A Comparative Study on Earthquake Prediction using Machine Learning Algorithms

Kaladevi.A.C[1], Gayathri.M.L[2], Gopika.I[3], Iswarya.K[4]

[1] Professor, Computer Science and Engineering, Sona College of Technology,Salem, Tamilnadu, India,
kaladeviac@sonatech.ac.in
[2,3,4] U.G Student, Computer Science and Engineering, Sona College of Technology,Salem, Tamilnadu, India,
gayathri.18cse@sonatech.ac.in, gopika.18cse@sonatech.ac.in, iswarya.18cse@sonatech.ac.in,

Abstract

Natural Disaster is the major reason for the huge loss of property, displacement of population, damage to financial and economy of the country or a certain area, etc. This can be reduced to a lesser amount by giving prior information about the disaster to the people and making them alert and can be evacuated from that place. But some failure cases in prediction of disaster may lead to unnecessary anxiety to the people, and further induce the people in loss of confidence in the system. Therefore, this article concentrates on reducing the false prediction and provides better accuracy in prediction. This article mainly focuses on Earthquake. Machine Learning provides a better algorithm for clustering and augmentation of the data. Here we have chosen three machine learning algorithms namely Logistic regression algorithms, XGBoost(Extreme gradient boost) and Light gbm (gradient boost machine). By comparing all the three algorithms of machine learning we have concluded that light gbm has more accuracy in the result.

Keywords: Logistic regression, light gbm, Xtreme boost gradient

1. INTRODUCTION

Natural Disaster events are purely unexpected events which cause drastic changes in human lives and their properties. In addition, earthquakes cause environmental issues such as surface faulting, ground failure and may also cause tsunamis. Ground shaking caused by earthquakes lead to collapse of buildings, bridges and dams which causes severe destruction to the peoples' livelihood. Vibrations caused by earthquakes generate surface rupture and displacement of ground which produce deep ruts and steep banks. This further makes the land prone to damage. A survey says economic damage due to earthquakes includes reduction in Gross Domestic Product (GDP) per capita by 1.6%. Developing countries and low-income countries face extreme economic fall in the global market. When GDP decreases there will be decrease in the earning capacity of the people which affects development of the country. Many governmental organizations and non-governmental organizations have taken steps in predicting the disaster beforehand and forecast the people about the disaster. Failure cases of prediction have increased which is a huge challenge in the prediction system. In the city of Parkfield, at San Andreas fault, a scientist has implemented his system for the prediction of earthquakes, he sowed a region of 40 km (25 miles), with seismic sensors to detect the movement of earth surface in real time. But an earthquake was recorded on 28 September 2004 with a measurement of 6.0 on the Richter scale and this earthquake has not been predicted by his system. Likewise, the United States Geological Survey(UGSC) has predicted an earthquake event at a right location and predicted the size of the event, which is the effect of the earthquake correctly, but the time of the occurrence was false which took 11 years to happen. These failure cases in earthquake forecasting are due to rare accuracy in the result.Machine Learning algorithm is defined as the target function (f) of the input variable (x) which maps to an output variable (y): Y=f(x). In this paper, we have implemented three algorithms Logistic regression, Light gbm and XGBoost and give the comparative results.

2.LITERATURE SURVEY

In the disaster prediction field, many scientists and researchers have printed their hands and given their ideas for the prediction methods. A research paper has been found with the prediction of precipitation forecasting. In this paper the researchers, T. Tang, D. Jiao, T. Chen, and G. Gui have proposed that machine learning algorithms are very useful in the prediction field. In their research they have implemented an extreme gradient boost algorithm which is a ML algorithm used for prediction processes. In this research they have used the k-means clustering algorithm for clustering the dataset for preprocessing. SMOTE (Synthetic Minority Oversampling Technique) algorithm is used for oversampling the data set; this is the unbalanced data can be converted into the balanced data. For the prediction process, an extreme gradient boosting algorithm has been used for the prediction process and they have implemented LSTM (Long-Short Term Memory) for storing more data and for quick access. For more accuracy RF (Random Forest) algorithm has been implemented [1]. FengChen has proposed a better method for data augmentation based on self-supervisedlearning. He has integrated instances and new categories and united them to bring a new boosting framework which brings robustness and extensive effectiveness in his work. Dataset such as Market-1501, DukeMTMC-reIDand CUHK03 are used in this paper where those datasets consist of large amount of data. Hehas also used GAN (Generative Adversarial Networks) based data augmentation which is a technique in data augmentation where it has a capability to create new sample data. [2]. Likewise, another researcher has proposed his idea by giving his paper about the prediction model, in his paper he has said that the vertical movement of the mass, moment, temperature, is governed by the planetary boundary layer height [BLH], [3]. This paper is a comparison between AdaBoost, Extreme Gradient Boosting XGBoost and Logistic regression algorithm. All these algorithms are a machine learning algorithm which is essentially used for the prediction process. This paper has proposed that the Logistic regression algorithm is working better for the datasets when the data is not hardly imbalanced, compared with AdaBoost and XGboost. They have also proposed that when compared with the low rate of input that is when the input is given with less percentage in number. It is said that the accuracy rate of all these algorithms, that is the logistic regression, AdaBoost and XGboost performs poorly with the minority of the input [4]. W. Han, Y. Gan, S. Chen and X. Wang are the researchers who have worked on the prediction process of the earthquake using the supervised learning methods [5].

3. PROPOSED METHOD

To predict the result the earthquake dataset is taken, the data is preprocessed and cleansed, data is clustered using k means clustering algorithm then the output is loaded on ML models where we use Light gbm, XG Boost, and logistic regression algorithms and the comparative result is predicted. The block diagram for comparative analysis of ML algorithm is in fig 1.

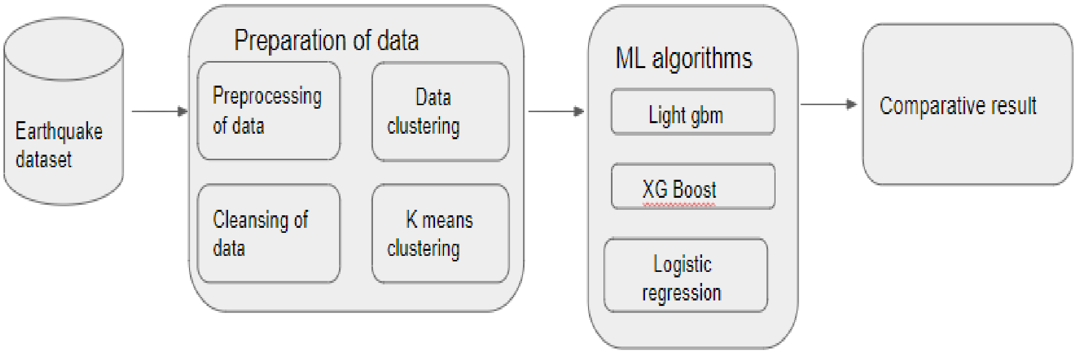


Fig 1: comparative analysis of machine learning algorithms

3.1 K-means Clustering

K-means clustering is the machine learning algorithm which is useful in clustering the data into K cluster categories. It is an unsupervised learning algorithm (i.e., it can cluster the unlabeled or unstructured data into k number of categories according to our requirement). This algorithm is very useful in solving the clustering problems. K-means clustering takes the given input data and finds the similar values or features and categorizes those values into the same group. Likewise, the given huge data will be clustered into k number of clusters as per required. The clusters categorized by the K-means clustering algorithm are named as k. This algorithm is mostly used for data mining purposes and statistics problems. The k-means algorithm initially presents the k points in the object space which represents the initial group of centroids. Centroids are nothing but the center of clusters which is represented by an imaginary or the real position. Then on k-means clustering each object will be allocated to the nearby or immediate k object in distance. On assigning the k objects, the position of k will be calculated further again and again. These steps will be processed repeatedly until the position of k remains unchanged.

3.2Data Pre-processing

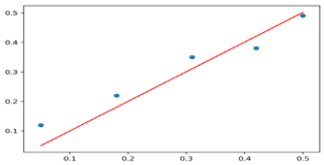
Pre-processing of data is nothing but making the dataset efficient to given as an input to the model. As the data from the dataset will be raw data and in a not understandable manner, so those data will be unfitted to give as an input to a machine learning algorithm. Pre-processing of data is a very essential step before giving it as an input data to a learning algorithm. We are importing the necessary library to run the python code such as NumPy which is a python library doing any kind of mathematical operation and scientificcalculation in the program, this is a fundamental open source library for python, then import matplotlib and in that import the sub library called PyPlot, thus install matplotlib.pyplot this is used for displaying a two dimensional plotting and for displaying any kind of plots, then we are going to install pandas which is also an fundamental open source library which is useful in managing and supervising the dataset for any machine learning algorithm. Once the dataset has been imported then the next step is extracting the dataset into dependent and independent variables. Extracting the dependent variables separately and independent variables separately. Then encode the categorized data, then extract the required attributes from the dataset in our project the required attributes are date, time, longitude, latitude, magnitude, and depth. With this extracted data we can further proceed with the prediction process.

3.3Training the dataset:

Getting the dataset as input we are going to train the model. It works well with Classification and regression problems. It classifies the dataset to n number of trees and predicts output based on each tree result. If there are more trees, then the accuracy will be increased. The main advantage of random forest is that it deals well with large datasets, and it prevents overfitting of data. The next step is to predict the values and create the confusion matrix which will decide whether the predicted value is correct or not. The last step of random forest is to visualize the result in an understandable form. For earthquake prediction we need to preprocess the dataset, fit it to a random forest algorithm, make a confusion matrix and finally we should visualize the result. The main applications of random forest are that it will work accurately in regression and classification problems. For sectors like banking, medicine, land use and marketing use random forest algorithms for better accuracy.

3.4 Logistic regression

Logistic regression is the machine learning algorithm which is specialized in the prediction process. This algorithm belongs to a supervised learning algorithm (i.e., labeled data or structured data) given as input. This algorithm works for predicting the value between a dependent variable. Logistic regression algorithm is a method for statistical analysis. Logistic regression algorithm predicts the output with the dependent variable comparing it with a relationship between one or more independent variables that are existing. The outcome value is a more straightforward result between the alternative cases. The result value will have only two possible results 0 or 1, yes or no and diseased or non-diseased. The result can be further done with feature scaling in need of more accurate results. In Logistic regression we will be fixing the logistic function in ‘S’ shape for predicting the two possible values (0,1). Logistic regression algorithm has the capability of providing probability and new data can be classified using continuous and discrete type datasets. Linear regression algorithm when plotted as a graph gives a straight line as shown in Fig.2. Logistic regression algorithm when plotted as a graph gives a curved shape ‘S’ as shown in Fig.3.



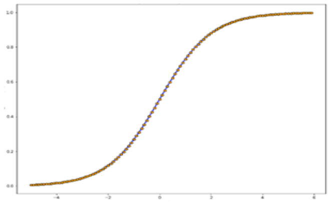


Fig2: Linear regression

Fig3: Logistic regression

As shown in above figures, we can see the difference between linear regression and logistic regression clearly. The ‘S’ shaped curve is also called a sigmoid function. This sigmoid function will be very useful in predicting the values to the probabilities. This curve easily maps any real value to any other value between the range 0 and 1. The value cannot go beyond the value of 0 and 1, thus ‘S’ shaped curve is formed in the graphical representation. And this graphical representation is known as the sigmoid function of the Logistic regression algorithm. In a logistic regression algorithm, there is a value called threshold value which acts as a center point

3.5 Light gbm algorithm

Light gbm (gradient boost machine) is a gradient boost algorithm that follows a set of rules that's primarily basedon the decision treemodel. By means that of computation it's thought about as a powerful algorithm.They are two types of techniques such asexclusive feature bundling (EFB) and gradient based one side sampling (GOSS) and e). This algorithm helps in reducing complexity of histogram building by down sampling data using EFB and GOSS. Light gbm algorithm grows vertically while all other algorithms grow horizontally. Light gbm chooses leaf-wise with large loss to grow while other algorithms choose level-wise. As the size of the data increases day by day all other algorithms computing speed is low. But in light gbm has high computational speed. 'Light' in light gbm denotes fast. Light gbm takes less memory to run and will be able to deal with a very large amount of data. There is more than 100+ numbers of parameters in light gbm documentation.

3.6 Extreme boost algorithm

Extreme Gradient Boosting (XGBoost) is an open-source library that comes under gradient enhancement algorithm which provides effective and efficient implementation. XG Boost algorithm comes under supervised algorithm which is based on decision tree and is effective when compared to all other algorithms. Even though there are many other algorithms available in machine learning XG Boost is considered as the fastest and efficient algorithm TheXGBoost algorithm has certain features such as it is highlyflexible, awareness of sparse data, implementation on both single and distributed systems and parallelization.For bagging, we need to give a large training set as input and those samples need to be bootstrapped. After these methods those samples should be classified, and prediction should be taken by the final ensemble classifier.in boosting original data is given as input in classifier and those data are changed to weighted data. Mathematically, we can write our model in the form

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \epsilon \mathcal{F}$$

4. RESULT AND DISCUSSION

Machine Learning algorithms used in this paper for the prediction of the earthquake disaster are Logistic Regression, XG Boosting and Light gbm. Comparing all these machine learning algorithms, the result is that the slight Gradient Boosting is faster than the others. Algorithm compared here, and while the XG boosting algorithm is little lag in the time, but the accuracy rate will be more when comparing the machine learning algorithms such as the logistic regression and light gbm with the extreme gradient boosting algorithm. And the Logistic regression algorithm gives the output in the binary form or in discrete form.

5. CONCLUSION

By comparing the three machine learning algorithms such as light gbm (gradient boost machine) XG Boost (Extreme gradient Boost) and logistic regression proves that lightweight gbm is faster and more efficient than XG Boost and logistic regression. XG Boost does not work well in probability problem which again may leads to false prediction While using Light Gbm increases efficiency and decreases memory usage.Light gbm provides better accuracy than XG Boost and logistic regression. Lightgbm uses leaf-wise technique while all other algorithm uses level-wise technique which makes Light gbm more effective. Logistic regression gives discrete values such as 0 and 1. logistic regression helps solve the classification problem and is subject to supervised learning.XGBoost comes under unstructured data and the main advantage is that it uses memory efficiently. Light gbm and XG Boost comes under decision tree while logistic regression comes under supervised learning. Summing up all three machine learning techniques it is evident that Light gbm gives better accuracy.

6. FUTURE WORK

Many algorithms are available in machine learning for the prediction process. Further we can also compare with other advanced algorithms in machine learning. Deep Learning also contributes many algorithms for the prediction model and prediction process. Thus, Deep Learning algorithms can also be implemented for the prediction process and can be compared with these algorithms and find the final algorithm which has highest accuracy rate. Deep Learning algorithms like Convolutional Neural Networks (CNN), Long-Short Term Memory Networks (LSTM), Recurrent Neural Networks (RNN) and other algorithms are effective in prediction process. Thus, these algorithms can be compared by implemented the algorithm in prediction model and find the accuracy rate.

REFERENCES

[1]T. Tang, D. Jiao, T. Chen, and G. Gui, "Medium- and Long-Term Precipitation Forecasting Method Based on Data Augmentation and Machine Learning Algorithms," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 15, pp. 1000-1011, 2022, Doi: 10.1109/JSTARS .2022.3140442.

[2] Sharmeen Binti Syazwan Lai, Nur Huda Nabihan Binti Md Shahri, Mazni Binti Mohamad, HezlinAryani Binti Abdul Rahman, Adzhar Bin Rambli , "Comparing the Performance of AdaBoost, XGBoost, and Logistic Regression for Imbalanced Data," Mathematics and Statistics, Vol. 9, No. 3, pp. 379 - 385, 2021. DOI: 10.13189/ms.2021.090320.

[3] M. R. Machado, S. Karray and I. T. de Sousa, "LightGBM: an Effective Decision Tree Gradient Boosting Method to Predict Customer Loyalty in the Finance Industry," 2019 14th International Conference on Computer Science & Education (ICCSE), 2019, pp. 1111-1116, doi: 10.1109/ICCSE.2019.8845529.

[4] L. Liu, "Research on Logistic Regression Algorithm of Breast Cancer Diagnose Data by Machine Learning," 2018 International Conference on Robots & Intelligent System (ICRIS), 2018, pp. 157-160, doi: 10.1109/ICRIS.2018.00049.

[5] X. Zhang, L. Liu, L. Xiao and J. Ji, "Comparison of Machine Learning Algorithms for Predicting Crime Hotspots," in IEEE Access, vol. 8, pp. 181302-181310, 2020, doi: 10.1109/ACCESS.2020.3028420.

[6] L. Li, Y. Wu, Y. Ou, Q. Li, Y. Zhou and D. Chen, "Research on machine learning algorithms and feature extraction for time series," 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), 2017, pp. 1-5, doi: 10.1109/PIMRC.2017.8292668.

[7] S. Ray, "A Quick Review of Machine Learning Algorithms," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 2019, pp. 35-39, doi: 10.1109/COMITCon.2019.8862451.

[8]V. Chamola, V. Hassija, S. Gupta, A. Goyal, M. Guizani and B. Sikdar, "Disaster and Pandemic Management Using Machine Learning: A Survey," in IEEE Internet of Things Journal, vol. 8, no. 21, pp. 16047-16071, 1 Nov.1, 2021, doi: 10.1109/JIOT.2020.3044966.

[9] P. M. Padmawar, A. S. Shinde, T. Z. Sayyed, S. K. Shinde and K. Moholkar, "Disaster Prediction System using Convolution Neural Network," 2019 International Conference on Communication and Electronics Systems (ICCES), 2019, pp. 808-812, doi: 10.1109/ICCES45898.2019.9002400.

[10]M. A. Kumar and A. J. Laxmi, "Machine Learning Based Intentional Islanding Algorithm for DERs in Disaster Management," in IEEE Access, vol. 9, pp. 85300-85309, 2021, doi: 10.1109/ACCESS.2021.3087914.

[11]W. Li, N. Narvekar, N. Nakshatra, N. Raut, B. Sirkeci and J. Gao, "Seismic Data Classification Using Machine Learning," 2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService), 2018, pp. 56-63, doi: 10.1109/BigDataService.2018.00017.

[12]R. Mallouhy, C. A. Jaoude, C. Guyeux and A. Makhoul, "Major earthquake event prediction using various machine learning algorithms," 2019 International Conference on Information and Communication Technologies for Disaster Management (ICT-DM), 2019, pp. 1-7, doi: 10.1109/ICT-DM47966.2019.9032983.

[13] R.Bharathi, T.Abirami,” Energy efficient compressive sensing with predictive model for IoT based medical data transmission”, Journal of Ambient Intelligence and Humanized Computing, November 2020, https://doi.org/10.1007/s12652-020-02670-z

[14] F. Chen et al., “Self-supervised data augmentation for person re-identification,” Neurocomputing, vol. 415, pp. 48–59, Nov. 2020, doi: 10.1016/j.neucom.2020.07.087.

[15]V. Nunavath and M. Goodwin, "The Use of Artificial Intelligence in Disaster Management - A Systematic Literature Review," 2019 International Conference on Information and Communication Technologies for Disaster Management (ICT-DM), 2019, pp. 1-8, doi:10.3390/su132212560.

[16] Z. Mushtaq and S. Su, “Environmental sound classification using a regularized deep convolutional neural network with data augmentation,” Appl. Acoust., vol. 167, Oct. 2020, doi:10.1016/j.apacoust.2020.107389.