Comparative Analysis of Different Data Mining Technique Based Intrusion Detection System: A Review

Sushmita Sagar
M.Tech. Research Scholar, Department of CSE,SIRTS (Bhopal)

Prof. Amit Shrivastava
Associate Prof., Department of CSE,SIRTS(Bhopal)

Prof. Chetan Gupta
Assistant Prof., Department of CSE,SIRTS(Bhopal)

Abstract-Increased connectivity and also the use of the web have exposed the security and safety issue in front of the organizations, therefore need to use of intrusion detection system to protect data system and communication network from malicious attacks or unauthorized access. Associate intrusion detection system (IDS) may be a security system that monitors computer systems and network traffic, analyze that traffic to spot attainable security breaches and raise alerts. Associate IDS triggers thousands of alerts per day that is tough for human users to investigate them and take acceptable actions. In this paper describe the different data mining technique based intrusion detection system and their problem to facet in the implementation of these system of intrusion detection. This paper also elaborate the optimization method like plant growth optimization and feature reduction based classifier like multiple support vector machine, that's our proposed method to improve the detection rate and accuracy of the system and overcome the problem which identified from the literature. It's vital to cut back the warning alerts, showing intelligence integrate and correlate them so as to present a high level read of the detected security issue to the administrator.

KEYWORDS:IDS, Data Mining Approach, PGO, SVM, KDD Cup-99.

I. INTRODUCTION

Detection and prevention of anomaly over the internet in real time scenario is a big challenge. The versatile feature and dynamic nature of anomaly attack emerges the issue of discovery and avoidance of attack. The abnormality assault is umbrella of different assault such DOS, Probe, U2R, R2L and numerous mixes of assaults. For the discovery of abnormality assault utilized firewall, interruption identification framework, antivirus and numerous more application programming are used [1, 2, and 3]. The handling of identification is ease back because of vast number of interruption characteristic, now different creators utilized machine learning method and highlight diminishment prepare for the characterization and detection in intrusion detection system. Big data is the accumulation of extensive informational collections and it winds up plainly hard handling utilizing reasonable customary information preparing applications or database administration instruments. The difficulties incorporate procuring, putting away, seeking, sharing, exchanging, breaking down and imagining.

Discussed assess the best in class thinks about on the issue of inconsistency recognition in PC systems. They give an intricate portrayal of the irregularity discovery issue, and delineate the distinctive arrangements of its answers. They likewise outline some current best in class arrangements on the system level, and portray current patterns in dealing with malware-instigated oddities in cell phone systems [1]. This Work is planned to exhibit grew inconsistency identification framework in secure distributed computing condition, demonstrate its hypothetical depiction and lead fitting reenactment. The outcome exhibits that the created framework gives the high rate of oddity discovery in secure distributed computing condition [2]. Described to the look at their approach against other oddity recognition frameworks on genuine and manufactured information with fluctuating time-arrangement qualities. They found that their system takes into account 50-60% change in accuracy and review for an assortment of utilization cases. Both the information and the structure are being publicly released [3].

A huge database, administrations, applications, programming and assets are a basic piece of this innovation. It has the ability to work a program or applications on various associated PCs at the same time and allows the clients to enter applications and assets through a web program or web benefit by means of the Internet whenever and anyplace [4]. The study recognizes the botnet inquire about into three fields: Anomaly Detection - Botnets, botnet assaults and most recent botnet practices, and methods for safeguarding against botnets. While botnets are limitless, the examination and illumination for botnets are still in their youth. The paper additionally compresses the current research and suggests future bearings for Botnet inquire about [5]. This talked about work displays a nonspecific sensor-situated data framework in view of Hadoop group (SOIS-Hadoop). NoSQL database is utilized to store and deal with the heterogeneous tangible information; Hadoop/MapReduce programming worldview is utilized to improve the parallelism of information recovery and examination [6]. A scaling heuristic is talked about to decide the degree of scaling important to diminish approaching execution punishment. FMR encourages down to earth appropriation by being actualized as an arrangement of libraries and scripts that require no progressions to the hidden source code of Hadoop [7]. They introduce Tejo, a regulated irregularity location plot for New SQL databases. Un-like universally useful irregularity ID for the cloud, Tejo portrays characteristics in New SQL database bunches in perspective of Service Level Objective (SLO) estimations. Their examinations with Volt DB, a discernible New SQL database, shed some light on the impact of irregularities on these databases and highlight the key outline decisions to upgrade oddity recognition [8]. They examined a versatile, cross breed Internet foundation for powerfully recognizing potential abnormalities progressively utilizing stream preparing. The framework

empowers scientifically investigating and looking at irregularities all-inclusive utilizing extensive scale cluster preparing. Conveyed on a genuine pipe organize topology of 1,891 hubs, this approach can successfully distinguish and describe peculiarities while limiting the measure of information shared over the system [9]. This paper, for the first time, exhibits the difficulties and openings in irregularity recognition for IOT and cloud. It first presents the unmistakable elements and application fields of IOT and Cloud, at that point talks about security and protection dangers to individual data and finally concentrates on arrangements from oddity recognition viewpoint [10]. They examined to exploit new circulated processing system keeping in mind the end goal to accelerate an Unsupervised Network Anomaly Detector Algorithm, UNADA. The assessment demonstrates that the execution time can be enhanced by a component of 13 enabling UNADA to process extensive hints of movement progressively [11].

Machine learning offers various algorithm for classification, clustering and combination of clustering and classification. The clustering and classificationtechniques provides various algorithm such as k-means, k-mod, and FCM and support vector machine, many more algorithm. Instead of clustering and simple classification the hybrid algorithm (plant growth optimization with support vector machine) gives more accuracy in terms of detection.

Rest of paper organized as follows: In section 2 describe the related work or literature survey and problem statement discussed in the section 3, section 4 describe the objective solution and last but not the least conclusion discuss in the section 5.

II. RELATED WORK

In this section of paper discusses the different data mining mechanism based intrusion detection system and made a comparison in term of accuracy, precision, and recall and probability of false alarm. In this section also defining the problem which have identified during implementation of this literature survey or research paper or different author or researcher.

ParisaAlaei and FakhroddinNoorbehbahani, et al. [12]:in this paper, a method is proposed to overcome this problem by performing online classification on datasets. In doing so, an incremental naive Bayesian classifier is employed. Furthermore, active learning enables solving the problem using a small set of labeled data points which are often very expensive to acquire. The proposed method includes two groups of actions i.e. offline and online. The former involves data preprocessing while the latter introduces the NADAL online method. The proposed method is compared to the incremental naive Bayesian classifier using the NSL-KDD standard dataset.

Amreen Sultana and M.A.Jabbar et al. [13]:proposed that with the tremendous growth of usage of internet and development in web applications running on various platforms are becoming the major targets of attack. New threats are create everyday by individuals and organizations that attack network systems. Intrusion is a malicious, externally induced operational fault. Intrusion is used as a key to compromise the integrity, availability and confidentiality of a computer resource. Hence intrusion detection systems (IDS) are becoming a key part of system defense, to detect anomalies and attacks in the network. Data mining based IDS can effectively identify intrusions. Average one dependence estimators (AODE) is one of the recent enhancements of naive Bayes algorithm. AODE solves the problem of independence by averaging all models generated by traditional one dependence estimator and is well suited for incremental learning. In this paper, we propose intelligent network intrusion detection system using AODE algorithm for the detection of different types of attacks. In order to evaluate the performance of our proposed system, we conducted experiments on NSL-KDD data set. Empirical results show that proposed model based on AODE is efficient with low FAR and high DR.

NazmulShahadat and Imam Hossain et al. [14]: As tremendous growth of information in theinternet, the importance of Network security also dramatically increases. Network and Host based Intrusion Detection System (IDS) are two primary systems in Network Security infrastructure. When new intrusion types are appeared in Network or Host, some serious problems are also appeared to detect these new intrusions. Due to this reason, IDSs demanded better than Signature based detection. The action of intrusion is represented by some features and collects the corresponding featured data from these uncertain feature characteristics. In last two decades, several techniques are developed to detect intrusion by using these data as human labeling which is very time consuming and expensive process. In this paper, we proposed a data mining rule based algorithm called Decision Table (DT) to detect intrusion and a new feature selection process to remove irrelevant/correlated features simultaneously. An empirical analysis on KDD'99 cup dataset was performed by using our proposed and some other existence feature selection techniques with DT and some others classification algorithms. The experimental results showed that proposed approach provides better performance in accuracy and cost compared among Bayesian Network, Naïve Bayes Classifier and other developed algorithms with data mining KDD'99 cup challenge in all cases.

SumaiyaThaseen and Aswani Kumar et al. [15]: Most of the intrusion detection systems analyze allnetwork traffic features to identify intrusions with different classification techniques. Any intrusion detection model developed has to provide maximum accuracy with minimal false alarms. Identifying the optimal feature subset for classification is an important task for improved classification. In this paper, consistency based feature selection is used to identify the optimal feature subset. The usage of training data improves the ability to differentiate the classes with similar behavior hence supervised classifiers are deployed to produce more reliable and accurate results. In this paper, we build a hybrid model for intrusion detection integrating consistency based feature selection and ensemble of classifiers such as SVM and LP-Boost. The objective of this paper is to determine the relevant features for the construction of model and classify the different network traffic classes using ensemble layer of classifiers. Our experimental analysis using NSL-KDD data set indicate that the proposed ensemble yields high accuracy though there is a huge class imbalance problem in network traffic.

Baojiang Cui and Shanshan He Et al. [16]: This Research exhibits a model which consolidate distributed computing with machine learning. Hadoop is a broadly utilized open source distributed computing structure to enormous information. The movement information put away in HDFS and handled by MapReduce. Other than these, machine learning module chose best execution calculation from various calculations by called Weka interface.

Table No. 1: Comparative Analysis of Different Data Mining Mechanism in Terms of Accuracy

Sr. No.	Approach	Accuracy			
		DOS	PROBE	R2L	U2R
01	NADAL	93.32	93.44	94.51	95.61
02	AODE	97.19	96.48	96.25	96.25
03	Naïve Bayes	89.90	90.48	95.47	95.47
04	SVM-LP	93.29	92.19	94.61	95.78
05	Incremental Naïve Bayes	89.96	90.64	89.35	90.41

III. PROBLEM INEXISTING SYSTEM

Most of the existing intrusion detection systems suffer from the following problems:

1. First, the information used by the intrusion detection system is obtained from audit trails or from packets on a network. Data needs to traverse an extended path from its origin to the IDS and within the method will probably be destroyed or changed by an assailant. Furthermore, the intrusion detection system needs to infer the behaviour of the system from the information collected, which may end in misinterpretations or lost events. This is often referred because the fidelity downside.
2. Second, the intrusion detection system endlessly uses further resources in system its observation even once there aren't any intrusions occurring, as a result of the parts of the intrusion detection system need to be running all the time. This is often resource usage downside.
3. Third, as results of the parts of the intrusion detection system are enforced as separate programs, they're vulnerable to tampering. Associate trespasser will probably disable or modify the programs running on a system, rendering the intrusion detection system useless or unreliable. This is often dependability downside.

IV. FUTURE WORK

In the future work we make hybrid technique of **SVM**(support vector machine) and **PGO** (plant growth optimization) by which we can improve the accuracy and detection rate.

Support Vector Machines: Support Vector Machines (SVMs) are extremely effective greatest edge straight classifiers

Plant Growth Optimization: The PGO take the solution space of the problem as the growth area of the artificial plant kingdom.

V. CONCLUSION

This paper presents the review of anomaly detection based on machine learning technique based on data mining. The detection of anomaly with accuracy is very critical task. The structure of data is very complex and arise huge traffic for the invitation of anomaly attack. For the detection of anomaly used various framework and model, but machine learning data mining algorithm is dominated algorithm. The machine learning data mining provides various classification and clustering algorithm. Some authors used hybrid technique for anomaly detection. Machine learning algorithm such as support vector machine achieve good classification accuracy in KDDCUP99 data. In future we can used ensemble based classification with optimization technique for the detection of anomaly.

REFERENCES

1) Sherenaz Al-Haj Baddar, Alessio Merlo and Mauro Migliardi “Anomaly Detection in Computer Networks: A State-of-the-Art Review”, Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 2015, Pp 29-64.
2) Zhengbing Hu, SergiyGnatyuk, Oksana Koval, Viktor Gnatyuk and SerhiiBondarovets “Anomaly Detection System in Secure Cloud Computing Environment”, I. J. Computer Network and Information Security, 2017, Pp 10-21.
3) Nikolay Laptev, SaeedAmizadeh and Ian Flint “Generic and Scalable Framework for Automated Time-series Anomaly Detection”, ACM, 2015, Pp 1-9.
4) Hasan Mahmoud Kanaker, MadihahMohd Saudi and MohdFadzliMarhusin “A Systematic Analysis on Worm Detection in Cloud Based Systems”, ARPN Journal of Engineering and Applied Sciences, 2015, Pp 540-548.
5) S. NagendraPrabhu and D. Shanthi “A Survey on Anomaly Detection of Botnet in Network”, International Journal of Advance Research in Computer Science and Management Studies, 2014, Pp 552-558.

6) Yu Liang and Chao Wu "A Sensor-Oriented Information System Based on Hadoop Cluster", Electronics and Energetics, 2016, Pp 437-450.
7) SelviKadirvel, Jeffrey Ho and Jose A. B. Fortes "Fault Management in Map-Reduce through Early Detection of Anomalous Nodes", ICAC, 2013, Pp 235-245.
8) Guthemberg Silvestre, Carla Sauvanaud, Mohamed Kaaniche and KaramaKanoun "Tejo: A Supervised Anomaly Detection Scheme for NewSQL Databases", SERENE, 2015, Pp 1-13.
9) DjellelEddineDifallah, Philippe Cudré-Mauroux and Sean A. McKenna "Scalable Anomaly Detection for Smart City Infrastructure Networks", IEEE, 2013, Pp 1-9.
10) Ismail Butun, BurakKantarci and MelikeErol-Kantarci "Anomaly detection and privacy preservation in Cloud-Centric Internet of Things", IEEE, 2015, Pp 2610-2615.
11) Juliette Dromard, Gilles Roudière and Philippe Owezarski "Unsupervised Network Anomaly Detection in Real-time on Big Data", Springer, 2015, Pp 197-206.
12) ParisaAlaei and FakhroddinNoorbehbahani, "Incremental Anomaly-based Intrusion Detection System Using Limited Labeled Data", 3th International Conference on Web Research (ICWR) IEEE 2017, Pp 178-184.
13) Amreen Sultana and M.A.Jabbar, "Intelligent Network Intrusion Detection System using Data Mining Techniques", IEEE 2016, Pp 329-333.
14) NazmulShahadat and Imam Hossain et al., "Experimental Analysis of Data Mining Application for Intrusion Detection with Feature reduction", International Conference on Electrical, Computer and Communication Engineering (ECCE), Bangladesh, Pp- 209-217, IEEE 2017.
15) SumaiyaThaseen and Aswani Kumar et al., "An integrated Intrusion Detection Model using consistency based feature selection and LP-Boost", Online International Conference on Green Engineering and Technologies (IC-GET), Pp-1-6 IEEE 2016.
16) Baojiang Cui and Shanshan He "Anomaly detection model based on Hadoop platform and Weka interface", Innovative Mobile and Internet Services in Ubiquitous Computing, 2016, Pp 84-89.