International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 11
November 2017

Prediction of Seismic Activities in Coal Mines using Machine Learning

Vedangi Godse[1], Karishma Kothari[2] ,Vaibhavi Nandode[3],Shrutika Ganesh[4] ,Nihar Ranjan[5]
Sinhgad Institute Of Technology And Science,Narhe,Pune.

ABSTRACT*-Seismic activities occurring in coal mines has constantly been a reason for loss of live as well as loss of property. Hence there arises a need to have monitoring and predicting system. This can be accomplished by identifying, planning and predicting seismic events for a safe and smooth working environment in coal mines. This paper proposes a solution based on an ensemble of classifiers that include Random Forest Tree and Extremely Randomized Trees. Dataset accessed by our system includes features recorded over a period of time. The features generated are in the form of time series data. The data collected can be represented as train datasets and test datasets. An ensemble of classifiers worked on the training datasets helps for extraction from training dataset. In this system, the ensemble of classifiers keeps intact both the quality and the differing qualities of individual models. Furthermore, we provide an all-around software framework to work with user interface model which with the ensemble of classifiers will include both the quality and user-friendliness of individual models.*

KEYWORDS: Feature generation, feature extraction, training dataset,test dataset, ensemble of classifier

I. INTRODUCTION

The coal mining activities are one of the major causes for seismic activities. The coal miners underground are endangered from sudden effects of seismic activities. These are mining induced seismic activities which changes energy level of earth's crust. The Poland industry in 2015 reported 2158 dangerous incidents with 19 causalities and 12 major injuries [1].There is a need to have a system that warns the coal miners about the danger and ensure security. A predicting system would play a major role in monitoring and preventing further dangers.

The increasing technologies like more and more sensors being deployed, number of classification model for prediction, data mining and analytics technology and large amount of data storage being available at less cost. In this paper we suggest a prediction system to predict the seismic activities from the dataset consisting of attributes such as seismic energy, total energy, seismic bumps and shift. Various algorithms like naïve bayes, neural networks, feature engineering, ensemble of classifiers have been suggested for training the dataset and testing. The major challenge is to obtain accuracy in prediction with considerable time complexity. We can say that the mining induced seismicity can be a mini model depicting earthquakes. Thus we can say that the proposed prediction system can be used for further earthquake prediction as well.

The paper is further outlined as first section describes problem statement and input dataset, second section discusses the about related work, third section gives the proposed flow system and fourth section gives concludes the paper.

II.PROBLEM STATEMENT AND INPUT DATASET

As per the AAIA'16 Data Mining Challenge the goal is to propose a system that is reliable for prediction of seismic activities with the provided dataset [1],[2],[5],[8].The task was basically a classification based task with time series data .It demanded prediction of energy level to warn the coal miners. The dataset provided in the competition was hourly aggregated multivariate time series data. The dataset is available from the sensors deployed in coal mines. Similar type of dataset with following attributes has been considered as input dataset for our system [7]:

seismic method (a - lack of hazard, b - low hazard, c - high hazard, d - danger state)

Seismoacoustic,shift,genergy,gpuls,gdenergy
gdpuls,ghazard,nbumps,nbumps2,nbumps3,nbumps

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 11
November 2017

4,nbumps5,nbumps6,Nbumps7, nbumps89,Energy,maxenergy

The dataset consists of summary about seismicity after every 8hours i.e one shift. The dataset may contain negative values along with positive value and also may have null values. The following screenshot gives the value of dataset [6]:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Seismic	genergy	gpuls	gdenergy	gdpuls	ghazard	nbumps	nbumps2	nbumps3	nbumps4	energy	maxenergy	class
2	1	15180	48	-72	-72	1	0	0	0	0	0	0	0
3	1	14720	33	-70	-79	1	1	0	1	0	2000	2000	0
4	1	8050	30	-81	-78	1	0	0	0	0	0	0	0
5	1	28820	171	-23	40	1	1	0	1	0	3000	3000	0
6	1	12640	57	-63	-52	1	0	0	0	0	0	0	0
7	1	63760	195	-73	-65	1	0	0	0	0	0	0	0
8	1	207930	614	-6	18	1	2	2	0	0	1000	700	0
9	1	48990	194	-27	-3	1	1	0	1	0	4000	4000	0
10	1	100190	303	54	52	1	0	0	0	0	0	0	0
11	1	247620	675	4	25	1	1	1	0	0	500	500	0
12	1	41950	135	-39	-36	1	1	0	1	0	5000	5000	0
13	1	53250	140	-19	-31	1	0	0	0	0	0	0	0
14	1	166180	448	-30	-19	1	1	1	0	0	400	400	0
15	1	64540	215	0	9	1	1	0	1	0	6000	6000	0
16	1	75670	207	18	4	1	1	1	0	0	600	600	0
17	1	225040	575	-3	5	1	2	0	2	0	7000	6000	0
18	1	31290	117	-52	-41	1	0	0	0	0	0	0	0
19	1	22890	81	-63	-58	1	0	0	0	0	0	0	0
20	1	13170	59	-77	-67	1	0	0	0	0	0	0	0
21	1	8900	38	-83	-77	1	0	0	0	0	0	0	0
22	1	13250	32	-71	-79	1	0	0	0	0	0	0	0
23	1	11710	70	-72	-47	1	1	0	1	0	4000	4000	0

seismic-1

Fig 2.1 Screenshot of input dataset

III.RELATED WORK

Monitoring and predicting system for coal mines could play animportant role in limiting the seismic hazards. In [1] they propose a system that generates features from input dataset, features are selected and time series generation algorithm is used. The features generated are basic statistics that include minimum, maximum, arithmetic mean, harmonic mean, skewness etc. The time series generation algorithm used are first derivatives of original series, amplitudes, frequencies and magnitudes obtained from Fast Fourier transforms delta series that will help remove variation in the data. Learning algorithms like random forest, feature generation and heuristics give prediction based on different feature subset. The framework is implemented in Python on top of scikit-learn library. The main advantage was drift distribution that helped for reduction in feature selection. To achieve robustness 5 prediction models have been used where 4 are ERT and 1 is logistic regression. The performance was significant though it used automatic feature engineering.

The [2] suggests a versatile solution using Deep Recurrent Neural Network with Long Short-term memory cells that require no feature engineering. The main advantage of RNN is that is capable of approximating arbitrary values and mapping them into sequence. But the drawback of RNN is vanishing gradient for smaller values .Hence they use Long Short-term Memory that has self-connected memory cells and 3 gates like input, output and forget. The preprocessing includes data normalization that optimizes loss function and makes easy regularization. The model is initialized with values between 0.1 to -0.1 and 5 fold cross validation model is used. The main aim is to use ensemble of logistic regression and RNN using rank of record. Thus, the solution does not depend on feature and gives good performance.

A system with feature extraction and selection framework designed with MapReduce model with ensemble of classifiers like Support Vector Machine, Regression tree has been proposed[3].The model starts with input phase where csv file is original dataset HDFS chunks are formed by splitting data to parallelize feature extraction. The map phase processes each row and multiple features are extracted. After reduce phase the significant and diverse objects and objects are passed to ensemble classifiers.A wide variety of training classifiers like SVM, Regression tree, GLM are applied over the subset of reducts.Again ensembles of filtered regression models are applied finally.The entire implementation is done in R-programming environment.

Predicting seismic activities based on transient activity features along with average indicators evaluated by a Fisher's linear discriminant analysis with maintaining low complexity [4].The method evaluates average risk parameters along with risk assessment values. The training data is labeled either warning or normal for individual samples. The Fisher's Linear Discriminant Analysis adds new hourly aggregated values to average values and risk assessment measures.

The [5] exploits selective Naïve Bayes classifier along with optimal preprocessing, variable selection and model averaging with automatic variable construction.Variables are preprocessed using supervised value grouping. Bayesian model selection is used to select best subset of variable. Along with this efficient search heuristics are used for greedy forward addition and backward elimination of values.Compression base model averaging is done instead of considering best subset. The automatic variables are constructed in form of data structure where there is root table with secondary tables that solve problem of over fitting

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 11
November 2017

as these are virtual infinite table. Resampling was proposed to remove drift between training and test dataset and achieve approximately same sampling rate. Applying this methodology 100 classifiers are trained, each exploiting subsamples of training instance.

Naïve Bayes Classifier along with negation handling is one of the suggested models. The negation handling is done for data cleaning and also that increases accuracy of the Naïve Bayes [6].Different smoothing techniques like Diriclet smoothing, Absolute Discounting smoothing for negation handling.The numeric attributes are taken as input and its value are Gaussian distributed which maybe mean standard deviation and Gaussian distribution. Accuracy of Naïve Bayes Classifier with negation handling is 76.983% while without negation handling is 64.5%.The paper tries to outperform the MATLAB Naïve Bayes classifier algorithm.

Thus we can conclude that device to device communication can be implemented using the 3GPP standardization for 5G network. The implementation of this enhanced network will increase the overall performance of the network .Applications of business market, social networking, online gaming, video streaming, IoT services all will experience faster output. Low latency, better spectrum frequency, higher network capacity are all advantages in new network. The main advantage of Device to Device communication that user can easily communicate with other users.All techniques have their pros and cons. The technique is more costly and also the existing mobile need to be adaptable to the changes.

IV.PROPOSED SYSTEM

We suggest a web based application that uses machine learning algorithm to train the input dataset and use the test dataset for prediction.

We provide an interface to login to the software for the coal miners. The software checks for authenticated user. The coal mine consist of seismic sensors deployed that register bumps and energy values. These values are taken as input dataset. The input dataset undergoes three processes to give prediction:

A. Feature Generation

B. Time series generation

C. Ensemble of Classifiers

D. Feature Extraction

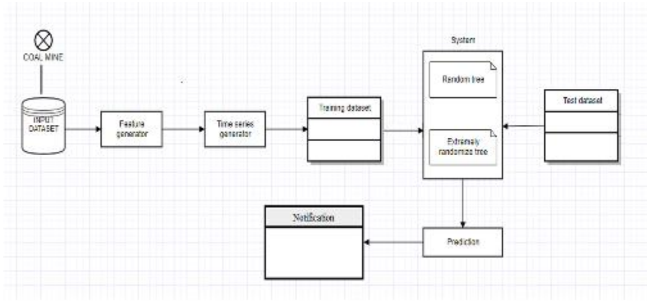

Fig 4.1: Proposed architecture

.

A. Feature Generation

The feature generation consists of two important goals. First the raw and unstructured data must be transformed to features such as max, min, average, harmonic mean, arithmetic mean, variance, standard deviation, etc. The second goal is to improve the accuracy of predictor which is obtained during feature selection. During feature selection there is need to check no extra features are generated. Also the features independently do not give that much indication as that given when correlated to one another. For example in our input dataset the attributes nbump and energy together as feature correlated would give much strong prediction about any seismic hazards.

B. Time series Generation

Time series is a sequence of observations

$s_t \epsilon R$

usually ordered in time. The time series data is data considered at equal interval of time forming a discrete time data. This data needs to be specified in a formatted pattern or cycles. We can also combine generators for faster generators.

Let an observed discrete univariate time series be s1,s2……..sT. This means we ‘T’ numbers which are observations on some variable made at ‘T’ equally distant time points, which for convince we label 1,2,….T.

A general model for the time series can be written as,

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 11
November 2017

st=g(t)+Φt   t=1,2,3….T

where g(t):deterministic function of time

and Φt: Residual term called as noise which follows a probability law.

C. Ensemble of classifiers

The combining of classifiers give much better performance than individual classifiers.We suggest a combination of Random Forest and Extremely Randomized tree that would form decision trees based on the values forming class.This algorithm provides over fitting hence unstructured data also forms deep tree.The Extremely Randomized Tree algorithm gives much smoother boundary to take decision.They are also much faster.Thus when both ensembles give much faster performance. The pseudo code for randomized tree can be given as:

1. Randomly select “k” features from total “m” features where k << m
2. Among the “k” features, calculate the node “d” using the best split point.
3. Split the node into daughter nodes using the best split.
4. Repeat 1 to 3 steps until “l” number of nodes has been reached.
5. Build forest by repeating steps 1 to 4 for “n” number times to create “n” number of trees.

To perform prediction using the trained random forest algorithm uses the below pseudocode.

1. Takes the test features and use the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome (target)
2. Calculate the votes for each predicted target.
3. Consider the high voted predicted target as the final prediction from the random forest algorithm

D. Feature Extraction

After the feature generation and selection, when the training dataset is processed under the ensemble of classifiers, test dataset is passed for feature extraction and determining the appropriate decision. The feature selection plays an important role for prediction using the test dataset.

The obtained result is displayed to the user in form of notification that would alert user if it is high danger else notify for low danger activities.

V.CONCLUSION

Thus the goal was topredict seismic activities from the input dataset with more accuracy and give warning alerts to coal miners for safety purpose. The proposed software framework gives a user friendly interface for coal miners that would either alert them or notify about dangers on basis of predictor system.

VI.REFERENCE

[1] Zdravevski, E., Lameski, P., & Kulakov, A. (2016). Automatic feature engineering for prediction of dangerous seismic activities in coal mines. *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, FedCSIS 2016*, *8*, 245–248. https://doi.org/10.15439/2016F152

[2] Kurach, K., & Pawlowski, K. (2016). Predicting dangerous seismic activity with Recurrent Neural Networks. *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, FedCSIS 2016*, *8*, 239–243. https://doi.org/10.15439/2016F134

[3] Grzegorowski, M. (2016). Massively Parallel Feature Extraction Framework Application in Predicting Dangerous Seismic Events, *8*, 225–229. https://doi.org/10.15439/2016F90

[4] Güzel, B. E. K., & Karaçalı, B. (2016). Fisher’s Linear Discriminant Analysis Based Prediction using Transient Features of Seismic Events in Coal Mines. *Proceedings of the Federated Conference on Computer Science and Information Systems*, *8*, 231–234. https://doi.org/10.15439/2016F116

[5] Boullé, M. (2016). Predicting Dangerous Seismic Events in Coal Mines under Distribution Drift,*8* 221–224. https://doi.org/10.15439/2016F21

[6]Netti, K., & Radhika, Y. (2016). An efficient Na??ve Bayes classifier with negation handling for seismic hazard prediction. *Proceedings of the 10th International Conference on Intelligent Systems and Control, ISCO 2016*. https://doi.org/10.1109/ISCO.2016.7726906

[7] Marek Sikora^{1,2} (marek.sikora '@' polsl.pl), Lukasz Wrobel^{1} (lukasz.wrobel '@' polsl.pl) (1) Institute of Computer Science, Silesian University of Technology, 44-100 Gliwice, Poland (2) Institute of Innovative Technologies EMAG, 40-189 Katowice, Poland(input data set)

[8] Bogucki, R., Lasek, J., Milczek, J. K., & Tadeusiak, M. (2016). Early Warning System for Seismic Events in Coal Mines Using Machine Learning, *8*, 213–220. https://doi.org/10.15439/2016F420

[9]Bontempi, G., & Group, M. L. (2013). MachineLearning Strategies for Time Series Prediction.