Performance Analysis of different Classifiers for Earthquake prediction: PACE

Arya P Menon[1], Abin Varghese[2], Joel P Joseph[3], Jofiya Sajan[4], Ninu Francis[5]
[1,2,3,4]*Student, Jyothi Engineering College*
[5]*Assistant Professor, Jyothi Engineering College*

***Abstract*— Earthquakes are catastrophic geo-hazards that endanger human life. Predicting the occurrence of earthquakes is very helpful to reduce the harmful effects. Therefore, a system to predict the forthcoming earthquakes and issues warning promptly are very appealing. There have been researches going on in the machine learning area to predict the earthquakes by the statistical methods based on the previous events recorded. However, the prediction of earthquakes suffers from the class imbalance problem as these events occur very rarely. This system is built to analyze the performance of various machine learning algorithms. The class imbalance problem of the data set is reduced using the resampling method. The system is trained using different algorithms namely: Support Vector Machine, K-Nearest Neighbour, Decision Tree, Logistic Regression and Naive Bayes. The performance is evaluated based on the values of accuracy, precision, recall, and f-measure. To increase the performance, k-fold cross-validation is implemented and performance is again evaluated. This cross-validation is carried out for three different values of k such as 5, 10 and 15. The system is evaluated with both class imbalance problem prevailing dataset and class imbalance problem resolved dataset. The performance is plotted and the optimum value of k for k-fold cross validation is found out. It also identifies which classifier is best for the prediction of earthquake.**

***Index Terms*— Decision Tree, Earthquake, K-fold cross-validation, K-Nearest Neighbour, Logistic Regression, Machine Learning, Naive Bayes, Support Vector Machine**

I. INTRODUCTION

Human faces many natural disasters like flood, earthquake, landslide and volcano in their life. These disasters cause great loss to human life. The main issue with these disasters is that they are unable to correctly predict. Investigations are going on in predicting these disasters based on the previously occurred events. Earthquakes are one of the major catastrophic geohazards and their unpredictability causes severe destruction in human life. Earthquakes are results of the sudden release of energy in the Earth's crust. This results in the shaking of earth which is named as the earthquake. This also creates elastic energy waves known as seismic waves. PACE is based on the quantitative earthquake dataset and the use of machine learning algorithms for differentiating the hazardous and non-hazardous region. Supervised learning technique is employed as earthquake prediction is a classification problem. Algorithms used for the study are SVM, Naive Bayes, K-Nearest Neighbour, Logistic Regression and Decision Tree. Even though logistic regression is considered as a regression algorithm, its output will be either 0 or 1. Thus it can be used for classification problem. Each algorithm will classify the data into the hazardous region or nonhazardous region. The splitting of the dataset into the training set and the test is done using the sampling method and kfold cross validation. Firstly the system is trained with the imbalanced dataset and then with the balanced dataset. Finally, the performance is evaluated based on accuracy, precision, recall and f-measure and the best classifier for the earthquake prediction problem is identified

K-fold cross-validation is carried out for three different values of k such as 5, 10, and 15. All the performance results are plotted and the optimum value of k for k-fold cross validation is also identified.

II. LITERATURE SURVEY

The literature review includes papers which covers almost all aspects of earthquake detection. The details of some papers are given here:

Reference [2] highlights the evaluation of different data mining algorithms to predict earthquakes. The purpose of this paper is to study and evaluate some of the data mining algorithms in terms of accuracy and computational time. These data mining algorithms include Artificial Neural Network, Decision Tree, Support Vector Machine (SVM) and Naive Bayes. Using the real data in the Rapid Miner platform, simulation results show that the Support Vector Machine (SVM) is the fastest algorithm but it has the lowest accuracy and multilayer perceptron has the highest accuracy. The drawbacks of this system are that the algorithms are evaluated based on the accuracy and computational time. Also, it uses the sampling method for dividing the data set into the training set and test set. Accuracy alone cannot determine the efficiency of the system and the sampling method cannot divide the data set effectively into the training set and test set.

Reference [3] uses the Extreme Learning Machine (ELM) in the modelling stage. ELM has a greater generalization performance than feed-forward networks learning by back-propagation algorithms and has the ability to learn quickly. 10-fold cross-validation is used to create the training set and test set. Three different activation functions are also used. The developed methodology exhibits high accuracy but low recall value which can be stated as its drawback. Since the earthquake detection is very important, the prediction system should have high value for recall.

Reference [4] proposes a novel method to improve the accuracy of Naive Bayes classifier. The author states that the assumption of conditional independence can be a reason for the loss of accuracy. Sometimes the factors may have some dependence among them which is not taken into account in Naive Bayes Classifier. The experimental results show that the suggested method exhibits good performance in increasing the accuracy of Naive Bayes classifier that the traditional Naive Bayes classifier. Though the proposed Naive Bayes classifier exhibits high accuracy, it cannot be considered as an effective model. Since the sampling method is used, the number of positive class data instances in the test set and the predicted class of those data instances are unknown. If their number is very low, the system will exhibit a better accuracy even if they are classified incorrectly. Thus, considering accuracy alone cannot be a good measure for analyzing performance.

III. PROPOSED SYSTEM

In general, earthquake is a word utilized to denote a seismic event which generates the elastic waves known as seismic waves. As earthquakes are destructive, a system to provide timely warning is a need of the hour. Thus this proposed system analyses which machine learning algorithm is effective for this problem statement. Performance of five different algorithms is also analyzed. Evaluation of algorithms is based on the values of accuracy, precision, recall and f-measure. More details about the implementation are discussed in the next section.

IV. METHODOLOGY

This system is made with a prime focus of finding the best classifier for the earthquake prediction problem. As mentioned earlier, 5 different algorithms are used to train the system and performance is evaluated based on 4 different values. This experiment is carried out on both imbalanced and balanced dataset. The different modules of the system are as follows:

A. Data Preprocessing

The efficiency of the system primarily depends on the dataset. The dataset used here consists of 18 features and 2584 samples [5]. Once all the relevant data are collected, then these data should undergo through the preprocessing stage. In this system, the preprocessing is applied to those data which are not numerical. That is, all the non-numerical values of features are converted to numerical values. Then the dataset is standardized. After the first stage of analysis, the class imbalance problem is resolved by resampling the data of the minor class. This also comes under the data preprocessing. The dataset is then split into the training set and test set by sampling method and k-fold cross-validation.

B. Training

After the data has been preprocessed, the system is trained using five different algorithms along with the training set. The different algorithms used are Support Vector Machine (SVM), K-Nearest

Neighbour (KNN), Decision Tree(DT), Logistic Regression(LR) and Naive Bayes (NB).

Support Vector Machine (SVM):
The purpose of the support vector machine algorithm is to seek out a hyperplane in N-dimensional space(N is the number of features, here N=18) that distinctly classifies the input points. There would be many hyperplanes present which can separate the two classes of data points. After the training, SVM returns the plane that has the maximum margin, i.e the plane with the maximum distance between data points of both classes. Maximizing the margin distance provides some assurance that future data points are often classified correctly.

K-Nearest Neighbour (KNN):
The k-nearest neighbour algorithm (KNN) is a method used for both classification and regression. The input consists of the k nearest training examples inside the feature space. The output depends on whether KNN is employed for classification or regression. In KNN classification, the output is the class label of the unknown tuple. The class label of the unknown tuple is assessed by the class labels of its k nearest neighbours. The class label with the majority among the k nearest neighbours is assigned as the class label of the unknown tuple.

Decision Tree (DT):
Decision tree learning is one of the predictive modelling methods employed in statistics, data mining and machine learning. It utilizes a decision tree (as a predictive model) to go from observations about an unknown tuple (depicted in the branches) to conclusions about the tuple's target value, that is the class label.

Logistic Regression (LR):
Logistic regression is a statistical model that in its elementary form uses a logistic function to model a binary variable, although numerous complicated extensions exist. In multivariate analysis, logistic regression is about evaluating the parameters of a logistic model (a sort of binary regression). Mathematically, a binary logistic model estimates a variable with two possible values"0" and"1". Here"0" denotes the non-hazardous region and"1" denotes the hazardous region.

Naive Bayes (NB):
Naive Bayes classifier is a probabilistic machine learning model used for the classification problem. The crux of the classifier is based on the Bayes theorem. This algorithm works on the assumption that the features are independent. That is the value of one feature does not affect any other feature. Hence it is called naive.
Each of the algorithm returns a model. Based on this model, the system is tested and performance is evaluated. Firstly the system is trained using the training set created using sampling method followed by k-fold cross validation. This is done on both balanced and imbalanced dataset.

C. Testing
In this module, the model of each algorithm got as the result of training is tested using the test set. Firstly, the test set created using sampling method is used and then by k-fold cross-validation. In the next stage, the same process is carried out for the class imbalanced dataset.

D. Performance Analysis
This is the last module. After the system is tested, in each case and for each algorithm, the values of accuracy, precision, recall and f-measure are found. Accuracy is measured in percentage. For precision, recall and f-measure, the values will be between 0 and 1, where 0 denotes the worst value and 1 denotes the best value.

V. RESULTS AND DISCUSSIONS

The Accuracy, Precision, Recall and F-measure are used for evaluating the performance of the system. The formulas for calculating them are shown below.

Accuracy can be defined as the percentage of correctly classified instances.
Accuracy = (True Positive + True Negative) / (True Positive + True Negative + False Positive + False Negative)
Precision is defined as the fraction of the relevant instances over the retrieved instances as true.
Precision = True Positive / (True positive + False Positive)

Recall is defined as the fraction of the relevant instances over the total relevant instances.

Recall = True Positive / (True positive + False Negative)

F1-score is based on both the precision and recall values. It is defined as the harmonic mean of precision and recall.

F1-score = (2 * Precision * Recall) / (Precision + Recall)

Both precision and recall are very important for a model because of which F1-score is used as a standard measure for the model comparison

Results of the study conducted are as follows:

Table I: Sampling Method (Imbalanced Dataset)

Different Classifiers	Performance Measures			
	Accuracy	*Precision*	*Recall*	*F1 score*
SVM	92.07%	0.2	0.03	0.05
KNN	92.65%	0.0	0.0	0.0
DT	91.10%	0.21	0.08	0.12
LR	92.26%	0.0	0.0	0.0
NB	85.49%	0.14	0.18	0.16

Table II: 5-Fold Cross Validation (Imbalanced Dataset)

Different Classifiers	Performance Measures			
	Accuracy	*Precision*	*Recall*	*F1 score*
SVM	92.80%	0.14	0.02	0.03
KNN	93.42%	0.0	0.0	0.0
DT	90.98%	0.10	0.04	0.06
LR	93.30%	0.3	0.01	0.02
NB	86.84%	0.14	0.19	0.16

Table III: 10-Fold Cross Validation (Imbalanced Dataset)

Different Classifiers	Performance Measures			
	Accuracy	*Precision*	*Recall*	*F1 score*
SVM	92.73%	0.10	0.02	0.03
KNN	93.42%	0.0	0.0	0.0
DT	90.87%	0.11	0.04	0.06
LR	93.27%	0.15	0.01	0.02
NB	86.07%	0.12	0.19	0.14

Table IV: 15-Fold Cross Validation (Imbalanced Dataset)

Different Classifiers	Performance Measures			
	Accuracy	*Precision*	*Recall*	*F1 score*
SVM	92.65%	0.09	0.01	0.02
KNN	93.42%	0.0	0.0	0.0
DT	90.48%	0.06	0.03	0.03
LR	93.23%	0.13	0.01	0.02
NB	85.92%	0.12	0.17	0.14

When all the results are analyzed, it is seen that all the classifiers exhibit a high accuracy. Even when the sampling method is used, the classifiers perform with high accuracy. But it is analyzed that, the values of precision, recall and f-measure are very low. Therefore, no classifier can be selected as the best classifier for earthquake prediction at this stage. When this behaviour of classifiers is analyzed, it is found that the imbalance in the dataset can be the reason for low values of precision, recall and f-measure. Thus the dataset is balanced and the performance of the classifiers is analyzed again.

Table V: Sampling Method (Balanced Dataset)

Different Classifiers	Performance Measures			
	Accuracy	*Precision*	*Recall*	*F1 score*
SVM	93.06%	0.88	1.0	0.94
KNN	51.66%	0.52	1.0	0.68
DT	97.10%	0.95	1.0	0.97
LR	93.06%	0.88	1.0	0.94
NB	96.58%	0.94	1.0	0.97

Table VI: 5-Fold Cross Validation (Balanced Dataset)

Different Classifiers	Performance Measures			
	Accuracy	*Precision*	*Recall*	*F1 score*
SVM	92.54%	0.87	1.0	0.93
KNN	50.12%	0.50	1.0	0.67
DT	95.63%	0.92	1.0	0.96
LR	92.61%	0.87	1.0	0.93
NB	95.61%	0.92	1.0	0.96

Table VII: 10-Fold Cross Validation (Balanced Dataset)

Different Classifiers	Performance Measures			
	Accuracy	*Precision*	*Recall*	*F1 score*
SVM	93.70%	0.89	1.0	0.94
KNN	50.17%	0.50	1.0	0.67
DT	96.12%	0.93	1.0	0.96
LR	93.85%	0.89	1.0	0.94
NB	96.12%	0.93	1.0	0.96

Table VIII: 15-Fold Cross Validation (Balanced Dataset)

Different Classifiers	Performance Measures			
	Accuracy	*Precision*	*Recall*	*F1 score*
SVM	93.91%	0.89	1.0	0.94
KNN	50.19%	0.50	1.0	0.67
DT	96.02%	0.93	1.0	0.96
LR	93.91%	0.89	1.0	0.94
NB	96.27%	0.93	1.0	0.96

When all the results are analyzed after balancing the dataset, it is seen that all the classifiers exhibit high performance even when the sampling method is used. It is analyzed that, the value of recall is returned as 1. Though 1 is the best value, it should be verified that the classifier is not always returning the positive class to the data points. This is checked with the help of the confusion matrix and found that no classifier always returns the positive class.

All the classifiers have satisfactory performance and best among them are Decision tree and Naive Bayes. KNN has comparatively low performance in terms of accuracy, precision and f-measure

VI. CONCLUSION

The system is trained with five different algorithms; SVM, NB, KNN, DT, and LR. After the training, the system returns a model for each algorithm. Algorithms are then tested using their returned model and test set. The training set and test set are created using the sampling method as well as the K-fold cross validation. System performance is evaluated using the values of accuracy, precision, recall, and F-measure. The performance analysis is done for both imbalanced and balanced datasets. Classifiers showed poor performance with the imbalanced dataset and good performance with the balanced dataset. The main objective of this project was to identify the best classifier for earthquake prediction problem. Based on the analysis, the algorithms which exhibit high performance are identified as DT and NB. Another objective was to identify the optimum value of k for k-fold cross validation among the three values considered and it is identified as 15.

In future, the system can be developed to identify the features that have greater influence in predicting an earthquake. This can be done using feature selection algorithms. The data set can be strengthened by obtaining data from various coals and fields. This system mainly focuses on the detection of earthquakes in coal mines. Hence, the system can be extended to identify all the other earthquakes and natural disasters like floods, landslides, etc. in the same manner. In addition to that, the performance of the system can be increased by using the ensemble algorithm. The performance of KNN algorithm and Naive Bayes classifier can be analyzed further by taking different values for k and different node splitting criteria respectively.

VII. ACKNOWLEDGMENT

With great respect, we extend our heartfelt gratitude and indebtedness to Mr. Unnikrishnan P, Assistant Professor of Computer science and Engineering Department in Jyothi Engineering College, for his continues support and assistance to complete this system.

REFERENCES

[1] Khawaja M. Asim, Adnan Idris, Talat Iqbal and Francisco Martinez-Alvarez, Earthquake prediction model using support vector regressor and hybrid neural networks, PLoS One, vol. 13(7), pp. 529–551, July 2018.

[2] Asadollah Shahbahrami and Zinat Mehdidoust Jalali, Evaluation of Different Data Mining Algorithms to Predict Earthquakes Using Seismic Hazard Data, J. Appl. Environ. Biol. Sci., 7(2)142-150, 2017.

[3] Musa Peker, Seismic Hazard Prediction Using Seismic Bumps: A Data Mining Approach, vol. 5, Issue-4, pp-106-111, 2016

[4] Kalyan Netti and Dr. Y Radhika, Minimizing Loss of Accuracy for Seismic Hazard Prediction using Naive Bayes Classifier, vol. 3 Issue: 04, April,2016.

[5] UCI Machine Learning Repository – Seismic Bumps data.

[6] C P Shabariram and K E Kannammal, Earthquake Prediction Using Map Reduce Framework, International Conference on Computer Communication and Informatics, 2017

[7] Jhon Veri and Teh Ying Wah, Earthquake Prediction based on the Pattern of Points Seismic Motion, International Conference on Advanced Computer Science Application and Technologies,

[8] Jozef Kabiesz, Beata Sikora, Marek Sikora and Łukasz Wrobel, Application of Rule-Based Models for Seismic Hazard Prediction in Coal Mines, Acta Montanistica Slovaca,

[9] https://en.wikipedia.org/wiki/Earthquake

[10] https://scikit-learn.org/stable/