IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

MALICIOUS ACTIVITY DETECTION

Annesha Sengupta, Aditya Pulivendula, Srajan Sadal, K. Vijay Naga Kumar,

Dr D. Ramakrishna

Student, Student, Student, Student, Assistant professor

Computer Science Engineering

GITAM Uinversity, Visakhapatnam, India

Abstract : Because of the rapid changes and variations in the internet over the past several years, the accuracy in detecting harmful behaviour online may be seen as a highly complex issue. In terms of spotting harmful behaviour, classification algorithms' potential is thriving. This research study adds to the use of several categorization algorithms for the purpose of detecting harmful activity in various internet zones. Algorithms for categorization include decision trees, support vector machines, and random forests. The information, which includes a number of characteristics, was gathered from several cyber security research organisations. The recall, precision, f1-score accuracy, and execution time of the classification algorithms were evaluated and trained using a variety of testing and training conditions. Decision trees are among the quickest models to run in terms of execution time, and decision tree classifiers are the superior choice in terms of run time.

INTRODUCTION

The study on the detection of the malicious activity demands the involvement of expertise in different fields that may cause a huge impact on the daily life. With the evolvement of the technology, machine learning can be considered as one of most helpful technology in the aspect of malicious activity detection. Different research studies suggest that applications of machine learning are deficient in some fields. detection of malicious activity can be considered as the crucial issue that has to be tackled by the cyber security experts as the issue is closely related to the human economy and daily life of the mankind. Heavy malicious activity can be considered as the cause for the online disasters such as data breaches and economical frauds that are being encountered by the mankind across the planet. When it comes to the accuracy regarding the malicious activity researching can be considered as a crucial aspect for many of the nations whose annual income is hugely impacted. Due to drastic and unstable behaviour of the hackers, few techniques like statistical techniques will obviously fail to provide the government and researchers with the accurate data regarding the malicious activity detection. Unstable and nonlinear data of malicious activity makes the machine learning techniques an impactful option for the sake of accuracy. The main intention of this research study is to provide an easy access regarding the data of machine learning techniques and approaches that are helpful in detecting the malicious activity.

detection of malicious activity can be helpful in preventing the nature complications like data breaches so that the properties and revenue of the people will be saved. This malicious activity detection system helps in looking after managing the water resources. Information regarding the malicious activity in prior also helps the farmers and the agriculture industry to manage their crops and fields in a much better way so that it results in greater economic returns. Unstable and fluctuated malicious activity timetable and the uncertainty regarding its quantity makes the detection of malicious activity a complicated and challenging issue to look after for the cyber security researchers. There are several services that cyber security researchers organisations provide like forecasting which stands out as the top more priority for the nations all over the planet. The task of malicious activity detection is pretty complicated as it demands a huge number of experts and specialised and also the calls were made without any sort of certainty.

RELATED WORK.

In the research paper researchers Goswami and Srividya have explained the results of the combination of RNN and TDNN features and they have concluded their work which says that the composite models provide us the better accuracy rates than the model that runs on the single algorithm or model. The researchers have used the networks like multilayer feed forward neural networks for the sake of detecting the malicious activity in Indian terrain. These network models can be considered as the models with numerous numbers of input parameters that are going to be analysed. Heavy malicious activity can be considered as the cause for the internet disasters such as data breaches and droughts that are being encountered by the mankind across the planet. When it comes to the accuracy regarding the malicious activity forecasting can be considered as a crucial aspect for many of the nation's whose annual income is hugely impacted by the agriculture especially the nations like India. Due to drastic and unstable behaviour of the atmosphere, few techniques like statistical techniques will obviously fail to provide the government and cyber security researchers organisations with the accurate data regarding the malicious activity forecasting. Unstable and nonlinear data of malicious activity makes the machine learning techniques an impactful option for the sake of accuracy.

Researchers M. Chattopadhyay have utilised few parameters with lesser or very minimum and higher for the sake of the malicious activity and forecasting of malicious activity. The researchers have used conjugate decent and Levenberg–Marquardt learning algorithm for the sake of learning algorithm for training. Taking its authentic acceptance in the field of epidemiological research, the method of logistic regression is usually involved in most of the aspects of the knowledge. Malicious activity can be considered as one of the most crucial happening of the internet system on this planet. It is now known that the intensity and variability of the malicious activity impact on the agriculture, nature, mankind and also on the biological balance of the globe. So we can clearly observe a essentiality where we have to be able to predict the malicious activity timings and quantity of the malicious activity by generating the predictors. In this research study we have utilised the logistic regression for detecting the malicious activity. We have observed that it is evident that the data regarding the internet were subjected towards the errors of gross recording as this complication goes unnoticed when it steps in to the stage of analysis. During our research study we have utilised very new screening methods to collect and to correct the data regarding the internet.

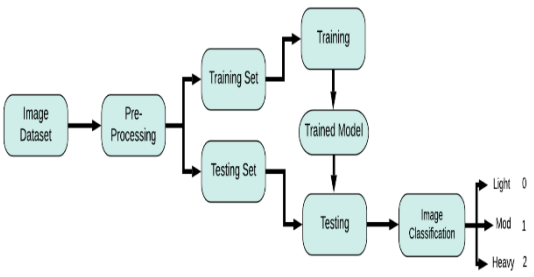
METHODOLOGY

Unstable and fluctuated malicious activity timetable and the uncertainty regarding its quantity makes the detection of malicious activity a complicated and challenging issue to look after for the cyber security researcher's. There are several services that cyber security researcher's organisations provide like forecasting which stands out as the top more priority for the nations all over the planet. The task of malicious activity detection is pretty complicated as it demands a huge number of experts and specialised and also the calls were made without any sort of certainty. Here are few machine learning techniques which can be used to predict the malicious activity.

1. Support vector machine
2. KNN and ANN
3. Random forest
4. Decision tree

SUPPORT VECTOR MACHINE(SVM)

Support vector machine can be considered as a type of feed forward network. SVMs can perform the tasks like nonlinear regression, classification of the patterns and so on. support vector machines are used for supervised learning because of features like the ability of better generalisation when compared to the NN models and the solution of SVM is completely identical, absent from the local minima and it is also optimal. SVM is also applicable for the data that is non vertical and very minimal parameters will be required for the sake of tuning the learning. Scientists have rarely applied this method as a malicious activity detection method and the results that were obtained were completely acceptable.



Numerous recent research papers regarding the usage of SVM on detection of malicious activity for the sake of different classifications and regression problems. We have studied the use of SVMs for the sake of classification of malicious activity

on very minute dataset with a training set 10 percent and the output that we obtained can be noted as the binary classification regarding rain or no rain daily.

DECISION TREE

Decision tree is one of the popular and powerful tool of machine learning that can be used for the sake of detection of malicious activity. Decision trees resembles the rules and instructions that can be understood by the mankind and can be utilised in the aspect of knowledge systems like database. Decision can be considered as the classifier in the form of the structure that resembles the tree that consists of different decision nodes that specifies a test regarding the tests on a attribute that is completely single. The leaf node works as the indication regarding the values of attributes that are collected for the sake of targets. The edge node splits the path of the attribute, a disjunction of a test to generate the final decision. Decision trees can be considered as impactful in classifying the examples or instances by initiating the root of the tree node and it moves through the leaf node. For the sake detectionas it generates not so complicated but potential tree with minimal depth possible.

Advantages of Decision trees

- decision rules are very much easy to interrupt.
- decision tree is completely nonparametric and thus it not really complicated to incorporate a wide range of categorical layers of the data.
- selecting the unimodal training data is not really needed.
- robust when it comes to the outliners in the training of data

Disadvantages of decision tree

- very poor results will be obtained when the decision trees tends to get along with the training data
- detection that exists beyond maximum and minimum limits of training data's responsible variables is not at all possible.

Application od decision tree.

- Decision tree can be used in differentiating the useful emails from spam emails
- Decision tree can be vividly used in the field of medicine.

RANDOM FOREST

You may think of random forests as the categorization approach that is employed. The creation of decision trees for the purpose of training and matching the same data with them is a component of the random forest algorithm's operation. To determine the significance of numerous factors for the purpose of classification difficulties, random forests are used. To comprehend the significance of a single variable in a certain data set, use the formula Dn=(Xi, Yi)i= 1n.

To this particular set of data, we have attempted to fit a random forest. The computation for each data error will be performed throughout the fitting phase, and it was also averaged across the random forest. The values of the i-th feature may be thought of as permuted inside the training data, and the error has been computed on this specific dataset, in order to grasp the potentiality of the i-th feature after the training phase. By calculating the mean values of the difference between the mistakes after and before permutation for the sake of trees, the essential scores for the i-th feature were determined. The standard deviation of these specific differences was used to generalise the scores.The characteristics that yielded the highest values for the purposes of this score are more important than the characteristics that produced lower values. Random forest provides us with information about a variable's importance as well as the closeness of the information points to one another.

Advantages of random forest

- Random forest algorithms helps in acquiring accurate detection for most of the types of applications
- Random forest helps in understanding the cruciality of every feature with respect to training data set

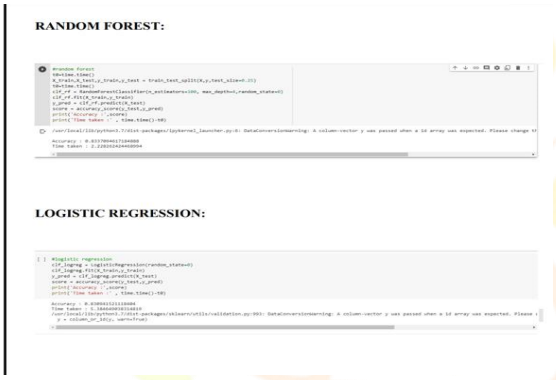
- Random forest helps in measuring the pairwise proximity within the samples

Disadvantages of random forests

- For the data that includes the variables or the categorical variables with various number of levels
- Random forest can be considered as the biased for the sake of those particular attributes with much number of levels.
- In in case the data consists the groups of connected features of frequent relevance for the sake of the output
- The process of integrating the randomness in to the trees
- The process of splitting the objective in to each node through the process of optimization

Application of random forests

- Random forests can be utilised for the sake of classification of image for analysation of pixels
- Random forests can be utilized in the aspects of bioinformatics for the sake of analysing the complicated biological data
- random forests can be utilised for segmentation of the videos



KNN

K-Nearest Neighbors (KNN) is a supervised machine learning algorithm that can be used for classification problems, including malicious activity recognition. In this context, KNN can be used to classify network traffic as either normal or malicious based on certain features or characteristics of the traffic.

To use KNN for malicious activity recognition, you would first need to gather a dataset of network traffic that has been labeled as either normal or malicious. You would then extract features from this dataset, such as packet size, protocol type, and source/destination IP address, and use these features to train the KNN model. Once the model has been trained, you can use it to classify new network traffic as either normal or malicious based on the extracted features.

One important consideration when using KNN for malicious activity recognition is selecting the appropriate value of k, which represents the number of nearest neighbors to consider when making a classification. Selecting the appropriate value of k is crucial to achieving accurate classification results. If k is too small, the model may be overly sensitive to noise and outliers, while if k is too large, the model may not be able to capture the nuances of the data.

Overall, KNN can be an effective algorithm for recognizing malicious activity in network traffic, but it should be used in combination with other techniques and tools to provide a comprehensive security solution.

ANN

Artificial Neural Networks (ANN) can also be used for malicious activity detection. ANN is a type of machine learning algorithm that is modeled after the structure and function of the human brain. ANN can learn complex patterns and relationships in data, making it useful for detecting malicious activity in network traffic.

To use ANN for malicious activity detection, you would first need to gather a dataset of network traffic that has been labeled as either normal or malicious. You would then extract features from this dataset, such as packet size, protocol type, and source/destination IP address, and use these features to train the ANN model.

The layers of artificial neurons that make up the ANN model process and examine the characteristics of network traffic. By changing the weights of the connections between the neurons in response to the training data, the model is trained. When the model has been trained, you can use it to categorise fresh network traffic as malicious or normal based on the attributes that were retrieved.One advantage of ANN for malicious activity detection is its ability to handle large amounts of data and detect complex patterns that may not be easily recognizable using other techniques.

Nevertheless, ANN may be computationally demanding and takes a lot of training data, therefore it's crucial to take into account the resources available before utilising this strategy.

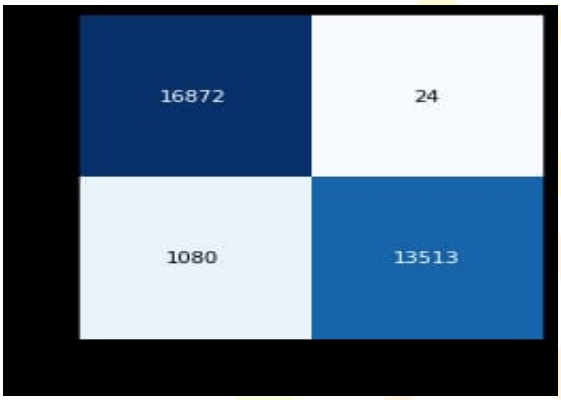
Therefore, ANN has the potential to be a potent tool for spotting malicious behaviour in network data, but it should be used in conjunction with other methods and technologies to offer a complete security solution.

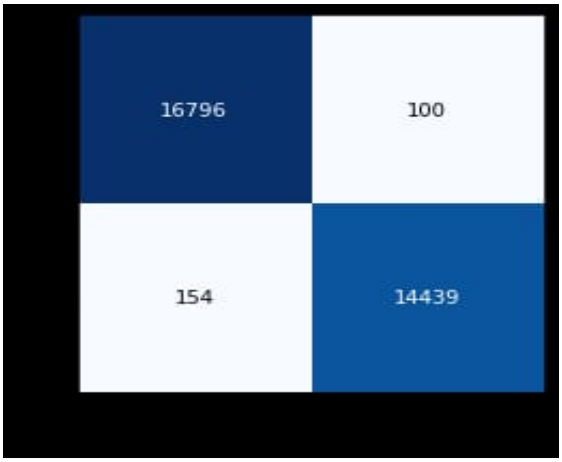
COMPARITIVE ANALYSIS

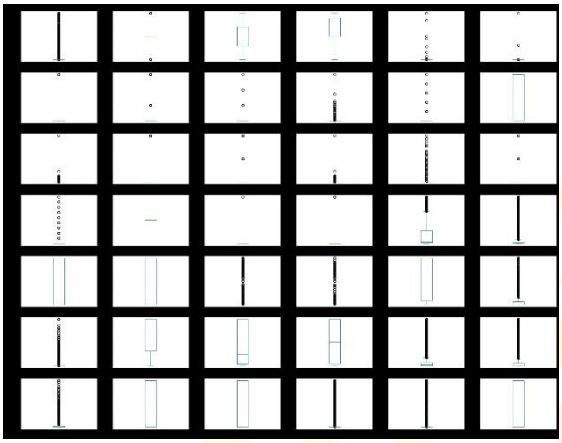
RESULTS

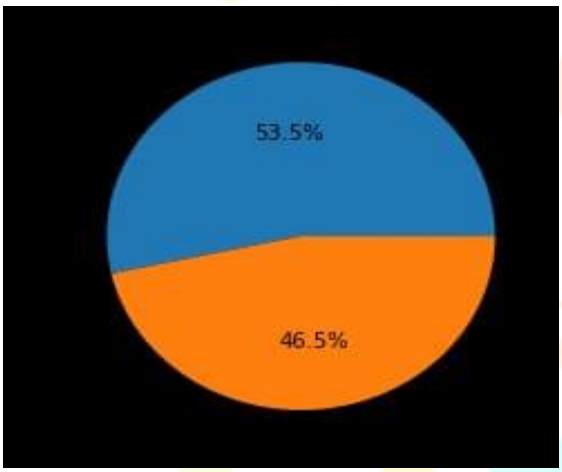
For malicious activity detection in different scenarios, we have used four different machine learning techniques such as random forests, decision tree and supervised vector machine. In our research study we have performed the testing and training different datasets and the results were analysed and the accuracy obtained by different machine learning techniques were shown below. The classification systems for the sake of successful detection in which the data that used as the input data was passed through pre-processing stage and was normalised and cleaned until the process of classification. The accuracy that we obtained with the machine learning techniques were shown below.

When it comes to future scope we propose that future forecasts should be performed by testing and training further the classification techniques and the attributes of internet on various dates.









Accuracy is completely dependent on the machine learning techniques that are mostly used for the estimating the accuracy. With this we conclude that decision tree can be considered as the best and accurate machine learning technique that detects the malicious activity

CONCLUSION

Malicious activity is really an important in every aspect that exists on the globe. It is obviously better and will be a helpful to predict the future rains and the quantity of the malicious activity and the steps that are needed to be taken to prevent the loss on the planet. In our research work, we have proposed a system that could predict the malicious activity or forecast the malicious activity of a particular geolocation where the classifier is trained with the techniques of machine learning with the help of prior internet data. The classification was done with the help of four different machine learning algorithms namely random forests, decision trees, supervised vector machines.. The algorithms that were used to extract the features of the datasets from the datasets like training sets and the algorithms of machine learning were able to predict the malicious activity with utmost accuracy. Several different parameters such as computational time or the efficiency of the processes which was found to be far better than the methods that were traditionally used for the sake of malicious activity detection. The future research should indulge the utilisation of many other machine learning and deep learning techniques and working on the security of the systems should also be considered.

REFERENCE

[1] P. Goswami and Srividya, “A novel Neural Network design for long range detection of malicious activity pattern,” Current Sci.(Bangalore), vol. 70, no. 6, pp. 447-457, 1996.

[2] C. Venkatesanet, S. D. Raskar , S. S. Tambe , B. D. Kulkarni , and R. N. Keshavamurty , "detection of all India summer monsoon malicious activity using Error-Back-Propagation Neural Networks," Meteorology and Atmospheric Physics, pp. 225-240, 1997.

[3] A. K. Sahai, M. K. Soman, and V. Satyan, "All India summer monsoon malicious activity detection using an Artificial Neural Network," Internet dynamics, vol. 16, no. 4, pp. 291-302, 2000.

[4] N. S. Philip and K. B. Joseph, "On the predictability of malicious activity in Kerala-An application of ABF neural network," Computational ScienceICCS, Springer Berlin Heidelberg, pp. 1-12, 2001.

[5] N. S. Philip and K. B. Joseph, "A Neural Network tool for analyzing trends in malicious activity," Comput. & Geosci.,vol. 29, no. 2, pp. 215-223, 2003.

[6] N. Chantasut, C. Charoenjit, and C. Tanprasert, "Predictive mining of malicious activity detections using artificial neural networks for Chao Phraya River," 4th Int Conf. of the Asian Federation of Inform. Technology in Agriculture and the 2nd World Congr. on Comput. in Agriculture and Internet Resources, Bangkok, Thailand, pp. 117-122, 2004.

[7] V. K. Somvanshi, O. P. Pandey, P. K. Agrawal, N.V.Kalanker1, M.Ravi Prakash, and Ramesh Chand, "Modeling and detectionNeural Network and ARIMA techniques," J. Ind. Geophys. Union, vol. 10, no. 2, pp. 141-151, 2006.

[8] S. Chattopadhyay, "Anticipation of summer monsoon malicious activity over India by Artificial Neural Network with Conjugate Gradient Descent Learning," arXiv preprint nlin/0611010, pp. 2-14, 2006.

[9] S. Chattopadhyay and M. Chattopadhyay, "A Soft Computing technique in malicious activity forecasting," Int. Conf. on IT, HIT, pp. 19-21, 2007.

[10] S. Chattopadhyay and G. Chattopadhyay, "Comparative study among different neural net learning algorithms applied to malicious activity time series, "Cyber security researchers applicat., vol. 15, no. 2, pp. 273-280, 2008.

[11] P.Guhathakurta, "Long lead monsoon malicious activity detection using deterministic Artificial Neural Network model," Meteorology and Atmospheric Physics 101, pp. 93-108, 2008.

[12] C. L. Wu, K. W. Chau, and C. Fan, "detection of malicious activity time series using Modular Artificial Neural Networks coupled with datapreprocessing techniques," J. of hydrology, vol. 389, no. 1, pp. 146- 167, 2010.

[13] K. K. Htike and O. O. Khalifa, "Malicious activity forecasting models using Focused Time-Delay Neural Networks," Comput. and Commun. Eng. (ICCCE), Int. Conf. on IEEE, 2010.

[14] S. Kannan , Subimal Ghosh, "detection of daily malicious activity state in a river basin using statistical downscaling from GCM output", Springer-Verlag, July- 2010.

[15] M.Kannan, S.Prabhakaran, P.Ramachandran, "Malicious activity Forecasting Using Data Mining Technique", International Journal of Engineering and Technology Vol.2 (6), 397-401, 2010.

[16] G. Geetha and R. S. Selvaraj, "detection of monthly malicious activity in Chennai using Back Propagation Neural Network model," Int. J. of Eng. Sci. and Technology, vol. 3, no. 1, pp. 211 213, 2011.

[17] M. A. Sharma and J. B. Singh, "Comparative Study of malicious activity forecasting models," New York Sci. J., pp. 115-120, 2011.

[18] J. Abbot and J. Marohasy, "Application of Artificial Neural Networks to malicious activity forecasting in Queensland, Australia," Advances in Atmospheric Sci.,vol. 29, no. 4, pp. 717-730, 2012.

[19] A. Kumar, A. Kumar, R. Ranjan, and S. Kumar, "A malicious activity detection model using artificial neural network," Control and Syst. Graduate Research Colloq. (ICSGRC), pp. 82-87, 2012.

[20] R. R. Deshpande, "On the malicious activity time series detection using Multilayer Perceptron Artificial Neural Network," Int. J. of Emerging Technology and Advanced Eng., vol. 2, no. 1, pp. 148-153, 2012.

[21] Soo-Yeon Ji, Sharad Sharma, Byunggu Yu, Dong Hyun Jeong, "Designing a Rule-Based Hourly Malicious activity detection Model", IEEE IRI 2012, August – 2012.

[22] G. Shrivastava, S. Karmakar, and M. K. Kowar, "BPN model for long range forecast of monsoon malicious activity over a very small geographical region and its verification for 2012," Geofizika, vol. 30, no. 2, pp. 155-172, 2013.