EARTHQUAKE PREDICTION

Vijay Kumar Kodari | Sri Krishna Tirthala | Dharma Teja Bandaru

[CS519 APPLIED MACHINE LEARNING I] [Open Machine Learning Project]

Introduction:

Man has always looked to computer systems for help in every area of human endeavours and even in understanding very complex systems as the human ability is somewhat limited in handling the amount of data generated every day. As the size of this data has increased considerably, the need to explore the use of computer system to learn from the data became imperative.

Machine learning [1], which is an application of artificial intelligence, provides computer systems the ability to learn autonomously from experience and improve as well without being explicitly programmed. Natural disasters cause massive casualties, damages and leave many injure. Human beings cannot stop them, but timely prediction and due safety measures can prevent human life losses and many precious objects can be saved. Earthquake is one of the major catastrophes. Unlike other disasters, there is no specific mechanism for earthquake prediction, which makes it even more destructive.

Though many of them say that it is impossible to make earthquake predictions, few Scientists declare that it is a predictable phenomenon.

Motivation:

Believing that earthquakes could be predicted, from the past experiences we can say that we can avoid life loss and property damage if we have an estimation on severity of earthquake that will occur in near future.

On a yearly basis, there are approximately 14,000—16,000 fatalities due to earthquakes [2]. The count of fatalities per year due to earthquakes when compared to other disasters and accidents might be less. Whereas, an abrupt and a huge earthquake in a city could take away thousands of lives at once. So, a solution to this problem using machine learning would save many lives and properties with no much of the damage.

The formal motivation behind the project is to use and learn various regression techniques which helps us to understand the appropriate Implementation of regression techniques upon the problems.

Problem Definition:

What could be the Solution?

- When the quake will occur.
- Where it will occur. (Future Implementation)

The above two factors could be able to give a proper intimation to the governing bodies and it could warn them to take the required actions accordingly.

But considering the data we have, the **WHEN** can only be predicted and that would help to some extent.

What are Solution Benefits?

Earthquakes are natural disasters and it could occur in any place around the earth, but we must be worried if the prediction is in an overcrowded place or in the place of historic sculptures or national monuments. In this case, predictions that are made accurately could give an intimation to evacuate the place and make amendments to the buildings with concrete and redesign them in a way that it's foundation can take that sort of punch.

Lifetime of the usage of solution and probability of solution to be correct is often based on hypothesis. As it is related to nature, sometimes the levels of earthquake might be more that of expected or it might not occur. The solution from the algorithm always depends on the balance of nature, i.e.; if the solution is made when the ecology system is good, the algorithm might not work as expected when the ecology system is not good in future. The solution might be useful for short-term or Long-term depending on the balance in nature and ecology system.

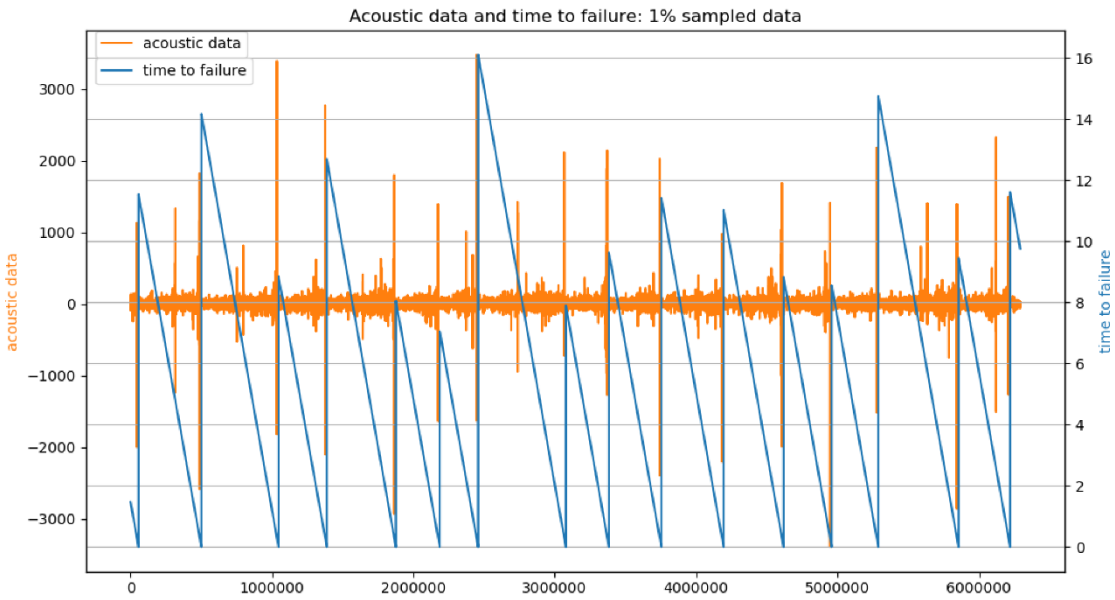
How would I solve the problem?

Understanding the data:

There are three kinds of datasets provided [sample, Training, Test]

In the train.csv, there is huge data i.e.; 600 million rows of data with two columns 'acoustic_data','time_to_failure' . acoustic_data : is the acoustic signal measured in the laboratory experiment. time_to_failure: the time until a failure will occur.

In the test data, it is divided into many segments but there is only data regarding the acoustic data with 15000 rows, and we need to predict the time_to_failure.



Processing the data:

As we need to predict time_to_failure only with acoustic_data present in test datasets, few common statistical features like mean, max, min, standard deviation must be calculated using the data in train.csv.

We can apply the regression and bring out the time_to_failure values.

The regression methods [3] we are trying to implement are the below three:

- Linear regression
- Decision tree regressor
- Random forest regressor

Linear Regression:

Linear regression models predict a continuous target when there is a linear relationship between the target and one or more predictors. In this method, the variable to be predicted depends on only one other variable. This is calculated by using the formula that is generally used in calculating the slope of a line.

Decision tree Regressor:

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

Random Forest Regressor:

A Random Forest [4] is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap Aggregation, commonly known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

Methodology:

Upon our research we understood that Algorithms like Cataboost [5], light-gbm [6]and Xgboost [7] works best for the above problem but first we are working on the basic structure of regressions and compare the results with the well defined algorithms as stated above so that It will help us to understand regressions in a better way.

Proposed solution:

Predicting the time remaining before the next laboratory earthquake i.e.; (time_to_failure) is the solution for this project, for which we are using the (acoustic_data) from the training data and are able to accomplish the task.

Future Implementation

Furtherly, we can also use classification techniques and work on generating seismic signals in (acoustic_data), so that the **WHERE** issue can also be solved as the values depend upon the laboratory locations.

As it was said the WHERE issue can also be predicted, it is done using the **Seismic data** provided in the dataset. The seismic data is generally gathered when geoscientists visualize the subsurface of the earth using waves of sound to “map” geologic structures. So, the value varies from one location to another. Predicting the time left using the seismic data would obviously give information about how much time left before an earthquake in a specific location from where the seismic data is obtained.

REFERENCES

1. https://expertsystem.com/machine-learning-definition/

2. https://www.researchgate.net/publication/307951466_Earthquake_magnitude_prediction_in_Hindukush_region_using_machine_learning_techniques

3. https://developer.ibm.com/technologies/data-science/tutorials/learn-regression-algorithms-using-python-and-scikit-learn/

4. https://www.geeksforgeeks.org/random-forest-regression-in-python/

5. https://towardsdatascience.com/catboost-vs-light-gbm-vs-xgboost-5f93620723db

6. https://medium.com/@hanishsidhu/whats-so-special-about-catboost-335d64d754ae

7. https://www.kaggle.com/c/LANL-Earthquake-Prediction/discussion/89909