Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Predictive Data Mining Model for Cyber Crime

Dr. Krishna Kumar Verma[1], Bharat Lal Tiwari[2], Dr. Prabhat Pandey[3]

[1]Govt. College Tala, Satna, A.P.S. University, Rewa (M.P.) India
[2]Research Scholar, A.P.S. University, Rewa (M.P.) India
[3]OSD, Additional Directorate, Higher Education, Division Rewa (M.P.) India
[1] krishnamscit.verma@gmail.com, [2]bltiwarics@gmail.com,
[3]pandeyprabhat51@gmail.com

Article Info
Page Number: 10194-10209
Publication Issue:
Vol. 71 No. 4 (2022)

Article History
Article Received:
12 September 2022
Revised: 16 October 2022
Accepted: 20 November 2022
Publication: 25 December 2022

Abstract
Application of data mining techniques (DMT) in different field like e-governance, e-agriculture, e-health, e-education, etc has contributed significant results. The results produced by DMT can be utilized by policy maker/ government officer for people welfare and making good governance. One of the applications of DM may be cyber crime classification. Internet usage has increased tremendously since the last few years, due to which the incidents of cybercrime have also increased rapidly. It causes the generation of new dataset in the field of cyber crimes. For these dataset, efficiency of the data mining algorithm is yet to be checked. Hence, this paper deals with the objective to develop efficient classification model to classify cyber crime incidents. We propose an iterative classification model using data mining techniques. Five classification models Naïve Bayes, Support Vector Machine, K Nearest Neighbors, Decision List and Decision Tree have been trained and tested on selected dataset. The iterative classification model relation with SVM was found best with 95.03% accuracy as compare to other classification models.

Keywords: Data Mining, Cyber Crime, Cyber Attack, Predictive Techniques

1 Introduction

Data mining has proven to be a proactive decision support tool to predict and prevent cyber crimes. Predicting cybercrime prototypes using data mining techniques has been a viable explanation to help law enforcement officers in mitigating the problems associated with cyber crimes and related security. In the modern era of rising cybercrime rates, crime prevention is one of the most important universal concerns, with the utmost concern for enhancing public safety. IT analysts/officers as well as law enforcement officers were quickly solving cybercrime cases through the use of advanced computerized systems to track cybercrimes and hunt down cybercriminals [15].

Cybercrime detection is the identification of signs of cybercrime when there is no prior mistrust. Initially, one must learn whether the given data samples are misleading or not. This can be done through supervised or unsupervised learning. Supervised learning of cybercrime dataset is a process where previously known data with classes are available. Unsupervised learning of cybercrime dataset is a case where data is available but we don't know what data is presenting, sometimes mimic the nature of fraud or crime. These data patterns are then

processed based on our actions. Various terms are used to accomplish this task and are described as techniques and methodologies for cyber crime detection. Information security policies improve the security and well-being of information assets. They form the foundation rules and regulations of information security within the organization. Fraud remains a challenge for businesses and organizations in many sectors. Data mining is an effective method for detecting various types of cybercrime, including telecommunications, credit card and health insurance fraud, as well as detecting intrusions into computer systems [11].

Methods are inadequate to deal with the volume of activity in cyberspace and its anonymous nature, and this manual effort must be accompanied by computational effort such as data mining. Cybercrime is not a problem of today's time, it is an age-old problem, through which people's data has been either destroyed and either it has been misused. Statistical methods were used earlier to detect and overcome this, but these methods were not sufficient for large scale data. In today's time, data mining and big data technology is being used to detect and reduce cybercrime [6].

2 Literature Review

A Novel approach has been developed which has included data mining technology and visualization techniques for prediction of cyber crime distribution over major areas of India. The developed cybercrime scanning tool provides a framework to visualize different types of cyber crimes and cyber crime prone areas in India and investigate them by data mining algorithms using Google Maps. This function allows law enforcement officers to investigate cybercrime networks through interactive visualizations. The relevance of the interactive and visual aspect will promote exposure and understanding of cyber crime prototyping. In an effort to help law enforcement officers protect humanity and consider cybercrimes, data mining algorithms and visualization techniques have been used. Based on the performance evaluation of existing and proposed classifiers, Enhanced Random-Forest achieved an accuracy rate of 99.58% with less computation time than Naïve-Bayes [13].

A data mining model has been proposed in [16] using mining tasks such as cleaning, Transformation etc. Data mining algorithms have been compared on the basis of accuracy. Data cleaning and pre-processing, experimental analysis, DM model development and evaluation of DM model have been done using python programming. Data analysis has been done using algorithms like Logistic Regression, KNN, Random Forest, KNN, DT and SVM. After comparison of results of these algorithms; it has been observed that Logistic Regression model has given higher precision to predict result.

A computational tool using ML algorithms for analysis of cyber-crime rate has been proposed in [3]. The proposed tool was flexible in nature and able to classify cyber-crimes state-wise in a country. The dataset has been downloaded from Kaggle and CERT_In. Data pre-processing has been done using "TFIDF weighted vector process" and chi-squared test has been performed. It has been found that MP state has max number of crime rate. DM model has been developed using logistic regression, random forest, linear SVM, multinominal NB. The accuracy of these developed model were logistic regression- 0.9938,

random forest- 0.8069, linear_SVM- 0.9923, multinominal_NB- 0.9895. Best model was linear SVM.

Researchers have discussed the application of GIS in the field of crime prediction using ML algorithms. Point of Interest (POI) features from Open_Stree_Map is useful prediction of cybercrime and help to distinguish between upper and lower regions with precision. SVM, Logistic regression, Decision tree and Random forest classification algorithms have been used to develop prediction model. Model has been compared with the help of ROC curves which has been created for every city and for every type of crime like shoplifting, theft, robbery etc. It has been found that Random Forest given best performance for urban area crime prediction. The SVM performed also performed best result in most of the cases [4].

Linear regression technique has been used in [19] to analyze crime data which help in prediction of crime rate. Secondary crime dataset of "Los Angeles Police Department (LAPD)" from 2010 to 2019 has been used for prediction of future crime pattern. The length of dataset was 2036897 rows having data related to crime occurrences. On selected dataset the linear regression algorithm has given best result with R^2= 0.19. It has been concluded that the developed model can be utilized to predict crime cases in Indonesian region because the attribute structure of Indonesian crime dataset is same as dataset that has been used in experiment/ model development.

Use sequential rule mining over "intrusion detection alerts" has been proposed in [9]. The used dataset has 12 million alerts which have been detected form 34 intrusion-detection systems of 3-organizations. This dataset has been processed with proposed analytical model. They have created predictive blacklist using sequential rules mining. Authors have observed that using proposed approach they were capable to analyze predictive-rules from dataset which can be used in prediction of future upcoming cyber-security actions. The predicted alert was 60% which can be utilized to develop predictive-blacklist of IP_addresses.

Decision Tree algorithm can be used to find the answer of question- why cyber crime is increasing? They have discussed that cyber crime is increasing because a group of people are trying to know "how cyber crime can commit?" This learning process produces cyber criminal. In this way the cases of cyber crime increase day by day in US. Finally authors have used prediction method to predict the cases of cyber crime that can be used to reduce cyber crime in future [17].

The use of Design Science Research (DRS) approach has been discussed in [2] to do their research. The main focus of this research was to build and evaluate artifacts. In this sequence "data analysis framework" and "classification model" have been proposed. Pre-evaluation of accuracy and post evaluation of performance of classification model has also been done by authors. The NB classifier has been used in proposed model. Study also covered a set of proposed definitions for type of Crime as Service (CaaS) and Crime_Ware. On the basis on literature author have built RAT_Based classification model. It has been concluded that the mostly target organizations are technology-companies having 28% as compared to other companies (content: 22%, finance: 20%, e-commerce: 12% & tele-communication: 10%).

In [18] it has been discussed that crime incident can affect "life quality", "economic growth" and "reputation of nation". They have developed a tool that provides a visualization framework to visualize crime networks and able to analyze them. This development has been done by the use of Google-Maps and R-Packages. The two main data mining algorithms are NB and KNN have been used for performing analysis task. Data analysis has been done on "U.K. Police department" data-set. From 11 attribute of dataset, 5 attributes has been selected to do experiment. Developed visualization tool has 6 modules which presents different types of analysis result. It has been concluded that there is 37.5% probabilities of reporting case is Anti-Social behavior.

It has been proved that Random Forest (RF) classifier gives best result to classify crime data. To find best classifier authors have used different classifier on dataset and results of these classifiers have been compared with the help of accuracy, precision, recall and F score. Classification, precision, generalization and error reduction increase efficiency through proper sample preparation and evaluation which appeared to aid this study by providing accurate and reliable efficacy. DT, RF, NB and Logistic_Regression classification algorithms are used for classification task. Classification accuracy of Clean data by DT was 75.90%, RF was 83.39%, NB was 77.64% and logistic regression was 64.72%. Classification accuracy of dirty data according to DT was 76.77%, RF was 81.35%, NB was 75.42% and logistic regression given 66.93% [22].

A novel DM technique has been proposed in [10] to analyze cyber crime datasets. The proposed method was combination of K-Mean, Influenced Association Rule and J48 Prediction Tree. Weighted support and confidence were used in Influenced Association Classification, which makes it an improved classifier. The K-mean groups dataset into n groups or clusters. These clusters can be used for further rule mining followed with j48 decision tree, which generate high precise and accurate results. Hence cyber crime detection and prediction can be done efficiently using proposed approach.

3 DM Algorithms & Performance Measures

DM Techniques has two categories: Predictive and Descriptive. Both techniques are very useful to analyze dataset and extracting knowledge from large volume of data stored in server. Classification is a process to categorize dataset into particular class level. The classification is a type of predictive techniques. Predictive techniques classifies data-set into pre-defined classes whereas descriptive techniques classifies dataset into undefined class levels which can be used to define a particular class levels [20].

An attempt has been made to develop classification model based on supervised learning. For this purpose five classification algorithms have been selected. These classification algorithms are "Naïve Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision List (DL) and Decision Tree (DT)". Various performance indicators have been used for evaluation of developed model. The common indicators are "True_Positive rate, False_Positive rate, Precision, Recall, F-Measure and ROC area". These performance indicators can be obtained through the confusion matrix and other statistical methods.

3.1 Naïve Bayes

Naïve Bayes classifier uses bayes algorithm. This algorithm makes independent assumption; on the basis of target value, individual attribute values are selected which are independent to each other [7]. In other words Naïve Bayes is based on simple Bayes Network model and under this model all variables are assumed as independent.

3.2 Support Vector Machine (SVM)

SVM is a predictive data mining techniques. It is extensively utilized in pattern identification, regression, or classification related problems [5]. We have used the SVM algorithm known as SMO function in weka [8].

3.3 K-Nearest Neighbors

In this classifier, each datasets are represented in the form of n-dimensional space point. Data points are categorized on the basis of distance between other data-set points. Euclidean distance is used in the process of K nearest neighbors to identify nearest neighbor. IBk algorithm is a type of KNN classifier, proposed by Aha and Kibler [1].

3.4 Decision Tree

A DT is a structural model like a tree. It has leaves, which represent specific class, and branches of tree which represent the set/ path of features that lead to reach in this particular classification [12].

We have selected five categories and their appropriate algorithm (available in weka) for data mining purpose. These techniques along with algorithm are Naïve Bayes->Naïve Bayes, SVM-> SMO, KNN-> IBK, DL-> PART and DT-> J48.

3.5 Performance Measures

To identify the best model among various available techniques/ algorithm over selected dataset there is a need to compare the results of every technique. Data Mining/ Machine Learning process provided some performance measures that can help to identify best data mining algorithm on selected dataset. We can use K Fold Cross Validation, Confusion Matrix, and area in ROC Curve etc. to compare results of classification techniques. K-Fold cross validation has been adopted to minimize the bias related to random sampling of training and test dataset. The K-Fold Cross Validation process makes random partition of dataset into k mutually exclusive sub-sets or "folds", like D_1, D_2, D_3, D_4, ….. D_n; each part approximately has equal size and training and testing process is performed K number of times [7].

3.6 Confusion Matrix

The primary source of performance measurements in classification is confusion matrix. It is also known as coincidence-matrix, contingency-table or classification-matrix. Olson & Delen [14] discussed that for n class problem, a confusion-matrix is a matrix in the form of two dimensional table of (m X m) size. The basic structure of confusion matrix has been shown in

Table 3.1, which presents a two class classification confusion matrix. Other parameters can be calculated using confusion matrix. The formulas are as follows [21]:

- **Accuracy** = (TP + TN)/ (TP + TN + FP + FN))
- **Specificity** = TN / TN + FP
- **Sensitivity** = TP / TP + FN
- **Precision** = TP / TP + FP
- **F-Measure** = 2* Precision* Recall/ (Precision + Recall))

		Actual Class	
		0	1
Predicted Class	0	True Negatives (TN): *"If the instance is negative and it is classified as negative, it is counted as a true negative (TN)"*	False Positive (FP): *"If it is classified as positive, it is counted as a false positive (FP)"*
	1	False Negatives (FN): *"If it is classified as negative, it is counted as a false negative (FN)"*	True Positive (TP): *"If the instance is positive and it is classified as positive, it is counted as a true positive (TP)"*

Table 1: Basis of Confusion Matrix

4 Proposed Methodology

Following algorithm has been used to classify dataset:

Step 1: Start

Step 2: Selection of Dataset

Step 3: Preprocessing of Data

Step 4: Training of Classification Model

Step 5: Testing of Classification Model

Step 6: If efficient result was found

Go to step 7

Else

Go to step 3 (for refining the dataset based on output)

Step 7: Result Interpretation

Step 8: Stop.

5 DM Model Development

Predictive algorithm has been used to develop classification model and analysis of data. Classification of dataset has been performed using five data mining techniques. Data mining model has been developed using Weka tool and try to classify cyber attack dataset. Ten fold cross validation parameter has been used for creating training and testing dataset. As we know that, 10 fold cross validation divides dataset into 10 parts. Every set of nine parts (90% of dataset) is used as training dataset and rest of one part (10% dataset) is used as testing dataset. This step is performed iteratively for every possible 90% and 10% dataset, which generate result more efficient rather than dividing dataset into ratio of 70:30 %. Initially, rests of parameters are set as default. Dataset has been collected from website http://hackmageddon.com/. Cyber incidents from 2013 to 2020 are available in this website. We have selected most recent cyber incidents of the year 2020 to conduct our experimental research. The results of first attempt of experiment have been discussed below.

5.1 NaiveBayes Classifier Iteration-I

		True/ Actual					
	CLAS	a=CC	b=CE	c=CW	d=N/	e=H	f= >1
Predicted	a	1960	49	21	12	23	1
	b	22	183	3	1	1	0
	c	5	1	18	1	1	1
	d	0	0	0	0	0	0
	e	18	1	1	1	8	0
	f	0	0	0	0	0	0

Table 2: Confusion Matrix of NB (Iteration-I)

5.2 SMO Classifier Iteration-I

		True/ Actual					
	CLA	a=CC	b=CE	c=C	d=N/	e=H	f= >1
Predicted	a	1985	48	16	11	27	1
	b	12	185	1	2	0	0
	c	6	1	23	0	2	1
	d	0	0	1	2	0	0
	e	2	0	2	0	4	0
	f	0	0	0	0	0	0

Table 3: Confusion Matrix of SMO (Iteration-I)

5.3 IBK Classifier Iteration-I

		True/ Actual					
	CLA	a=C	b=C	c=C	d=N/	e=H	f= >1
Predicted	a	1971	61	21	10	24	2
	b	19	172	4	1	0	0
	c	5	1	18	0	2	0
	d	1	0	0	4	0	0
	e	9	0	0	0	7	0
	f	0	0	0	0	0	0

Table 4: Confusion Matrix of IBK (Iteration-I)

5.4 PART Classifier Iteration-I

		True/ Actual					
	CLAS	a=CC	b=CE	c=CW	d=N/	e=H	f= >1
Predicted	a	1988	58	40	15	33	2
	b	17	176	2	0	0	0
	c	0	0	1	0	0	0
	d	0	0	0	0	0	0
	e	0	0	0	0	0	0
	f	0	0	0	0	0	0

Table 5: Confusion Matrix of PART (Iteration-I)

5.5 J48 Classifier Iteration-I

		True/ Actual					
	CLAS	a=CC	b=CE	c=CW	d=N/	e=H	f= >1
Predicted	a	1980	48	21	13	31	2
	b	19	186	2	2	0	0
	c	6	0	19	0	2	0
	d	0	0	0	0	0	0
	e	0	0	1	0	0	0
	f	0	0	0	0	0	0

Table 6: Confusion Matrix of J48 (Iteration-I)

In first attempt we observed that some class levels have very low frequency in their tuple value hence algorithm may not able to classify these tuples in particular class level. In the next experiment class levels having very low frequency in their tuple value has been

removed. The result of next experiment has been dictated as algorithm_name followed by term "Iteration-II". Experimental results of different classifier and different iterations are discussed in this chapter.

5.6 Result and Discussions (Iteration-I):

In first iteration of model development, the percentage of correctly classified instances given by selected algorithm was NB - 93.0103%, SMO - 94.2967%, IBK -93.1389, PART - 92.8388 and J48 - 93.6964. The percentage of incorrectly classified instances given by selected algorithm was NB - 6.9897%, SMO - 5.7003%, IBK - 6.8611, PART - 7.1612% and J48 - 6.3036%. The comparisons of correctly and incorrectly classified instances are presented in Figure 1 and 2. Comparison shows that SVM (SMO) algorithm has produced best result as compare to other classifier. Additionally, It has been observed that, the algorithms used to predict cyber attack was not able to classify class level "NA" and ">1" correctly, because these class levels have very few records in dataset. Therefore, experiment has been extended for second iteration. In next iteration of model development, the class level "NA" and ">1" and instances related to these class levels have been removed; and the model development process for updated dataset has been performed again. Results of iteration-II is presented and discussed in next section.

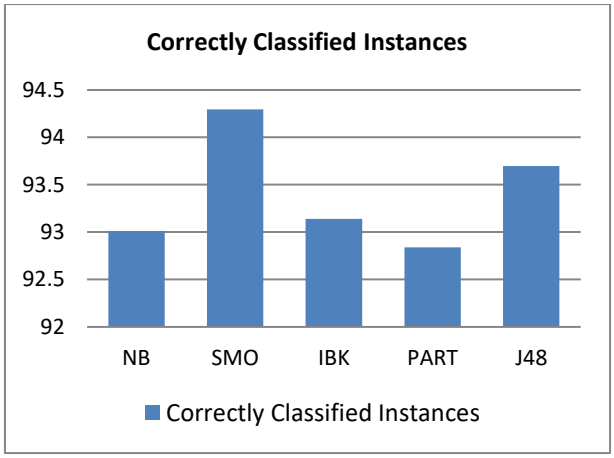

Figure 1: Correctly Classified Instances (Iteration-I)

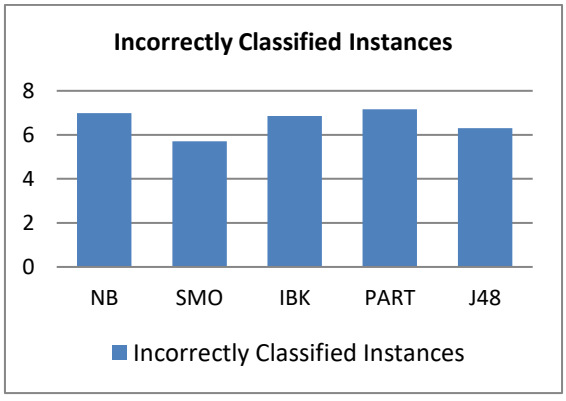

Figure 2: Incorrectly Classified Instances (Iteration-I)

5.7 NaiveBayes Classifier Iteration-II

		True/ Actual			
	CLASS	a= CC	b= CE	c= CW	e= H
Predicted	a	1963	50	22	23
	b	23	183	2	0
	c	2	1	18	1
	d	17	0	1	9

Table 7: Confusion Matrix of NB (Iteration-II)

5.8 SMO Classifier Iteration-II

		True/ Actual			
	CLASS	a= CC	b= CE	c= CW	e= H
Predicted	a	1985	48	15	27
	b	15	185	1	0
	c	4	1	26	2
	d	1	0	1	4

Table 8: Confusion Matrix of SMO (Iteration-II)

5.9 IBK Classifier Iteration-II

		True/ Actual			
	CLASS	a = CC	b = CE	c = CW	e = H
Predicted	a	1974	63	24	26
	b	16	170	5	0
	c	5	1	14	1
	d	10	0	0	6

Table 9: Confusion Matrix of IBK (Iteration-II)

5.10 PART Classifier Iteration-II

		True/ Actual			
	CLASS	a = CC	b = CE	c = CW	e = H
Predicted	a	1988	67	34	33
	b	15	167	1	0
	c	2	0	8	0
	d	0	0	0	0

Table 10: Confusion Matrix of PART (Iteration-II)

5.11 J48 Classifier Iteration-II

		True/ Actual			
	CLASS	a = CC	b = CE	c = CW	e = H
Predicted	a	1980	48	21	31
	b	19	186	2	0
	c	6	0	20	2
	d	0	0	0	0

Table 11: Confusion Matrix of J48 (Iteration-II)

5.12 Result and Discussion (Iteration-II):

In Iteration-II, after removing low frequency instances; the model development process performed again. The output produced by this process shows that for every data mining algorithm, the accuracy has been increased. The correctly classified instances and incorrectly classified instances produced by second iteration have been presented in Figure 3 and 4 respectively. Comparison of F-Measure and IR-Precision are also presented in Figure 6 and 7. Additionally, comparison of accuracy produced by algorithms during in iteration-I and II has been presented in Figure 7. These comparisons show that SMO algorithm has produced best result (with highest number of correctly classified instances 95.03% and minimum number of incorrectly classified instances 4.96 %) as compare to other DM algorithms.

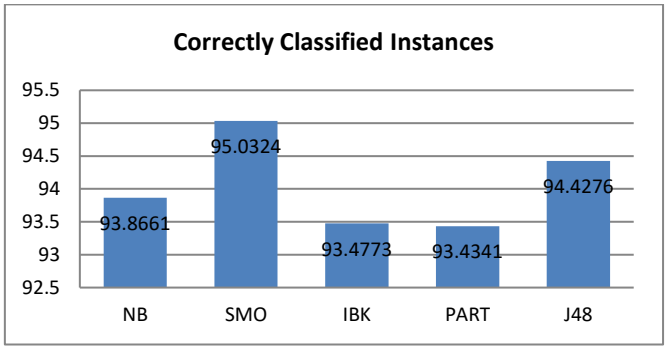

Figure 3: Comparison of Correctly Classified Instances

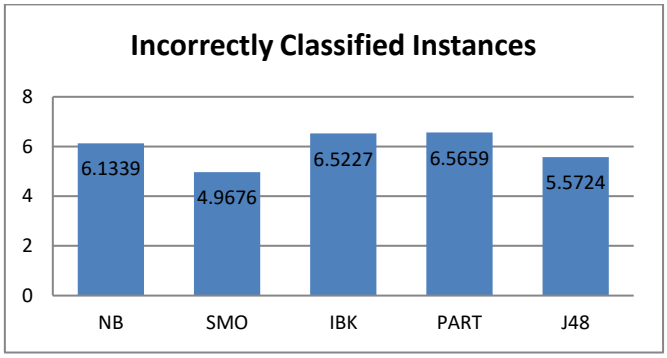

Figure 4: Comparison of Incorrectly Classified Instances

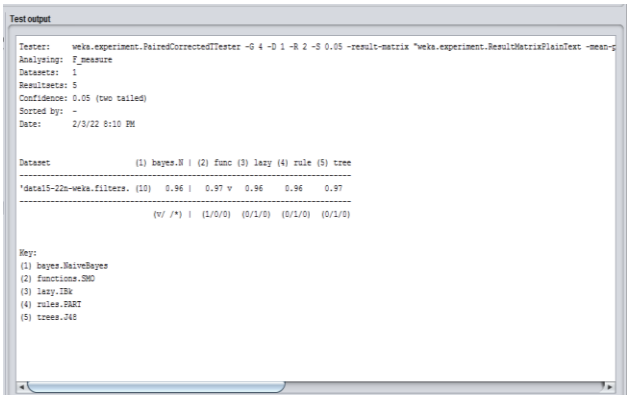

Figure 5: Comparison of F_Measure

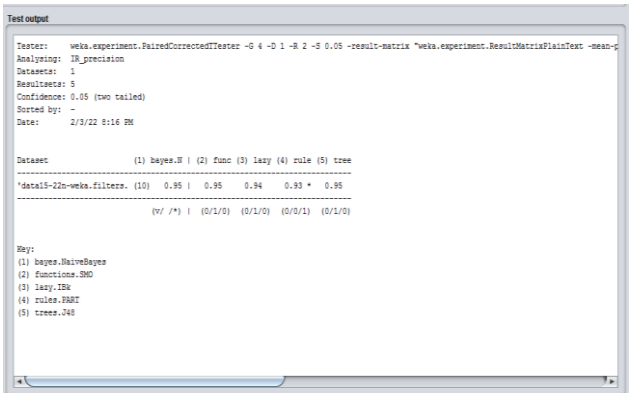

Figure 6: Comparison of IR_Precision

5.13 Comparison of results of Iteration-I and Iteration-II

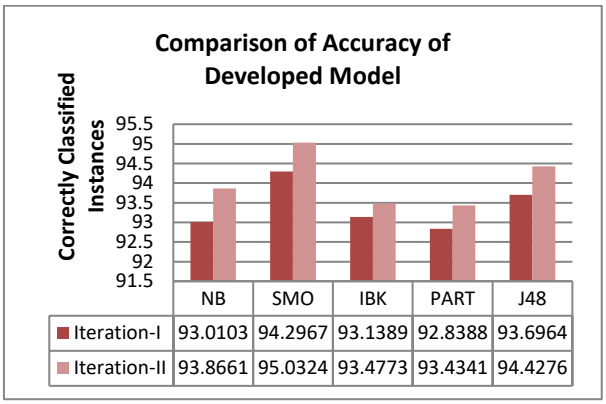

Figure 7: Comparison of Accuracy of Developed Model

Comparison of accuracy of both iterations has been presented here. Graph shows that SVM (SMO) algorithm has produced best result with 94.2967% (Iteration-I) and 95.0324% (Iteration-II) accuracy as compared to other algorithms. Hence we can say that SVM algorithm can be utilized to predict cyber attack which will definitely produce accurate result. Using these results Cyber officer may able to take effective decision to reduce cyber crime and draft new rules related to cyber crime and security.

6 Conclusion

Data mining techniques are very useful to produce knowledge form large dataset stored in server. Prediction and description of dataset can be performed using these techniques. This research paper has presented prediction techniques to classify cyber attacks categories. The concept of iterative classification model has been proposed. Same classification techniques have been utilized twice (Iteration I and II). In the result of first iteration, it has been found that instances related to some class levels have very few frequency in dataset that are not able to classify correctly in particular classes. In Iteration II, these low frequency data and their class levels have been removed. The classification model development process is performed again on refined dataset. Five classification models NB, SVM (SMO), KNN (IBK), Decision List (PART) and Decision Tree (J48) have been trained and tested on selected dataset. The interactive classification model relation with SVM was found best with 94.296% in Iteration I and 95.032% in Iteration II accuracy as compare to other classification models. Using these results Cyber officer may able to take effective decision to reduce cyber crime and draft new rules related to cyber crime and security. The proposed technique can be utilized in development of real time cyber crime classification framework.

Reference

[1] Aha, D. W., Kibler, D., And Albert, M. K. Instance-based learning algorithms. *Machine learning*, *6*(1), p. 37-66, 1991.

[2] Avinash, G. R., Scholar, P., And Rao, M. V. Data Analytics Approach To The Cybercrime, Underground Economy. *Complexity International*, *24*(01), 2020.

[3] Ch, R., Gadekallu, T. R., Abidi, M. H., And Al-Ahmari, A. Computational system to classify cyber crime offenses using machine learning. *Sustainability*, *12*(10), p. 4087, 2020.

[4] Cichosz, P. Urban Crime Risk Prediction Using Point of Interest Data. *ISPRS International Journal of Geo-Information*, *9*(7), p. 459, 2020.

[5] Cortes, C., & Vapnik, V. Support-vector networks. *Machine learning*, *20* (3), p. 273-297, 1995.

[6] Foster, J. 21 Terrifying Cyber Cyber Crime. Retrieved from Data Connectors: https://www.dataconnectors.com/technews/21-terrifying-cyber-crime-statistics/, (April 17), 2020.

[7] Han, J. and Kamber M. Data Mining Concepts and Techniques, Second Edition: Morgan Kaufmann Publishers, San Francisco, 2006.

[8] Hastie T. and Tibshirani R. Classification by pairwise coupling. *Proceedings of the 1997 conference*, 1998.

[9] Husák, M., Bajtoš, T., Kašpar, J., Bou-Harb, E., and Čeleda, P. Predictive cyber situational awareness and personalized blacklisting: a sequential rule mining approach. *ACM Transactions on Management Information Systems (TMIS)*, *11*(4), p. 1-16, 2020.

[10] Lekha, K. C., and Prakasam, S. Data mining techniques in detecting and predicting cyber crimes in banking sector. In *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, IEEE, p. 1639-1643, (August), 2017.

[11] Lekha, K. C., and Prakasam, S. Implementation of data mining techniques for cyber crime detection. *International Journal of Engineering, Science and Mathematics*, *7*(4), p. 607-613, 2018.

[12] Malathi, A., and Baboo, S. S. An enhanced algorithm to predict a future crime using data mining, 2011.

[13] Michael, G. Knowledge Based System For Predicting Cyber Crime Patterns Using Data Mining Techniques. *Journal of Critical Reviews*, *7*(10), p. 2043-2053, 2020.

[14] Olson, D. L., and Delen, D. Advanced data mining techniques. *Springer Science & Business Media on Advances in neural information processing systems 10*, Denver, Colorado, USA, p. 507-513, 2008.

[15] Prabakaran, S., and Mitra, S. Survey of analysis of crime detection techniques using data mining and machine learning. In *Journal of Physics: Conference Series* (Vol. 1000, No. 1, p. 012046). IOP Publishing, (April), 2018.

[16] Prithi, S., Aravindan, S., Anusuya, E., and Ashok Kumar, M. Gui Based Prediction Of Crime Rate Using Machine Learning Approach, 2020.

[17] Toapanta, S. M. T., Gallegos, L. E. M., Andrade, B. E. C., and Espinoza, M. G. T. Analysis to predict cybercrime using information technology in a globalized environment. In *2020 3rd International Conference on Information and Computer Technologies (ICICT)*, IEEE, p. 417-423, (March), 2020.

[18] ToppiReddy, H. K. R., Saini, B., and Mahajan, G. Crime prediction & monitoring framework based on spatial analysis. *Procedia computer science*, *132*, p. 696-705, 2018.

[19] Trisnawarman, D., and Imam, M. C. Prediction Analysis Of Criminal Data Using Machine Learning. In *IOP Conference Series: Materials Science and Engineering* (Vol. 852, No. 1, p. 012164). IOP Publishing, (July), 2020.

[20] Umadevi, S., and Marseline, K. J. A survey on data mining classification algorithms. In *2017 International Conference on Signal Processing and Communication (ICSPC)*, IEEE, p. 264-268, (July) 2017.

[21] Witten, I., Frank, E., Hall, M. and Pal, C. DATA MINING: Practical Machine Learning Tools and Techniques. *San Francisco, CA, USA: Morgan Kaufmann Publishers Inc*, 2016.

[22] Yerpude, P. Predictive Modelling of Crime Data Set Using Data Mining. *International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol*, *7*, 2017.