

Minimizing Loss of Accuracy for Seismic Hazard Prediction using Naive Bayes Classifier

Kalyan Netti¹, Dr. Y Radhika²

¹Senior Scientist, NGRI, Hyderabad

²Associate Professor, GITAM University, Visakhapatnam

Abstract - Classification is the one of the most important techniques in Datamining for data analysis. In Datamining, different Classification Techniques are available to predict the outcome for a given dataset. There are many classification methods for predicting and estimating accuracy; one such famous method is Naive Bayes Classifier. Naive Bayes is very popular as it is easy to build, however to the assumption of conditional independence among predictor's results in loss of accuracy. In this paper, we propose a technique to minimize loss of accuracy when predicting Seismic Hazard Activity. Hazard indicates a possible threat to life, health, property and environment. Mitigation of hazard when crossing stipulated level is paramount; otherwise, it may lead to an emergency. One of the most dangerous hazards in mining activities is Mining Hazard. Mineral, Diamonds/Gold and Coal exploration involves mining in a big way where hazard occurrence is quite common and addressing this mining hazard are a challenging task. A substantial threat of Mining Hazard is Seismic Hazard which is normal in underground mines. Thus, Predicting Seismic Hazard is one of the most important aspects in countering Mining Hazards. In this paper, the authors are proposing a novel method for minimizing loss of accuracy in Naive Bayes Classifier. The proposed novel technique used in NBC gave better accuracy even with Conditional Independence

Key Words: Data Mining, Classification, Naive Bayes Classifier, Conditional Independence, Accuracy, Hazard, Seismic Hazard

1. INTRODUCTION

Data Mining is a process of extracting useful and relevant information from data [1]. There are many techniques in Data Mining to extract information from data. With different advanced technologies employed in the areas of engineering, finance, health, etc., the data collected/accumulated, resulting in the exploration, is increasing exponentially. Now, with all the massive amounts of data available, the primary task is to understand the data and extract knowledge from the data. In the current scenario, obtaining useful information from massive amounts of data is a complex task and need very efficient algorithms/techniques. This area is explored in a big way by employing new processes, methods along with statistical techniques. One such effective technique in Data Mining is Classification. There are many

Classification techniques available; like Bayesian Networks, Decision Trees, Nearest Neighbour, and Neural Networks. In general, classification is one of the analysis techniques, used to derive models by bringing in prior observations to predict the outcome. Naive Bayes classifier (NBC) is a very modern and efficient method in data classification. Naive Bayes, a Supervised Classification Technique, is an effective one because it is easy to build, computationally not sophisticated and is capable of handling massive datasets. Moreover, Naive Bayes Classifier performs well compared to other predictive models as it assumes conditional Independence among predictors [2][4][7]. One of the main reasons for better performance of Naive Bayes Classifier is the assumption of independence among predictors. This very assumption of Independence sometimes leads to loss of accuracy in Naive Bayes Classifier. The loss of accuracy can be more when data sets have attributes with strong inter-relation among themselves. Thus, improving Naive Bayes classifier with the assumption of Independence among predictors is a challenging task [4] [5].

In this paper, seismic hazard data downloaded from UCI repository[6] is chosen to estimate accuracy hazard occurrence using NBC. There is an urgent need to mitigate seismic hazard event and classification techniques may address this issue. One of the most dangerous hazards is Mining Hazard who is common in mining activities. A substantial threat of Mining Hazard is Seismic Hazards which is normal in underground mines. Thus, Predicting Seismic Hazard is one of the most important aspects in countering Mining Hazards and any loss of accuracy in the algorithm used will result in wrong estimation of hazard.

In this paper, the authors present a novel method to minimize this loss of accuracy in Naive Bayes Classifier due to the assumption of Independence among predictors [2][3]. The experimental results show that the proposed method performed well and improved the accuracy when compared to the traditional Naive Bayes Classifier.

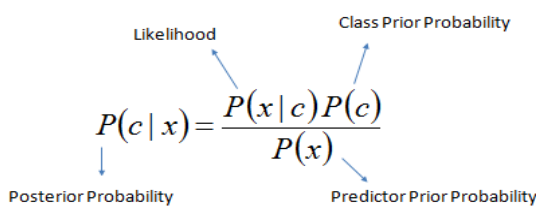
The next Section i.e. Section-II, discusses the Naive Bayes Classifier, Data Set is explained in Section-III; Implementation is presented in Section - IV, Section-V discuss the results and the last section presents the conclusions.

2. NAIVE BAYES CLASSIFIER

The Naïve Bayes Classifier is a datamining classification method which takes probabilities of attributes belonging to class for prediction. NBC is a supervised classification approach which can be used effectively to model a predictive problem probabilistically [1].

Naïve Bayes classifier is based on Bayes' Theorem where predictors are treated as Independent. In Naïve Bayes method the overall probabilities of attributes belonging to a class are calculated by presuming that the likelihood of an attribute on a given class value is not dependent on other attributes. This presumption leads NBC to better results and is called conditional independence [1][7].

Naïve Bayes theorem is described as follows,



$$P(c | X) = P(x_1 | c) \times P(x_2 | c) \times \dots \times P(x_n | c) \times P(c)$$

The posterior probability, P(c/x) can be calculated, from P(C), Class Prior Probability, P(x) Predictor Prior Probability and P(x/c) Likelihood. The conditional Independence is explained in this scenario as, the predictor (x) value on a class (c) has no effect on the other predictor's values.

The Naïve Bayes Classifier in this paper takes numeric attributes as input and the values of each numeric attribute are Gaussian distributed. This was considered for robust results.

Gaussian Distribution

Gaussian distribution is defined by mean and standard deviation, which were defined as below

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{Mean}$$

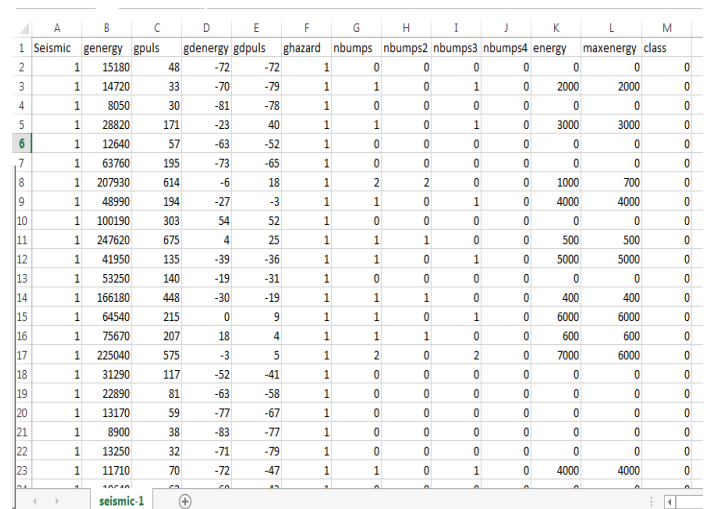
$$\sigma = \left[\frac{1}{n-1} \sum_{i=1}^n (x_i - \mu)^2 \right]^{0.5} \quad \text{Standard Deviation}$$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{Gaussian Distribution}$$

3. DATA

Data in .csv format which was given as input to NBC for accuracy estimation was download from UCI Machine Learning Repository website which falls under Multivariate category. This data is a collection of forecasted seismic bumps in a coal mine, collected from two of long walls of a Polish coal mine [6].

Each row of the data describes the seismic activity in the rock mass within one shift (8 hours). If decision attribute has the value 1, then any seismic bump with energy > 10^4 J is registered in the next shift. This is the primary attribute for accuracy estimation in this paper. A sample screen shot of data for accuracy estimation is shown in Fig-1.



Seismic	genergy	gpuls	gdenery	gdpuls	ghazard	nbumps	nbumps2	nbumps3	nbumps4	energy	maxenergy	class
1	15180	48	-72	-72	1	0	0	0	0	0	0	0
1	14720	33	-70	-79	1	1	0	1	0	2000	2000	0
1	8050	30	-81	-78	1	0	0	0	0	0	0	0
1	28820	171	-23	40	1	1	0	1	0	3000	3000	0
1	12640	57	-63	-52	1	0	0	0	0	0	0	0
1	63760	195	-73	-65	1	0	0	0	0	0	0	0
1	207930	614	-6	18	1	2	2	0	0	1000	700	0
1	48990	194	-27	-3	1	1	0	1	0	4000	4000	0
1	100190	303	54	52	1	0	0	0	0	0	0	0
1	247620	675	4	25	1	1	1	0	0	500	500	0
1	41950	135	-39	-36	1	1	0	1	0	5000	5000	0
1	53250	140	-19	-31	1	0	0	0	0	0	0	0
1	166180	448	-30	-19	1	1	1	0	0	400	400	0
1	64540	215	0	9	1	1	0	1	0	6000	6000	0
1	75670	207	18	4	1	1	1	0	0	600	600	0
1	225040	575	-3	5	1	2	0	2	0	7000	6000	0
1	31290	117	-52	-41	1	0	0	0	0	0	0	0
1	22890	81	-63	-58	1	0	0	0	0	0	0	0
1	13170	59	-77	-67	1	0	0	0	0	0	0	0
1	8900	38	-83	-77	1	0	0	0	0	0	0	0
1	13250	32	-71	-79	1	0	0	0	0	0	0	0
1	11710	70	-72	-47	1	1	0	1	0	4000	4000	0

Fig -1: Screenshot of Seismic Bumps Dataset

Value - '1' in the decision attribute, i.e., Class in the above dataset means a high energy seismic bump occurred in the next shift that means 'hazardous state,' '0' indicates a not high energy seismic bumps happened on the next shift that means a 'non-hazardous state' [6].

3. IMPLEMENTATION

The stepwise implementation of the proposed model to minimize the loss of accuracy is as follows,

Step-1: Dataset in CSV is given as input and split into Training and Test datasets.

In this paper, a ratio of 67% and 33% is considered for Training and Test Sets.

Step-2: Differentiate the Training data set as per the Class values. i.e. 1, 2 & 3.

Step-3: Calculate the Mean and Standard Deviation for each data instance in the order of class values.

Step-4: Use the above values to calculate probabilities corresponding to class values using Gaussian Distribution Function.

Step-5: Generate probabilities for all attributes of a class belonging to Training set to the data instances of test dataset.

Step-6: If the resultant probability is '0', after Step 5 for a particular data instance, which is not first in the list, then the mean value of preceding probabilities of the attributes is taken as the current probability.

If the data instance is itself in the top of the list and probability is '0' then an equivalent value of '1' is added to attribute values of that particular data instance of the Training dataset.

Step-7: Generate Predictions by comparison between probabilities of data instances of each class values belonging to Training Dataset.

Step-8: Evaluate the accuracy of predictions by comparing with the class values of test dataset. The accuracy is computed regarding ratio between 0 to 100%.

The addition of Step-6 to the existing Naïve Bayes Classifier increases the accuracy to a great extent by considering the probabilities of each data instance belonging to Training Set, even with the assumption of Conditional Independence [2][4].

5. RESULTS AND COMPARISONS

The proposed model described in the section –IV was executed on the Seismic Bumps Dataset with Step-6 and without Step-6. In each case, the results were analyzed by accuracy of Naïve Bayes Classifier. The experimental result shows the accuracy of Naïve Bayes Classifier improved when Step-6 is executed (81.1%) when compared to Classifier without Step-6 (64.5%). The Native Matlab Naïve Bayes Classifier, when applied on the same data set, gave an accuracy of 65.09%.

A comparative analysis of the proposed method, regarding accuracy, with other models, is shown in Table-1. Based on the results as shown in Table-1 it is evident that the proposed novel method minimizes the loss of accuracy in Naïve Bayes Classifier even with the assumption of Conditional Independence. The proposed model in this paper further improved the performance of Naïve Bayes Classifier.

Table -1: Sample Table format

Classifier	Performance		
	Training (60%)	Test (40%)	Accuracy in %
NBC with Conditional Independence and with executing Step-6	1550	1034	81.1
NBC with Conditional Independence and without executing Step-6	1550	1034	64.5
Native Matlab NBC Algorithm	1550	1034	65.09

As per the experimental results, the NBC classifier algorithm presented in this paper performs better when compared to the native MATLAB Naïve Bayes Function without Negation

Handling, which gave an accuracy of around 65.09% when carried out on the same dataset with the same ratio of Training and Test data. Table-I illustrates that the proposed method has improved accuracy (+10%), due to preprocessing of input data using negation handling, thus reducing the impact of Class Conditional Independence, when compared to Matlab Native NBC (65.09%) and NBC (64.5%) without Negation Handling.

6. CONCLUSIONS

The proposed Naïve Bayes Classifier algorithm used for predicting for Seismic Hazard, with 60%-40% ratio of Training and Test Datasets, yielded a good accuracy of 81.1% which is better when compared to the native MATLAB Naïve Bayes Classifier. Experimental results indicate that using proposed NBC algorithm has significantly improved accuracy. Our future work will be exploring the feasibility of applying various distributions like Multinomial, Bernoulli, etc. and using other smoothing techniques as mentioned in Section-I to further improve the performance of Naïve Bayes Classifier.

ACKNOWLEDGEMENT

We sincerely acknowledge the creator and donors, Marek Sikora (marek.sikora@polsl.pl), Institute of Computer Science, Silesian University of Technology, 44-100 Gliwice, Poland and Lukasz Wrobel (lukasz.wrobel@polsl.pl), Institute of Innovative Technologies EMAG, 40-189 Katowice, Poland, which was download from UCI Machine Learning Repository.

REFERENCES

- [1] J. Han, M. Kamber, 'Data Mining Concept and Techniques', Morgan Kaufmann, 2001.
- [2] Improving Naive Bayes Classifier Using Conditional Probabilities , Sona T, M Mammadov, A M. Bagirov, Proceedings of the 9-th Australasian Data Mining Conference (AusDM'11), Ballarat, Australia.
- [3] Scaling Up the Accuracy of Naïve Bayes Classifiers a Decision Tree Hybrid, Ron Kohavi.
- [4] Novel Frequent Sequential Patterns based Probabilistic Model for Effective Classification of Web Documents, H Haleem, P K Sharma, M M S Beg, Proceeding in 2014 5th International Conference on Computer and Communication Technology.
- [5] An Improved Computation of the PageRank Algorithm, S J Kim, Sang H Lee, Springer-Verlag Berlin Heidelberg 2002.
- [6] UCI Machine Learning Repository – Seismic Bumps data
- [7] Xi-Zhao Wang, Yu-Lin He and Debby D. Wang, "Non-Naïve Bayesian Classifiers for Classification Problems with Continuous Attributes", IEEE Transactions on Cybernetics, Vol.44, No.1, January 2014.