Winter Precipitation Forecasting Based on Random Forest Algorithm

Kelly Lavertu
Kennesaw State University
Marietta, GA USA
klavertu@students.kennesaw.edu

Brian Rawls
Kennesaw State University
Norcross, Georgia USA
brawls3@students.kennesaw.edu

Christopher Sefcik
Kennesaw State University
Woodstock Georgia, USA
csefcik@students.kennesaw.edu

ABSTRACT

Our paper will build on the research performed by Kanavos et al. in the research paper “Forecasting Winter Precipitation based on Weather Sensors Data in Apache Spark”. We will be using data gathered from weather sensors to forecast precipitation outcomes: rain, freezing rain and snow. We will use the classical random forest algorithm and compare the results with the algorithms used by Kanavos’ et al. research paper. The performance of the algorithm will be measured based on accuracy and computation time. We will then compare our results with the results found in their research paper.

RESEARCH STATEMENT AND CONJECTURE

Our research team is comprised of several computer science master’s students with an interest in algorithms, computing, and machine learning.

Predictive Forecasting in Modeling Precipitation Outcomes

This area of our work, modeled using automated surface observing systems (ASOS) weather datasets, observes the outcome of performance of random forest and several other algorithms against the proposed dataset. Our research is aiming to answer questions such as: is Random Forest more accurate in predictive power given our model is trained using a system with sufficient storage and memory? Does Random Forest prove to take less computational time than any of the observed algorithms in the “Forecasting Winter Precipitation based on Weather Sensors Data in Apache Spark” research paper?

We hypothesize that random forest will prove to be more optimal, and performant based on accuracy in predictive power and computation time. However, there is a known drawback: a trained forest will require significant memory for storage due to the need to retain information from a vast number of individual trees. We will additionally need to deploy our model on a system with sufficient resources given our project budget.

PRELIMINARY LITERATURE SURVEY

The “Forecasting Daily Stock Trends Using Random Forest Optimization” research paper discusses using Random Forest to predict daily stock market trends. The research team captures its stock price datasets from the Korea Composite Stock Price Index (KOSPI) list and compares the stocks of Celltrion and Pharmicell. One of the challenges is cleaning the initial dataset and selecting a strong set of predictors for training. The Random Forest classifier adaptively determines the optimization parameters to improve the predictive accuracy while avoiding overfitting issues. The researchers found it best to reduce its stock variables down to a small subset with the least number of uncertainties. Due to the highly volatile nature of stock prices, all variables still had an unacceptable amount of uncertainty. This paper is relevant in our research in that our team is researching trends, specifically weather forecasts, and looking to predict the forecast outcomes using random forest. The paper details the importance of cleaning the dataset and reducing any imbalance before utilizing the Random Forest algorithm. Our research is relying on a consistent trend among the datasets we will train, reduction of noisy variables. and grouping of seasons. This will also help effectively reduce the possibility of an overfitting issue. It is our belief that with the given preliminary

parameters input for our algorithm we should have a trainable model and an acceptable predictive output of forecast.

In "A comparison between Machine learning algorithms for the application of micro-grids Energy management", Khoshlessan et al. compares the performance of support vector machine, logistic regression, decision tree, and gradient boosting with random forest with respect to classifying the operation modes of microgrids. Khoshlessan et al. found that the model with the best performance is random forest - which was able to classify with 93-95% accuracy when used with the Scikit-Learn PolynomialFeatures package for preprocessing data. For our research, this paper suggests an optimal preprocessing technique to use in combination with Random Forest.

In the research performed by Dai et al. “Using Random Forest Algorithm for Breast Cancer Diagnosis”, the Random Forest algorithm is used to diagnose breast cancer. The type of Random Forest algorithm used, classification and regression trees (CART), consists of four steps: determining the number of attributes, splitting the node using the Gini coefficient, performing training tasks on the decision tree, and voting for the best solution. The researchers chose the Random Forest algorithm due to its high predictive accuracy in combining results of multiple decision trees. Also, weak predictors can be combined to produce an accurate predictor. The overall result of the paper was that the Random Forest algorithm performed well and produced accurate predictors. The results of this paper helped reinforce our decision to apply the Random Forest algorithm to predicting weather precipitation.

METHODOLOGY

Our objective, which is to compare performance measurements of different classification algorithms against the Random Forest algorithm, will require us to create a controlled environment in order to make the comparisons. The measurements used: lowest accuracy, highest accuracy, difference, average and time, will be computed for each algorithm after completing a fixed number of trials. Overall, the measurements will provide detail on which algorithms are faster to run, how well they create classifications and how well they create the classifications consistently. The performance measurements of the Random Forest algorithm will be compared against the performance statistics for: Naive Bayes’, Hoeffding Tree and the Adaboost algorithm.

The algorithms will be run using the Spark ecosystem. The hardware that our Spark ecosystem resides on consists of 64 GB RAM, 1 TB storage SSD, M1 Max Chip, 10 Core CPU and a 32 Core GPU.

Within our controlled environment, all our algorithms will be run against three standard datasets. We are currently in the process of refining our standard datasets but we do have a good baseline for the variables we are going to use and which records we are going to proceed with from the (ASOS) weather data provider. Each dataset will contain a different number of records but with the same records from the smaller dataset. Running the algorithms against three different sized datasets will allow us to capture any performance loss due to an increasing record count. While our performance loss might not be substantial on a smaller test level, when dealing with millions of rows, the overall performance loss could be substantial if exponential. During the execution of each algorithm, the dataset will be split into a training set and a test set. We will split the dataset randomly using an 80% training split and a 20% test split.

While executing our algorithms on the standard datasets within our controlled environment, measurements will be computed during the execution of each run. These measurements will then be combined, averaged and tabulated by each algorithm for each of the three different sized datasets.

PROGRESS

Note: Project task will be agile based with 2-week sprint interactions

1. Sprint 1 (Feb 21st – March 6th)
    - Obtain forecast dataset → **finished** (https://mesonet.agron.iastate.edu/request/download.phtml)
    - Clean forecast dataset → **finished**
    - Record observations, modifications, and expected outcomes → **finished**
2. Sprint 2 (March 7th - 21st)

- Random Forest proof of concept with sample size of dataset (Scala) → **in progress**
- Setup deployment environment → **in progress**

3. Sprint 3 (March 21st – April 4th) → **in progress**
    - Coding and running a subset of algorithms listed in our primary research source against the complete forecast dataset
    - Running Random Forest algorithm against the complete forecast dataset
    - Observing results of all algorithms and comparing against our conjecture
    - Preparing progress report
4. Sprint 4 (April 4th – May 5th)
    - Preparing project report
    - Prepare presentation

PRELIMINARY RESULTS AND ANALYSES

Our initial implementation, of several algorithms, focuses on the Random Forest algorithm to establish a baseline. A reliable prediction output from Random Forest requires a robust, cleaned, dataset. In cleaning our dataset we focused on several requirements - code our dependent variable to quantitative values, remove all variables with only qualitative values, and for our resulting variable set we perform median imputation on all missing values. Our initial dataset included 31 variables and after this feature process the dataset was reduced to 15 variables with no missing values (i.e. a missing value is one that was reported as missing or a value that was set to missing after meeting some general quality control check, or a value that was never reported by the sensor).

Random Forest requires features and a dependent variable to make initial predictions. The variable representing forecasted weather (wxcodes) was chosen as our dependent variable as its value is computed from an analysis of each variable's output in a unique observation. Additionally our dataset was transformed into an array of vectors and split into 70% training data and 30% testing data.

In our initial run we processed over a million observations with 40 trees, 42 bins, a max depth of 30, and impurity set to gini. The resulting accuracy of our prediction model was 0.001138699 revealed by the multi class classification evaluator. This lower percentage generally would mean our model is close to perfect, however that result could be misleading to which case we would like to perform gradient boosting on our model.

Currently, our dataset's time frame spans five years, from 2017-2022. Our goal is to eventually reduce the dataset to a more specific time frame to yield better prediction over the targeted forecast codes. This reduction should remove a number of variables that have missing data, and consequently, potentially reveal more variables that can assist in accurately predicting winter precipitation. In future runs, the overall dataset should be reduced from a million observations to a few thousand and we will narrow our window to five specific months, November through March, over those five years.

CONCLUSION AND FUTURE WORK

With our current output, we are still in the process of drawing concrete conclusions to determine what variables have high predictive power to make classification decisions between rain, snow and freezing rain. When we have fully reduced our variables and filtered the observations, we will then use this dataset to run against other algorithms and measure the performance. Having a defined dataset with a fixed number of records will allow us to create a controlled environment in which we will use to make performance comparisons between each algorithm.

Future work in the area could consist of expanding on our results to include other classification algorithms. In our research paper, we provide our machine specifications and we can provide our dataset upon request and or our filter criteria to assist future researchers in mimicking our controlled environment to run performance tests against.

REFERENCES

[1] B. Dai, R. Chen, S. Zhu and W. Zhang, "Using Random Forest Algorithm for Breast Cancer Diagnosis," in 2018 International Symposium on Computer, Consumer and Control (IS3C), Taichung, Taiwan, 2018 pp. 449-452.doi: 10.1109/IS3C.2018.00119keywords: {decision trees;classification algorithms;training;machine learning

algorithms;machine learning;breast cancer}url: https://doi.ieeecomputersociety.org/10.1109/IS3C.2018.00119

[2] A. Kanavos, T. Panagiotakopoulos, G. Vonitsanos, M. Maragoudakis and Y. Kiouvrekis, "Forecasting Winter Precipitation based on Weather Sensors Data in Apache Spark," 2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA), 2021, pp. 1-6, doi: 10.1109/IISA52424.2021.9555553.

[3] M. Khoshlessan, B. Fahimi and M. Kiani, "A comparison between Machine learning algorithms for the application of micro-grids Energy management," 2020 IEEE International Conference on Industrial Technology (ICIT), 2020, pp. 805-809, doi: 10.1109/ICIT45562.2020.9067203.

[4] J. S. Park, H. Sung Cho, J. Sung Lee, K. I. Chung, J. M. Kim and D. J. Kim, "Forecasting Daily Stock Trends Using Random Forest Optimization," 2019 International Conference on Information and Communication Technology Convergence (ICTC), 2019, pp. 1152-1155, doi: 10.1109/ICTC46691.2019.8939729.

[5] Daryl Herzmann Akrherz@iastate.edu. IEM :: Download ASOS/AWOS/Metar Data. Retrieved from https://mesonet.agron.iastate.edu/request/download.phtml