IJITKM 0973-4414 Volume-10, Number-1 Jun-Dec 2016 pp. 1-5 **Impact Factor 2.8** available online at www.csjournalss.com

Prediction of Respondants' Knowledge towards Cyber Security measures using various Classification Algorithms

[1]K. Chitra Lekha, [2]Dr. S. Prakasam
[1] Ph.D Research Scholar, [2] Associate Professor
Department of Computer Science and Applications, SCSVMV University, Enathur, Kanchipuram.

Abstract: Cyber security involves safeguarding of sensitive, personal and business information through prevention, recognition and retort to unusual online attacks. The main objective of this work is to find the best classifier from the performance evaluation of different classifiers of data mining techniques. The purpose of this work is to predict how the respondents of various categories are alert about cyber security measures using the best classifier. A study has been conducted during November 2016 with different category respondents of 189. The questionnaire was planned to forecast the respondents' attitude towards cyber security measures. The WEKA tool is used for comparing the performance evaluation of different classifiers for the purpose of concluding best classifier so that it can further be used for prediction. In this work, NaiveBayes, WAODE, SimpleLogistic, SMO, JRip, NNge, NBTree, RandomTree classifiers were used for the intention of analyzing the best classifier for the prediction of cyber security measures dataset.

Keywords: Cyber security, JRip, NaiveBayes, NBTree, NNge, RandomTree, SimpleLogistic, SMO, WAODE, WEKA.

1. INTRODUCTION

Cyber security is important because it helps in defending the computer system against different types of destructive technologies and protects the PC from damage (viruses, worms, bugs etc). Computer security is vital for protecting the confidentiality, integrity and availability of computer systems, resources and data [2]. The global spam rate, malware rate and phishing rate is increasing rapidly and also there is a potential impact of cyber crime on economics, consumer trust and production time. The counter measures like GPRS Security architecture, Intrusion Detection and prevention System and Agent based Distributed Intrusion Detection system are used for security purposes [4]. The WEKA tool is chosen for study execution, as it contains in-built data preprocessing features, classifiers, clustering techniques, Association techniques and also data visualization techniques. The questionnaire was framed and was distributed to face-to-face contact to the respondents in and around Kanchipuram. The cyber security measure dataset contained 15 attributes with 189 instances.

Classification Algorithms:

For the purpose of comparative analysis, a sum of 8 classification algorithms have been used for cyber security measures dataset. Various groups of classifiers such as Bayes, Functions, Lazy, Meta, Rule, Tree etc are available in WEKA. A fine set of classification algorithms such as NaiveBayes and WAODE from Bayes; SimpleLogistics and SMO from functions; JRip and NNge from rules; NBTree and RandomTree have been selected for evaluating their performance on the dataset.

1.1NaiveBayes: This is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. Its applications include text classification, spam filtering, sentiment analysis.

1.2WAEDO: It constructs the model called Weightly Averaged One-Dependence Estimators.

1.3SimpleLogistic: This is used for building linear logistic regression models. LogitBoost with some regression functions as base learners is used for fitting the logistic models.

1.4SMO: This implements John Platts's Sequential Minimal Optimization algorithm for training a support vector classifiers. It globally replaces all missing values and transforms nominal attributes in to binary ones.

1.5JRip: This implements a propositional rule learner. Repeated Incremental Pruning to produce Error Reduction(RIPPER) which was proposed by William W. Cohen as an optimized version IREP.

1.6NNge: This is Nearest Neighbor-like algorithm which uses non-nested generalized examplars.

1.7NBTree: This generates a decision tree with NaiveBayes classifier at the leaves.

1.8RandomTree: Performs no pruning. Class for constructing a tree that considers k randomly chosen attributes at each mode.

In this work of predicting the respondents' attitude towards cyber security measures all the above classification algorithms are used on datasets and the outcomes have been analyzed.

2. LITERATURE REVIEW

Shiju Sathayadevan and Surya Gangadharan(2014) have introduced an approach between computer science and criminal justice to develop a data mining procedure so that it would help solving crimes faster there by focusing mainly on crime factors each day[1].

Vinit Kumar Gunjan, Amit Kumar and Shrada Avdhanam(2013) have presented a brief overview of all about cyber criminals and crime with its evolution , types, case study, preventive majors and the department working to combat those crime[3].

Ahmed Lebbe Sayeth Saabith, Elankovan Sundararajan and Azuraliza Abu Bakar(2014) have said that feature selection would increase the accuracy of the classifier because it eliminates irrelevant attributes; reduce the Median Standard Error(MSE) and increases ROC to diagnosis the breast cancer dataset[6].

P Thamilselvan and Dr. J. G. R. Sathiaseelan(2015) have considered the performance of data mining algorithms in image classification which is analyzed based on classification accuracy and kappa coefficient.[8]

3. METHODOLOGY AND TOOL

WEKA tool is used for analyzing the performance of various classification algorithms by evaluating the parameters like accuracy, error rate, sensitivity, specificity, precision, F-measure, ROC area for the dataset. The implementation of this work was partitioned in to two phases:

Phase-I

In this phase, classification algorithms like NaiveBayes, WAODE, SimpleLogistic, SMO, JRip, NNge, NBTree, RandomTree were executed on cyber security measure dataset by various parameters and the best classifier for the data set is chosen. The parameters that are used to evaluate are:

1. Accuracy that calculates the proportion of correctly classified instances.
2. Error rate that calculates the proportion of incorrectly classified instances.
3. Sensitivity that evaluates the classifiers' capability to discover positive results.
4. Specificity that evaluates the classifiers' capability to discover negative results.
5. Precision that evaluates the retrieved instances that are significant.

Phase-II

In this phase, the best classification algorithm selected from Phase-I is used to predict the respondents' attitude towards cyber security measures.

4. PERFORMANCE EVALUATION OF VARIOUS CLASSIFICATION ALGORITHMS

The questionnaire has been designed to forecast the attitude of respondents towards cyber security measures. The main objective of this work is to find the best classifier from the performance evaluation of different classifiers of data mining techniques. The author collected 189 samples from the data among which there are 93 male respondents and 96 female respondents. 38 samples have been collected from respondents placed in government sectors, 40 samples from respondents doing business, 23 samples from students of schools and colleges, 36 sample from home makers and 52 samples from respondents placed in private sectors.

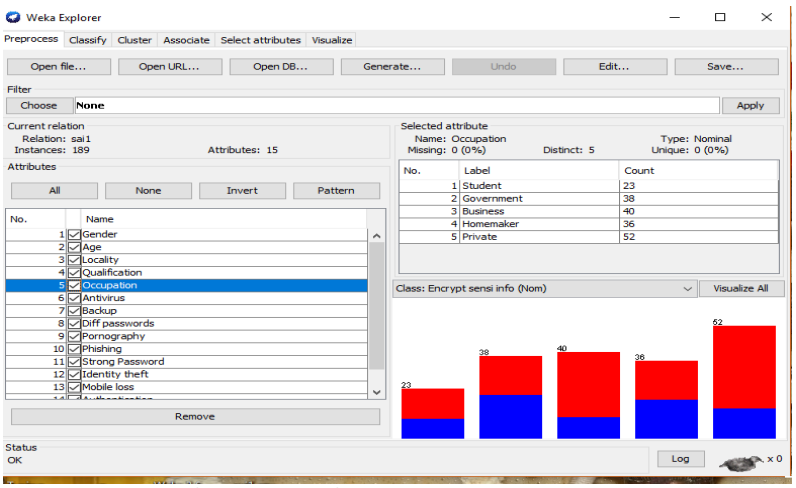

Figure1. Screen shot for categories of respondents based on their occupation in WEKA

Phase – I

Selecting the Explorer option in Application window of WEKA, the collected dataset have been put forward to a set of classification algorithms of WEKA. The classifier tab in WEKA enables us to access different classification algorithms for our dataset. Classification algorithms like NaiveBayes, WAODE, SimpleLogistic, SMO, JRip, NNge, NBTree, RandomTree are evaluated based on their accuracy, speed, error rate, sensitivity, specificity, precision, F-measure and ROC area.

Table 2: Comparison based on accuracy and error rate

Classifier	Correctly classified instances	Incorrectly classified instances	Speed (in sec)	Accuracy (%)	Error rate
NaiveBayes	156	33	0.01	82.53	0.1747
sWAODE	187	2	0.02	98.94	0.0106
SimpleLogistic	187	2	0.85	98.94	0.0106
SMO	186	3	0.32	92.63	0.0737
JRip	187	2	0.11	98.94	0.0106
NNge	186	3	0.1	98.41	0.0159
NBTree	185	4	1.1	97.88	0.0212
RandomTree	187	2	0.0	98.94	0.0106

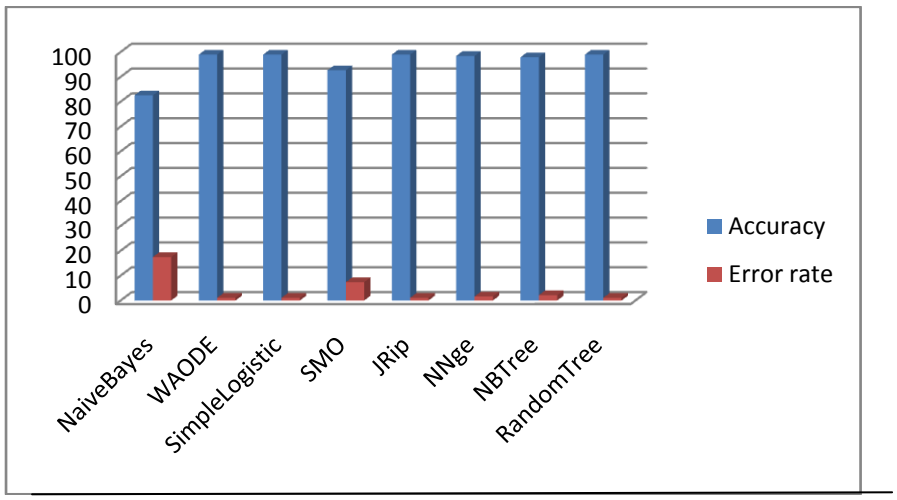

Figure2: Comparison of Accuracy and Error rate of different classification algorithms

From Table2, various classifiers were compared in terms of speed (time taken to build model), accuracy and error rate. WAODE, SimpleLogistic, JRip, RandomTree. An algorithm which has a worse error rate and highest accuracy will be chosen as it has more dominant classification capability. Though WAODE, SimpleLogistic, JRip and RandomTree has maximum accuracy than other classifiers regarding the time taken to build model RandomTree consumes less time from which it is concluded that RandomTree is best classifier concerning speed and accuracy. Moreover, NaïveBayes has least accuracy than other classifiers.

Table 3: Performance evaluation of different classification algorithms

Parameters	NaiveBayes	WAODE	Simple Logistic	SMO	JRip	NNge	NBTree	Random Tree
Confusion matrix	a b 54 17 a 16 102 b	a b 70 1 a 1 117 b	a b 70 1 a 1 117 b	a b 70 1 a 2 116 b	a b 70 1 a 1 117 b	a b 69 2 a 1 117 b	a b 69 2 a 2 116 b	a b 70 1 a 1 117 b
Sensitivity	0.7605	0.985	0.985	0.9859	0.9859	0.9718	0.9718	0.9859
Specificity	0.8644	0.9915	0.9915	0.9830	0.9915	0.9915	0.9830	0.9915
Accuracy	0.8253	0.9894	0.9894	0.9263	0.9894	0.9841	0.9788	0.9894
Precision	0.7714	0.9859	0.9859	0.9722	0.9859	0.9857	0.9718	0.9859
F-measure	0.7658	0.9854	0.9854	0.9789	0.9857	0.9784	0.9717	0.9857
ROC area	0.879	0.993	0.989	0.984	0.995	0.982	0.99	0.997

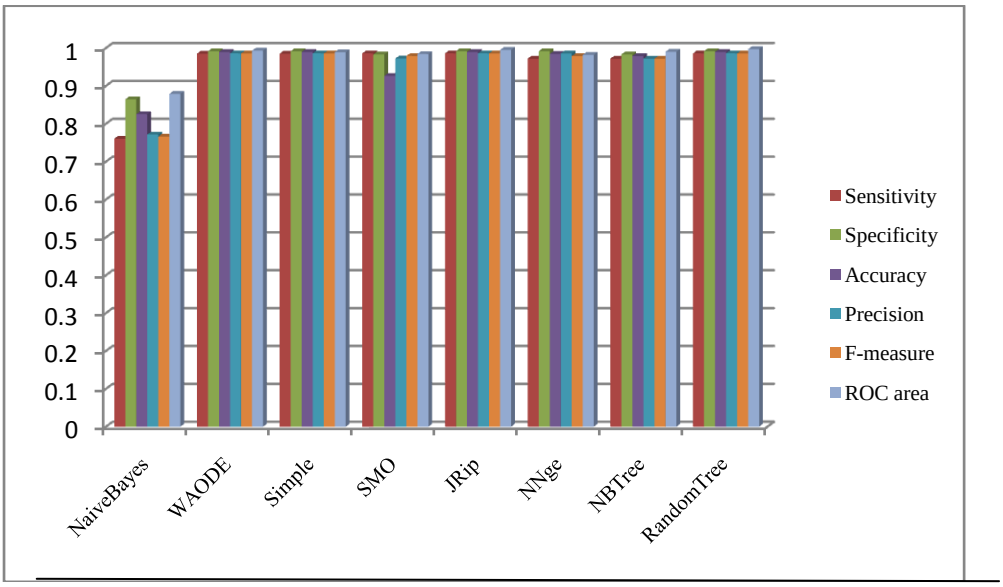

Figure3: Performance evaluation of different classification algorithms

From Table3, the confusion matrix of all classifier can be utilized which shows sensitivity, specificity, accuracy, precision, F-measure.

Phase – II

Prediction of Respondents' attitude towards Cyber Security measures using RandomTree

RandomTree can transact with both classification and regression problems. The working principle of classification is that the RandomTree classifier takes the input characteristic vector, classifies it with every tree in the forest, and outputs the class label that acknowledged the greater part of votes. All the trees are skilled with the similar parameters but on various training sets. The confusion matrix showed the measures behind the attitude of respondents towards cyber security. Here the prediction is done by the attributes occupation, different passwords, pornography, phishing, strong password, identity theft and mobile lost and the predicted occupations result is described below. The decision tree is formed based on which classification on the test data is done. In our study, confusion matrix (contingency table) has five classes, and so a 5*5 confusion matrix. The sum of diagonals in the matrix is the number of correctly classified instances, all others are incorrectly classified instances.

```
=== Confusion Matrix ===
a   b   c   d    e <-- classified as
8   2   0   0    0 | a = Private
1   31  0   1    0 | b = Government
0   1   1   0    0 | c = Business
1   0   0   130  0 | d = Home maker
0   0   0   0   13 | e = Students
```

The correctly classified instances are 183(96.8254%) which is the sum of the diagonals of confusion matrix(8+31+1+130+13) and the incorrectly classified instances are 6(3.1746%).

Table 4: Prediction based on respondents' occupation

Observed	**Occupation**	**Predicted values**					**% of correctly predicted**
		a	**b**	**c**	**d**	**e**	
	Private	8	2	0	0	0	80.0
	Government	1	31	0	1	0	93.93
	Business	0	1	1	0	0	50.0
	Homemaker	1	0	0	130	0	99.23
	Student	0	0	0	0	13	100

Among 183 correctly classified instances, 80% of respondents working in private sectors strongly agree; 93.93% of respondents working in various government sectors agree; 50% of respondents involved in business strongly disagree; 99.23% of respondents as home makers disagree and 100% of respondents as students are neutral.

5. CONCLUSION

Among NaiveBayes, WAODE, SimpleLogistic, SMO, JRip, NNge, NBTree and RandomTree classifiers, the RandomTree is chosen as best classifier for cyber security measures dataset since it has maximum accuracy, worse error rate and takes least time to construct the model and NaiveBayes has least accuracy. The performance evaluation of different classification algorithms on cyber security measures dataset is done on the basis of sensitivity, specificity, accuracy, precision and F- measure. A perfect classifier will have ROC area of 1 from which RandomTree is concluded as perfect classifier since it has 0.997. Finally, the RandomTree classification algorithm was used for predicting of respondents' attitude towards cyber security measures. In future this work can be extended by comparing classification algorithms between various data mining tools; by increasing more number of instances and more number of attributes.

6. REFERENCES

[1]. Shiju Sathyadevan and Surya Gangadharan, "Crime Analysis and Prediction using Data Mining", IEEE Trans, First International Conference on Networks and Computing, 2014.

[2]. M. Lakshmi Prasanthi, Tata A S K Ishwarya, "Cyber Crime: Prevention & Detection", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 3, Marh 2013.

[3]. Vinit Kumar Gunjan, Amit Kumar, Sharda Avdhanam, "A Survey of Cyber Crime in India", IEEE, 2013.

[4]. Janhavi J Deshmukh and Surbhi R Chaudhari, "Cyber crime in Indian scenario – A literature snapshot", International Journal of Conceptions on Computing and Information Technology, Vo;. 2, Issue 2, April 2014.

[5]. Hemlata Sahu, Shalinl Shrma and Seema Gondhalakar, "A Brief Overview on Data Mining Survey", International Journal of Computer Technology and Electronics Engineering, Vol. 1, Issue 3,2012.

[6]. Ahmed Lebbe Sayeth Saabith, Elankovan Sundararajan and Azuraliza Abu Bakar, "Comparative study on different Classification Techniques for Breast cancer dataset:, International Journal of Computer Science and Mobile Computing, Vol. 3, Issue 10, October 2014.

[7]. Vikas Chaurasia and Saurabh pal, "Early Prediction of Heart Diseases using Data mining Techniques", Caribbean Journal of Science and Technology, Vol.1, 2013.

[8]. P Thamilselvan and Dr. J. G. R. Sathiaseelan, "A comparative study of Data mining Algorithms for Image Classification", Inter Journal of Education and Management Engineering, Issue 2, 2015.

[9]. Karan Pruthi and Dr. Parteck Bhatia, "Application of Data mining in Predicting Placement of Students", IEEE, International Conference on Green Computing and Internet of Things, 2015.

[10]. Aman Kumar Sharma and Suruchi Sahni, "A Comparative Study of Classification Algorithms for Spam Email Data analysis", International Journal on Computer Science and Engineering, Vol. 3, Issue 5, May 2011.

[11]. M. Mayilvaganan and D. Kalpanadevi, "Comparison of Classification Techniques for predicting the performance of Students Academic Environment", IEEE, international Conference on Communication and Network Technologies, 2014.