EARTHQUAKE PREDICTION

Group 6, IIIT Guwahati

Abstract- Earthquake prediction is a branch of the science of seismology concerned with the specification of the time, location, and magnitude of future earthquakes within stated limits, and particularly "the determination of parameters for the next strong earthquake to occur in a region". Earthquake prediction is sometimes distinguished from earthquake forecasting, which can be defined as the probabilistic assessment of general earthquake hazard, including the frequency and magnitude of damaging earthquakes in a given area over years or decades. In this work, sixty seismic features are used which were previously computed using seismological concepts, such as Gutenberg-Richter law, seismic rate changes, foreshock frequency, seismic energy release, total recurrence time. Maximum Relevance and Minimum Redundancy (mRMR), PCA feature selection algorithms were used to extract the relevant features. Traditional Machine learning algorithms like Random Forest and XGBoost were used to classify and predict whether earthquake occurs with an applied threshold magnitude of 5.0. The obtained numerical results show improved prediction performance for all the considered regions, compared to previous prediction studies.

I. INTRODUCTION

Men-Andrin Meier, a seismologist at Caltech, says that his “best guess is that earthquakes are inherently unpredictable.” Many of the sounds and small movements along tectonic fault lines where earthquakes occur have long been thought to be meaningless. But machine learning—training computer algorithms to analyze large amounts of data to look for patterns or signals—suggests that some of the small seismic signals might matter after all. Earthquake prediction is necessary to undertake disaster preparedness measures, reducing the damage from the earthquakes. This requires that the accuracy of prediction be known, but, contrary to common belief, a timely prediction of low accuracy may be very useful. Earthquake prediction is necessary also for fundamental understanding of the dynamics of the lithosphere, particularly in the timescales of 102 years and less. So far, this problem is in the same stage as the theory of gravity was between T. Brahe and J. Kepler: the study of heuristic regularities that are necessary to develop a fundamental theory.

In this model, we used the dataset generated by Khawaja M.Asim, in which he calculated sixty on the principle of retaining maximum information. Hindukush, Chile and Southern California are three of the most active seismic regions in the world, thus considered in this model for studying problem of earthquake prediction. The obtained numerical results haven shown a huge improvement compared to the previous models proposed for these regions.

II. RELATED LITERATURE

This section reviews a number of publications where applications of machine learning methods to the task of earthquake prediction on various temporal and spatial intervals have been researched. An earthquake prediction system (EPS) named EPGPBoost was described in a paper, which was published in Soil Dynamics and Earthquake Engineering in 2018. This system is a classifier based on a combination of genetic programming (GP) and a boosting algorithm named AdaBoost. A total of 50 features were calculated, based on such geological concepts such as Gutenberg-Richter’s law, release of seismic energy, foreshock frequency, etc. The experiments have shown outstanding performance in all three observed regions both in terms of low false alarm ratio (the precision values were 74.3%, 80.2% and 84.2% for Hindukush, Chile and Southern California, respectively) and in terms of other metrics considered for evaluation, such as MCC and R score.

Another paper which uses the same USGS catalog was published in 2018. But this time, 60 seismic parameters were computed using various concepts of seismology. Again, some specific features were calculated via different approaches to retain the most complete information about the observed seismic zones. The proposed system is multistep, unlike previous other predictors proposed in literature which are mainly simple. The system is a combination of different machine learning algorithms, and on each step, one algorithm uses the knowledge obtained through learning of a previous one. Firstly, two-step feature selection is used to choose the most relevant parameters for training a model. Specifically, relevance and redundancy checks are performed (Minimum Redundancy Maximum Relevance criteria, denoted as mRMR, is applied). The resulting set of parameters is passed to a support vector regressor (SVR), and the trend predicted by SVR is then used as a part of input data for a hybrid neural network (HNN). The resulting values of performance measures (for instance, R score increased from 0.27 to 0.58 for Hindukush, from 0.344 to 0.603 for Chile, from 0.5107 to 0.623 for Southern California) showed that proposed multistep methodology improved prediction performance in comparison with individual machine learning techniques.

III. PROBLEM FORMULATION

In this project, earthquake prediction problem is designed/modeled as a binary classification problem. Every earthquake magnitude is converted to Yes, No (1, 0) through applying threshold on magnitude 5.0.The model is evaluated on various evaluation metrics such as sensitivity, specificity, positive predictive value or precision (P1), negative predictive value (P2), Matthews correlation coefficient (MCC) and R score.

IV. PROPOSED SOLUTION

Dataset obtained for all the three regions, is divided into training and testing sets. For training and validation purposes, 70% of the dataset is selected, while testing is performed on rest of 30% hold out dataset. The reason for separate training is that every region has different properties and can be classified tectonically into different categories, such as thrusting tectonics, strike-slip tectonics and so forth. Therefore every type of region possesses different behaviors and relations to the earthquakes. Thus separate training for every region is meant to learn and model the relationship between seismic features and earthquakes for that particular region.

The proposed methodology includes the use of two step feature selection process. The features are selected after performing relevancy and redundancy checks, to make sure that only useful features are employed for earthquake prediction. The selected sets of features are then passed to XGBoost Classifier. We have used XGBoost since it uses a more regularized model formalization to control over-fitting, which gives it better performance.

V. EXPERIMENTS

Initially Cut-off magnitude was applied on the dataset. Cut-off magnitude corresponds to the earthquake magnitude in the USGS catalog, above which USGS catalog is complete and no seismic event is missing. This depends upon the level of instrumentation. The cut-off magnitude for Southern California region is found to be less than 2.6, for Chile it is 3.4 and for Hindukush it is 4.0.Infinity and null values present in Chile and Southern California datasets were dropped from the dataset. Every earthquake magnitude is converted to Yes, No (1, 0) through applying threshold on magnitude 5.0.

The following sets of models were applied on the dataset: Logistic regression, Random Forest, XGBoost, PCA with Logistic regression, PCA with Random Forest, PCA with XGBoost, mRMR with XGBoost.

1) Chile Region

Model	Accuracy	MCC	Feature count
Logistic Regression	22.65	-0.07	60
Random Forest	94.25	0.83	60
XGBoost	95.54	0.87	60
PCA with Logistic regression	84.77	0.50	16
PCA with Random Forest	95.14	0.85	28
PCA with XGBoost	95.28	0.86	28
mRMR with XGBoost	96.3	0.89	23

2) Hindukush Region

Model	Accuracy	MCC	Feature count
Logistic Regression	72.84	0	60
Random Forest	87.28	0.66	60
XGBoost	88.30	0.69	60
PCA with Logistic regression	62.95	0.08	10
PCA with Random Forest	89.17	0.71	45
PCA with XGBoost	88.93	0.71	39
mRMR with XGBoost	89.58	0.72	18

3) Southern California :

Model	Accuracy	MCC	Feature count
Logistic Regression	84.65	0	60
Random Forest	96.01	0.84	60
XGBoost	97.72	0.91	60
PCA with Logistic regression	78.15	0.264	23
PCA with Random Forest	96.15	0.84	28
PCA with XGBoost	96.76	0.87	28
mRMR with XGBoost	98.29	0.93	15

VI. LIMITATIONS OF THE PROPOSED SOLUTION

- Null and Infinity values were dropped for Southern California and Chile regions.
- Average accuracies were calculated for only 10 simulations.

VII. FUTURE WORK

In future actual magnitude of the earthquake can be predicted by modeling it as a regression problem. Deep Learning methodologies can be applied.

VIII. CONCLUSION

Region	SA accuracy	Model accuracy	SA MCC	Model MCC
Chile	84.9	96.3	0.613	0.89
Hindukush	82.7	89.58	0.60	0.72
Southern California	90.6	98.29	0.72	0.93

SA – State of art ; Model – mRMR with XGBoost.

Logistic regression has performed very poor (for instance MCC score (Chile) - negative 0.07) due to multi-collinearity features which depicts there are redundant features in the data. So, we have selected useful features using both PCA and mRMR. . Later we passed maximum relevant features to XGBoost since it uses a more regularized model formalization to control over-fitting and gives a better performance. Gradient boosting (GB) methods are powerful classifiers that typically perform very well on structured data, and the XGBoost library is an excellent implementation of this algorithm. Hence, we have achieved better scores both in terms of accuracy and MCC scores.

REFERENCES

1. https://www.smithsonianmag.com/science-nature/could-machine-learning-be-key-earthquake-prediction-180972015/
2. https://www.kdnuggets.com/2019/05/xgboost-algorithm.html
3. https://www.ncbi.nlm.nih.gov/pmc/articles/pmc6033417/
4. https://machinelearningmastery.com/principal-components-analysis-for-dimensionality-reduction-in-python/
5. http://ceur-ws.org/vol-2372/seim_2019_paper_31.pdf
6. https://www.annualreviews.org/doi/full/10.1146/annurev.earth.30.100301.083856
7. https://github.com/jundongl/scikit-feature/blob/master/skfeature/function/information_theoretical_based/mrmr.py

AUTHORS

KANDADI VENKATA SHRAVAN - 1801081

ATMAKURU DHATRISH - 1801037

THIRUMURUGAN R - 1801185

HALAVATH VENU - 1801067