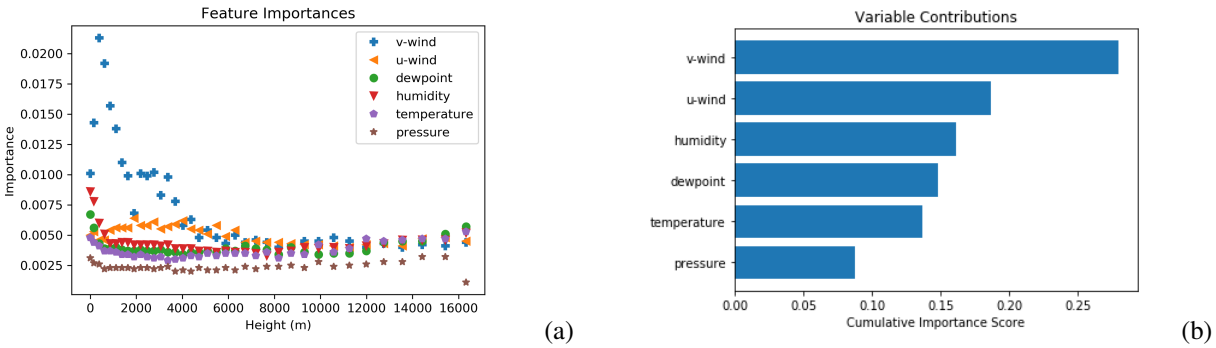

Fig. 1: Feature importance scores in (a) and variable contributions in (b).

Performance testing using various ranges of hyperparameters indicate that the model is performing optimally given the current settings and the current form of the input data. In testing a range of max_depth settings from 10 to 400, the minimum accuracy score is 64.79% (for max_depth = 10), the maximum is 71.6% (for max_depth = 200), the average accuracy score is 69.78%, and the standard deviation is $\pm 1.65\%$. We also experiment with increasing the size of the forest (n_estimators), varying choices between 100 to 2000. There were no significant gains in accuracy for larger forests, and the best accuracy is observed at around n_estimators = 200.

The ability to easily and efficiently perform feature importance analysis is our primary motivation for using the Random Forest model on this data set. Feature importance analysis can lead to insights regarding our data and can lead to model improvements. After calculating the tree-specific feature importance for each tree in our forest, we average the results to compute a feature importance score. This is done automatically using the `sklearn` function `feature_importances_`.

Our data has 222 features, however, these features are split equally across six variables, temperature, dewpoint, relative humidity, u-wind, v-wind, and pressure, corresponding to thirty-seven standardized heights ranging from 10 m to 16.3 km. Since each feature corresponds to a specific height, we plot the feature importance scores of each variable across the range of heights in Figure 1 (a), along with the cumulative feature importance scores for each variable in Figure 1 (b), in which we add up each variable's feature importance scores for every height. These figures indicate the v-wind variable scored significantly higher than other variables, specifically at heights below 4 km. We also note that the first few humidity readings were of more importance than subsequent humidity readings. But the pressure variable scores are consistently lower than other variables across all height levels, as seen in Figure 1 (a).

These feature importance results are consistent with known key features of tornado formation. Low-altitude relative humidity is decent predictor of downdraft coldness. Lower relative humidity allows for more evaporation and typically colder downdrafts. This colder, denser air at the surface resists being converged and stretched into a tornado. Along these lines, low-altitude vertical wind shear is well correlated with the strength of the convergence and stretching by the supercell on developing vortices [6].

Accurate prediction of significant tornadoes using machine learning algorithms is a relatively new and challenging data science problem. Solutions to this problem could, in time, provide a useful tool in severe weather forecasting and may provide additional insight into conditions surrounding tornado formation. We show an initial exploration of using Rapid Update Cycle (RUC) sounding data for supercell storms, that describes the pre-storm atmospheric conditions of a supercell, to train Random Forest classification for tornado prediction. Feature analysis of the model uncovers that the pressure variable has little impact on the classification process, which is consistent with known key physical attributes of tornado formation, demonstrating the ability of machine learning techniques to provide insight solely based on the data.

Acknowledgements This work is supported in part by the U.S. National Science Foundation under the CyberTraining (OAC–1730250) and MRI (OAC–1726023) programs. The hardware used in the computational studies is part of the UMBC High Performance Computing Facility (HPCF). Co-author Carlos A. Barajas was supported as HPCF RA.

References

[1] B. T. Smith, R. L. Thompson, J. S. Grams, C. Broyles, and H. E. Brooks, Weather Forecast. **27**(5), 1114–1135 (2012).
[2] R. L. Thompson, B. T. Smith, J. S. Grams, A. R. Dean, and C. Broyles, Weather Forecast. **27**(5), 1136–1154 (2012).
[3] L. Breiman, Machine Learning **45**, 5–32 (2001).
[4] A. Cutler, D. Cutler, and J. Stevens, Machine Learning **45**, 157–176 (2011).
[5] B. Coffer, M. Kubacki, Y. Wen, T. Zhang, C. A. Barajas, and M. K. Gobbert, Using machine learning techniques for supercell tornado prediction with environmental sounding data, Tech. Rep. HPCF–2020–18, UMBC High Performance Computing Facility, University of Maryland, Baltimore County, 2020. http://hpcf.umbc.edu.
[6] P. M. Markowski and Y. P. Richardson, Atmospheric Research **93**(1), 3–10 (2009), 4th European Conference on Severe Storms.