Analyze US climate Data to Predict the Occurrence of Tropical Storm

Seunghyuk Baek (sbaek44@vt.edu)
Jeevan Thapa (gforest5@vt.edu)
CS5644, Fall 2018

Introduction

Every year, tropical storm or hurricane is developed in ocean surface near equator. When a tropical cyclone falls onto US soil, it often causes massive damage to human life and property. The development of a tropical cyclone depends highly on the environmental temperature. First, warm and moisture air near ocean surface rises upwards due to the less density of the air. If massive air on the ocean surface is missing because of this air convection, surrounding air will fill the empty space then heated. Again, enough heat will cause new air to rise upwards. If this meteorological phenomena repeats for a long time, a tropical storm forms. However, sometimes a cyclone can be strengthened when it contacts warm fresh water while hovering over ground. Therefore, It is plausible to correlate temperature with development of a hurricane.

Project Problem Statement

Occasionally People live near coastal area suffer from hurricane. States near Caribbean sea often announce evacuation to their citizens to prevent human life loss. This evacuation leads chaotic escape from beloving home. If a government knows when storm will land on us soil, the evacuation can be organized efficiently. Moreover, in government's perspective, it is easier to allocate aid. Therefore, the task of this project is to find any correlation between the climate condition of US land and the occurrence of a tropical cyclone. The climate data is complex mixture of features. The feature contains average temperature, highest/lowest temperature, wind speed, and so on. Fortunately these features are in numeric values. Therefore, data analytic approach can reveal hidden correlation between the US climates and the occurrence of tropical storms.

Data Set

Data set contains location specific data and is collected from NOAA (National Ocean and Atmospheric Administration). Since, hurricanes are developed from ocean, we picked random states and their weather observation stations located near US east coasts. Each station provide

different data points and features. In average, there are 20 different features for each data set. The number of data points depends on requested range of period. The data we collected from Jan 1, 1950 to Oct 31, 2018 has 829 instances for one location. Another dataset that acquired from NOAA is historic data of landfall hurricanes. This is a table from NOAA's website with 345 data points and 7 features. For this project, data from following 5 states were considered which are: Alabama, Florida, Georgia, Louisiana and South Carolina.

Preprocessing Steps

The climate dataset is a csv file which contains string and numeric values. The first 3 features of this dataset are: state code, station division and date. Because this is a monthly data, there is only year and month for the date. After these 3 columns, actual measurement data is provided with data attributes. Data attributes describe the characteristic of this collected data in a given month. For example, TAVG is average air temperature of the month. For the missing value, we replaced with the average value of that attribute. For this analysis, first 3 columns (state name, division and date) of data are deleted. The list of hurricane data is trimmed so that only data from 1950 to 2018 is available. Hurricane data is divided by states. Therefore, hurricane data results in 5 different tables that has 2 features; states and hurricane landing month. In last step, US climate data by month and hurricane data are merged so that US climate data by state. The final data set is represented by 5 tables grouped by state name having one extra feature which indicates whether a hurricane landed on particular state at the given month of the year.

Methods and Models

1. Decision Tree Classifier
2. Bayes Classifier
3. Linear Regression
4. Neural Network

The models were chosen based on the theory that past climate data could be used to predict the occurrence of tropical storm. Since historic data is accessible, we choose supervised learning method for this project. After skimming through each state's data, we found that there is significant differences between the number of hurricanes that landed on each states. Therefore, each states were evaluated separately. For neural network classifier, 3 hidden layers with 30 nodes was used. Decision tree and neural network classifiers were cross validated with 5 fold

cross-validation. For cross validation, accuracy, precision, f1 score and recall values were averaged.

Results

Alabama Decision Tree			Florida Decision Tree	
	Average	**Standard Deviation**	**Average**	**Standard Deviation**
Accuracy	0.97	0.007	0.93	0.035
Precision	0.0	0.0	0.35	0.35
F1 Score	0.08	0.16	0.17	0.17
Recall	0.0	0.0	0.20	0.12

Georgia Decision Tree			Louisiana Decision Tree	
	Average	**Standard Deviation**	**Average**	**Standard Deviation**
Accuracy	0.99	0.011	0.95	0.019
Precision	0.0	0.0	0.14	0.13
F1 Score	0.0	0.0	0.19	0.11
Recall	0.0	0.0	0.17	0.11

South Carolina Decision Tree		
	Average	**Standard Deviation**
Accuracy	1.0	0.0
Precision	0.0	0.0
F1 score	0.0	0.0
Recall	0.0	0.0

Alabama Neural Network			Florida Neural Network	
	Average	**Standard Deviation**	**Average**	**Standard Deviation**
Accuracy	0.99	0.0045	0.96	0.010
Precision	0	0	0	0
F1 Score	0	0	0	0
Recall	0	0	0	0

Georgia Neural Network			Louisiana Neural Network	
	Average	**Standard Deviation**	**Average**	**Standard Deviation**
Accuracy	0.99	0.0059	0.97	0.011
Precision	0	0	0	0
F1 Score	0	0	0	0
Recall	0	0	0	0

South Carolina Neural Network		
	Average	**Standard Deviation**
Accuracy	1	0
Precision	0	0
F1 Score	0	0
Recall	0	0

Regression Model	Alabama	Florida	Georgia	Louisiana	South Carolina
Residual sum of squares	0.01	0.04	0.00	0.03	0.00
Variance score	0.05	0.11	0.02	0.06	1.00

Bayes Classifier		
Alabama	Predicted	[0. 0. 0. 0. 0. 0. 0. 1. 0. 1.]
	Truth	[0. 0. 0. 0. 0. 0. 0. 0. 1. 0.]
Florida	Predicted	[0. 0. 0. 1. 0. 1. 1. 0. 0. 1.]
	Truth	[0. 0. 0. 0. 0. 0. 0. 0. 1. 0.]
Georgia	Predicted	[0. 0. 0. 0. 0. 0. 0. 0. 0. 1.]
	Truth	[0. 0. 0. 0. 0. 0. 0. 0. 1. 0.]
Louisiana	Predicted	[0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
	Truth	[0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
South Carolina	Predicted	[0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
	Truth	[0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]

Conclusion

In conclusion, The data set we used for this project seems relatively large for the hurricane data analysis. In fact, all of neural network machine learning algorithms resulted in almost 100% accuracy and 0 for other statistic test. This reflects that the occurrence of hurricane in each states is far scarce (11 storms for AL, 37 for FL, 4 for GA, 23 for LA and 11 for SC) compare to the total instance of data which is 829. Other classifiers resulted very similar, except with decision tree classifier. With decision tree classifier, FL and LA data showed some reasonable statistic data. According to the result from these two states, the machine learning model is not suitable to detect upcoming tropical storm since recall and precision values were less than 0.5. For future projects, collecting data from the region that suffers many hurricane may be needed.

Member 1

Seunghyuk Baek

- Analyzing Data
- Write up draft report / proofread

Member 2

Jeevan Thapa

- Collect Data
- Write up draft report / proofread