Shalini Panthangi

March 17, 2019

Mrs. Craun

10th Grade Research

Power Grid Reliability Prediction Using Machine Learning Algorithms

**Background:**

A power grid is a connected network of power towers that delivers electricity between the utility companies and the people that use the power. Resilience of a power grid is the rate and speed at which the power grid can return back to its original conditions after a power outage or an extreme weather event. Power grid reliability is the ability of the power grid system to deliver electricity as demanded by the people using the power and electricity (Ensuring Electricity System Reliability, Security, and Resilience, 2017). There are many causes of a power outage, mainly including extreme weather events such as hurricanes, thunderstorms, or tornados. However, power grids can lose reliability due to the number of people using power from the grid at once. It is possible to look at past data regarding power outage events to determine future power outage events through the use of machine learning. Machine learning is a modern math modeling approach which is essentially the use of specific past data in algorithms that are able to predict future occurrences of that event by training the model with the data, fine-tuning the model, and analyzing sources of error in the model to improve the level of prediction. Machine learning can be used on data from a multitude of sources and fields, including genetics data, environmental data, utility data, and much more.

The most important part of machine learning research is the specific types of machine learning algorithms to predict the reliability of power grids. The two algorithms being considered are the Support Vector Machine Model and the Bayesian Additive Regression Tree Model. A support vector machine is a type of machine learning algorithm approach involving the use of a separating hyperplane (Fouodo, König, Weihs, Ziegler, & Wright, 2018). A hyperplane is a constant value that represents a subspace of a set vector space. It is also usually one dimension less than the defining dimensions of the set space. The dimension of a mathematical space is the number of coordinates in the data; this can change as a result of different amounts and types of data. The two main types of hyperplanes are vector hyperplanes, which are created for data going through the origin, and affine hyperplanes, which are essentially translated vector hyperplanes. In this research, an affine vector plane will be used since the data does not perfectly go through the origin (Huang, Cai, Pacheco, Narandes, Wang, & Xu, 2018).

Given supervised training, meaning that the data passed through the algorithm is labeled, the algorithm creates a hyperplane that categorizes the outputs of a certain set of inputs into two parts per each dimension. Essentially, a support vector machine creates a hyperplane that separates different classes in a multi-dimensional vector space. In order to train the data used to fit a support vector machine model, the data needs to be transformed into another dimension using transformations called kernels and then transformed back to the original data set. Kernels are applied on each data set to map the original multi-dimensional, randomly placed data into a space with higher dimensions to allow the separation of classes by the hyperplane. Not using these kernels confuses the classification due to the vast amounts of data, and a proper hyperplane will not be formed (Afonja, 2017).

In a support vector machine, there are also tools called tuning parameters which include regularization parameters. Tuning parameters are used since any one model will never fit all of the data provided. So, the model needs to be fine-tuned and trained over and over. Regularization parameters prevent the overfitting of the model to the data. This type of overfitting is when the model fits the data too well to the point where it negatively impacts the ability of the model to predict new data because it so perfectly fits the old data. This overfitting can also be prevented through the use of a cross-validation model, which past research has used and future research will use to verify and test the models. However, regularization parameters tackle this problem by discouraging the learning of a more complex model on the data to avoid overfitting. In mathematical terms, the coefficient estimates of the vectors used to fit the data will be shrink toward zero, involving the residual sum of squares function to figure out how to best minimize these coefficient values (Regularization).

In addition to a support vector machine, another machine learning algorithm being considered is the Bayesian Additive Regression Tree Model (BART). The BART model is a nonparametric regression approach incorporating the structure of a decision tree (Self, 2018). At the basic level, a decision tree or regression tree is a machine learning algorithm used to differentiate between the different outcomes of a set of data. Decision trees are incredibly beneficial as they are able to achieve high accuracy and are relatively easier to interpret than some of the other models (Linero, 2017). To create this model, the decision tree first needs to be built, considering all of the possible outcomes and variables, but only using the best features to classify the data, and then needs to be streamlined to eliminate all unnecessary decisions from the tree, reducing the model’s level of complexity (Jost, 2017). This model especially uses nodes, which are places in a model where computation takes place to provide results that are used later in the model. Bayesian trees are different from other decision tree models such as random forest regression as it fully explores the model space. The BART model uses more sophisticated regression models in place of the simple decision nodes for some of the other models. (Chipman, George, & McCulloch, 2001). Lastly, the BART model uses dimensionally adaptive random basis elements to determine the estimates of the function and the potential causes of error. For both of these models, the exact math behind the model is dependent on the type of data being trained. Because of this, no explicit functions or regression models can be explained.

**Previous Research:**

Previous research by Eskandarpour, Khodei, and Arab (2016) used the support vector machine model to improve the resilience of a power grid. That research did not just use one type of support vector machine, but used a Linear SVM, Quadratic SVM, Cubic SVM, and Gaussian SVM. This research also used a k-cross validation model to verify and fine-tune their model and results. All four of these models had greater than 80% accuracy at predicting these outage events. It was also seen that the Gaussian SVM gave the highest F1-Score, which incorporates both precision and recall scores. This research only looked at hurricanes causing power outages instead of smaller events, and they did not account for environmental factors for power outages.

Past research conducted by Nateghi, Gulkema, and Quiring (2011) has found that the BART model gives the most accurate and precise predictions for the power outage events. Quantitatively, the BART model was found by previous data to have the least amount of error for the model but still give the most accurate predictions. For example, while all other residual values were in the thousands, the BART model was the only one with residual values less than a thousand. Past research was also able to find the most important elements to determining the power outage event using the BART model. Some of these important variables were wind speed, outage event, and location of the power grid.

Some problems in the machine learning field is that no model is perfect. With the constant, spontaneous changing of data, the models have to constantly be fine-tuned to figure out a correlation between the different variables in the data. Past research has only looked at data regarding the specific hurricane that caused the power outage event, but failed to include vital data regarding population where the power grid is located, environment type, precipitation in the area, and location of the power grid. This causes many important factors of a power outage event to be overlooked because the conditions of the power grid are very important to determine its reliability (Nateghi et al, 2011). Poor conditions could cause an increased failure in the grids, while good conditions increase reliability.

**Future Research:**

Past research has looked at power grid resilience as the subject to improve, while future research will improve reliability. This past research also looked at power outage events as a result of big disasters like hurricanes, while future research will focus on smaller outage events that need more attention than the big events that people are already able to notice and predict. Future research will also use data from different outage events, not just limited to hurricanes or extreme weather events like past research. Future research will include data not just related to the cause of the outage, but also data related to population, location in the US, soil type, depth to bedrock, and annual precipitation as factors of power outage. These are important to the prediction of the outage event as it is not just a weather event that reduces the reliability of a power grid, but also the conditions of the power grid in its specific location. In the later years of this research, these predictions will be used to figure out which power grids to restore or improve, considering all of the factors that may cause power grids to not be as reliable, including the increase in population, climate change, etc. Future research will also use the System Average Interruption Duration Index and frequency as a measure of reliability. The greater the frequency and the longer the duration, the less reliable to power grid system is. Lastly, future research will steer away from finding optimal locations to place new power grids, as that is not practical due to economic limitations by countries. Instead, it will look at legacy of the power grid and predict which power grids currently existing to improve.

References:

Afonja, T. (2017, January 02). Kernel Functions. Retrieved from

https://towardsdatascience.com/kernel-function-6f1d2be6091

Fouodo, C.J.K. König, I.R., Weihs, C., Ziegler, A., Wright M.N. (2018 July). Support Vector

Machines for Survival Analysis with R. The R Journal Vol. 10/1.

Chipman, H., George, E., McCulloch, R.E. (2001, February). Bayesian Treed Models.

Retrieved from http://www-stat.wharton.upenn.edu/~edgeorge/Research\_papers/treed-

models.pdf

Ensuring Electricity System Reliability, Security, and Resilience. (2017, January).

Retrieved from https://www.energy.gov/sites/prod/files/2017/01/f34/Chapter%20IV%20 Ensuring%20Electricity%20System%20Reliability%2C%20Security%2C%20and%20Resilience.pdf

Eskandarpour, R., Khodei, A., Arab, A. (2016). Improving Power Grid Resilience Through

Predictive Outage Estimation. US National Science Foundation.

Gupta, P. (2017, November 15). Regularization in Machine Learning. Retrieved from

https://towardsdatascience.com/regularization-in-machine-learning-76441ddcf99a

Huang, S., Cai, N., Pacheco, P.P., Narandes, S., Wang, Y., Xu, W. (2018). Applications of

Support Vector Machine Learning in Cancer Genomics. Cancer Genomics and

Proteomics 15: 41-51

Jost, Z. (2017, July 02). “Bayesian Additive Regression Trees” paper summary.

Retrieved from https://towardsdatascience.com/bayesian-additive-regression-trees-paper-

summary-9da19708fa71

Linero A.R. (2017) A review of tree-based Bayesian methods. Communications for Statistical

Applications and Methods; 24: 543-559

Nateghi, R., Guikema, S., Quiring, S.M. (2011). Comparison and Validation of Statistical

Methods for Predicting Power Outage Durations in the Event of Hurricanes.

Self, G. (2018, November 03). A Guide to Decision Trees for Machine Learning and Data

Science. Retrieved from https://towardsdatascience.com/a-guide-to-decision-trees-for-

machine-learning-and-data-science-fe2607241956