Comparative Analysis of Machine Learning Algorithm for Anomaly Detection: A Literature Survey

Manohar Vemula[1], Shaik Balasaidulu[2]
[1]*MCA, MCPD, CSM, Software Engineer, Hyderabad*
[2]*MCA, Osmania University*

Abstract- **Intrusion detection system used to discover illegal and avoidable activities at accessing or manipulating computer system. To improve system correctness and decrease false positive rate based on DARPA 98 and later the updated version which shoes some arithmetic issues, researcher focuses on various dataset. Manu researchers give their efforts to explore the dataset by different techniques. In this paper machine learning based methods which are one of the types of anomaly detection, techniques are discussed.**

Index Terms- Intrusion detection, KDD Cup 99 dataset.

1. INTRODUCTION

With the growing number of internet users, the threats to it also increase day by day. Intruders may be from external the host or the network or legal users of the network. To identify various threats, intrusion detection systems are necessary. Intrusion detection is the process of monitoring the actions that are happening in the system or network and analyse them for sign of possible incidents, which are infringement or threats to computer security policies [3]. There are two types of intrusion detection approaches first is misuse detection where we model attack behaviour or quality using intrusion inspection data and anomaly detection, which is model used to normal convention behaviour. Normally in the commercial intrusion detection system, the signature or misuse based approach is followed but the anomaly based approach is efficient using the machine learning method [2].

2. LITERATURE REVIEW

In the year 2011 K. Hanumantha Rao, G. Srinivas, Ankam Damodhar and M. Vikash Krishna[4] states that with the help of data mining it is possible to search large amount data of characteristic rules and patterns which is use to detect intrusion, if applied to network monitoring data recorded on a host or in a network. In this paper, they present machine learning method to classifying anomalous and normal activities in a computer network with supervised and unsupervised algorithm that have the best efficiency. They used K-means clustering and the Id3 decision tree which improves the system classification performance.

In the year 2013 Sneh Lata Pundir and Amrita[2] states that intrusion detection is very common and dangerous problem in recent scenario and they have many mechanism to solve it but we need to maintain system performance of the intrusion detection system. In this paper they used feature selection approach, for the intrusion detection system should be fast and effective an optimal feature subset should be made. Out of all the subset which they certain out of 41 features, the best presentation is given by the subset of 15 features which is almost equal to the performance given by the set of 41 features and time taken to construct the model by the subset of 15 features is less than the time taken by the set of 41 features.

In the year 2013 Kriangkrai Limthong [3] states that Zero day attack is one of the variety of anomalies and it is very important and real time challenge for both network operators and researchers. They proposed an alternative detection technique which is based on combination of feature space and time series to automatically detect anomalies in real time for using machine learning algorithm. They conducted experiments with real time traffic in real world and compared the detection performance and their result show proposed technique do better in task of anomaly detection and has a good possibility for applying in real time system.

In the year 2012 Riti Lath and Manish Shrivastava proposed analysis of kdd cup'99 dataset using classification and normalisation to generate good result. Analysis of the dataset is performed using different classification techniques that is K-Mean which is based on clustering, K-Nearest neighbour, Support vector machine. To classify and normalise that data first analysed that flat result then preprocessed data is used and for preprocessing statistical normalisation has been used. By applying classification algorithm in data without any preprocessing they generate good result but when data get normalised then degrades potential of classification techniques. Classification of anomaly separately gives not excellent result and takes much more time for execution. They have concluded that after evaluation of all classification algorithm K-Nearest neighbour provides better result as compared to both K-means and SVM but it takes more execution time.

In the year 2013 Mr. Suresh Kashyap, Ms. Pooja Agrawal, Mr. Vikas Chnadra Pandey, Mr. Suraj Prasad Keshri proposed comparison between different algorithms for training neural network with big amount of data and find which algorithm is more suitable for intrusion detection data. The two algorithms that is Error back propagation (EPB) which is the most worn training algorithm for feed forward neural network algorithm and the Redial basis function (RBF) neural network which is based on supervised based neural network learning algorithm are compared in this paper and finds Redial basis function gives better result than Error back propagation.

In the year 2013 Amira Sayed A. Aziz, Aboul Ella Hassanien, Sanna EL-Ola Hanafy, M.F.Tolba proposed multilayer hybrid machine learning intrusion detection system to achieve high efficiency and improve the detection and classification rate accuracy stimulated by immune system with negative selection move towards. In the first layer, for feature selection principal component analysis algorithm was used. After that genetic algorithm was applied to generate anomaly detectors, which are able to find difference between noram and abnormal behaviour in second layer. It is followed by applying several classifiers like naive bayes, multilayer perceptron neural network and decision tree to improve the detection accurateness. In this paper they find Decision tree generates best result for anomaly detection. In the year 2014 Goverdhan Reddy Jidiga, and P. Sammulal proposed a brief study about performance criteria used in anomaly detection to specify boundaries based on statistical mathematical statistics in emerging application used in real world. To classify the, records new rule based decision tree (RBDT) machine learning approach can be used. In this paper 2012 Devendra Kailashiya and Dr. R. C. Jain proposed intrusion detection system with the help of decision tree algorithm to improve their accuracy rate. In this paper they used supervised learning neural network with preprocessing step for intrusion detection. To generate the sample for the original dataset he used stratified weighted sampling technique and this sample applied on the proposed algorithm. The result showed the proposed system created higher accurateness and low error in identifying whether the records are usual or attack one. Mohammad Khubeb Siddiqui and Shams Naahid proposed analysis of dataset to improve security for intrusion detection. In this paper they present they focused on generate relationship between the attack types and the protocol used by the attacker, by using clustering technique which is k-means clustering algorithm. The investigation exposed many investing result about the protocol and attack type used by the attacker. In the year 2013 S. Revathi and Dr. A. Malathi states that a system to improve system accuracy and to reduce false positive based on DARPA 98 and later the updated version of KDD cup 99 dataset. In this paper they analyse the NSL-KDD dataset that solves some of the issues of KDDcup99. The analysis result shows that NSL-KDD dataset is very ideal for comparing more intrusion detection model.

3. MACHINE LEARNING

Machine learning studies computer algorithms for learning to do stuff. We might for instance, be interested in learning to complete a task, or to make accurate predictions, or to behave intelligently. The learning that is being done is always based on some sort of observations or data. Machine learning is about learning to do better in the future based on what was experienced in the past. The emphasis of machine learning is on automatic methods. In other words, the goal is to devise learning algorithms that do the learning automatically without human intervention. Machine learning is a core subarea of artificial intelligence. It is very unlikely that we will be able to build any kind of intelligent system capable of any of the facilities that we associate with intelligence. Machine learning techniques based on explicit or implicit model. The explicit is actual values seen in the data points and the implicit patterns that are seen across the values.

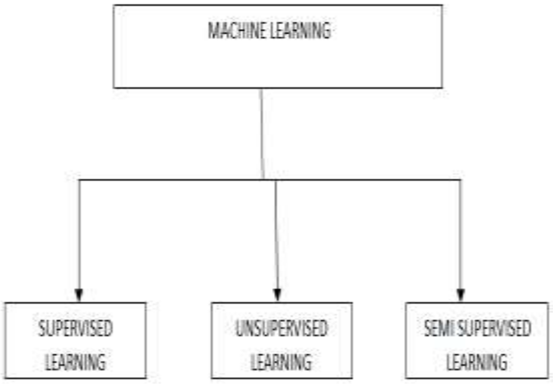

Fig. Types of Machine Learning

3.1 Types of Machine Learning:
Machine learning basically divided into three types:
3.1.1 supervised learning,
3.1.2 unsupervised learning,
3.1.3 semi-supervised learning.

3.1.1 Supervised learning: - supervised learning is the data mining task of inferring form labeled training data. The training data consist of a set of training examples. Each example is a pair consisting of an input object and a desired output value. A supervised learning algorithms analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. This requires the learning algorithm to generalised from the training to unseen situations in a “reasonable ” way.

3.1.2 Unsupervised learning: - in data mining, the problem of unsupervised learning is that trying to find hidden structure in unlabeled data. Since the example given to the learner are unlabeled, there is no error or reward signal to evaluate a potential solution. Unsupervised studying models a set of information, like clustering. In machine studying, without supervision studying is a type of problems in which one looks for to determine how the information are structured. It is recognized from supervised studying in that the student is given only unlabeled illustrations. Unsupervised studying is carefully relevant to the issue of solidity evaluation in research. However without supervision studying also involves many other techniques that seek in conclusion and explain key feature of the information.

3.1.3 Semi-supervised learning :- semi supervised learning is class of supervised learning tasks and techniques that also make use of unlabeled data for training –typically a small amount of labelled data with a large amount of unlabeled data. Semi supervised learning falls between unsupervised learning (without any labelled training data) and supervised learning (with completely labelled training data). Many machine learning researchers have found that unlabeled data, when used in conjunction with a small amount of labelled data, can produce considerable improvement in learning accuracy. The acquisition of labelled data for a leaning problem often requires a skilled human agent or a physical experiment.

4. CONCLUSION AND FUTURE WORK

In this paper, we have studying machine learning techniques. Anomalies in website performance are very regular. As we have established select studies on by one in the area of community scheme detection machine and we have got mentioned their result one at a time also. Most of time they may be brief and most successful have an effect on a small part of the customers. However, in e-trade an anomaly may be very extremely –priced. Just one minute with underperformance website means enormous loss for a enormous ecommerce retailer. As in line with this survey paper we have reached the give up that we are allowing an internet mostly based gadget for emerging E-trade that uses rule mostly based algorithm and SSH-2 set of rules to protect our statistics and transaction towards unauthorised get entry to and user.

REFERENCES

[1] K. Hanumantha Rao, G. Srinivas, Ankam Damodhar and M. Vikas Krishna, "Implementation of Anomaly Detection Technique using Machine Learning Algorithms" international journal of computer science and telecommunications 2011.

[2] Sneh Lata Pundir and amrita, "Feature selection using Random forest in intrusion detection system" international journal of advances in engineering and technology 2013.

[3] Kriangkrai limthong, "Real Time Computer Network Anomaly Detection using Machine Learning Techniques" journal of advances in computer networks 2013.

[4] Riti Lath, Manish Shrivastava, "Analytical Study of Different Classification Technique for KDD Cup data'99" international journal of applied information system 2012.

[5] Mr. Suresh kashyap , Ms.Pooja Agrawal , Mr. Vikas Chandra Pandey , Mr. Suraj Prasad Keshri, "Soft computimg based Classification Technique using KDD 99 Data Set for Intrusion Detection System" international journal of advanced research in electrical, electronics and instrumentation engineering, 2013.

[6] Amira Sayed A. Aziz, Aboul Ella Hassanien, Sanna EL-Ola Hanafy, M.F.Tolba, "Multi-layer hybrid machine learning techniques for anomalies detection and classification approach" IEEE 2013.

[7] Goverdhan Reddy Jidiga, P.Sannulal, "Anomaly Detection using Machine Learning with a case study" IEEE international conference on advanced communication control and computing technologies(ICACCCT) 2014.

[8] Devendra kailashiya, Dr. R.C. jain, "Improve Intrusion Detection using Decision tree with Sampling" international journal computer technology and applications 2012.

[9] Mohammad Khubeb Siddiqui and shams Naahid, "analysis of KDD CUP 99 Dataset using Clustering based Data Mining" international journal of database and application.

[10] S. Revathi, Dr. A. Malathi, " A Detailed Analysis on NSL-KDD dataset using various Machine Learning techniques for intrusion Detection" international journal of engineering research and technology 2013.