ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 10, Issue 3, pp: (26-33), Month: July - September 2022, Available at: www.researchpublish.com

Cyber Attack detection Using Big data analysis

Yazeed Al Moaiad[1], Yasser Mohamed Abdelrahman Tarshany[2],
Nasir Ahmed Algeelani[3], Wafa Al-Haithami[4]

Faculty of computer science and Information technology,
Faculty of Islamic Sciences,
AL-Madinah International University, Kuala Lumpur, Malaysia

DOI: https://doi.org/10.5281/zenodo.6924399
Published Date: 28-July-2022

Abstract: **Network-based Intrusion Detection System is a threat caused by the explosion of computer networks and the myriad of recent content-based threats, which occur daily. As well as an overview of machine learning approaches for signature and anomaly detection methods, this article discusses several machine learning strategies applied to intrusion detection and preprocessing. The NIDS taxonomy and attribute classifier have created classifications and outlines. Machine learning methods are widely utilized in anomaly detection using many data sets. Additional preprocessing methods have been added, for example, sorting and discretization have been applied to the data collection of measured values. Custom methods focused on search algorithms using machine learning that uses novel search algorithms are vulnerable to being revealed. This analysis is highly relevant to the use of machine learning methods used in computer security, which furthers their cause.**

Keywords: **Intrusion Detection, K-means Algorithm, Machine learning, Swarm Intelligence.**

I. INTRODUCTION

If the number of network providers grows, their stability, integrity, and availability often become an issue, which is an increasingly important concern. The 2014 Cisco Security Report (CISCO, 2014), where the increase in weaknesses is noted, that taking advantage of the latest assault approaches and revamped tactics has resulted in a corresponding rise in security threats, backs up this information. Lastly, the report suggests that organizations are no longer able to protect their networks. In comparison, 100% of the world's networks were found to have harmful material on their web servers, and 96% of the traced networks were found to be “phishing”.

Distributed denial of service (DDoS) attacks that target websites or the Internet have been even more common in 2013 and have risen in both frequencies and sophistication. Disorganized assaults have given way to more focused activities by cybercriminals, so complex that they might undermine the national security of both private and public institutions, as well as the country's prestige. Additionally, there is an increase in the vulnerability and response footprint as a result of the rapid increase in connected devices and virtualized Cloud computing environments. Vulnerabilities and security technologies have been supplemented by various kinds of mobile devices and infrastructures, making attacks by previously unseen adversaries possible.

As cybercriminals have found, the Internet's technology offers much more advantages than only looking for individual targets. These infrastructure attacks attempt to access the main web, DNS, and data center servers to potentially propagate risks to numerous smaller properties that are dependent on them. Warming up diminishes faith in the services which support the Internet [5].

It's perpetrated by trespassers. There are two groups of attackers: those that target unauthenticated computer networks, and others that have authenticated access to authenticated ones. For this purpose, thus, a shield is needed to defend the device from intruders [14]. Any effort to attack the resources, such as secrecy or availability, and honesty that undermines or

weakens security to find device vulnerabilities, we are employing IDS to check the operation of computer systems for unusual trends or behavior. Anomalies and misuse identification [14, 16]. Recent research exposes a growth of unknown content-based attacks, which is why new data preprocessing techniques are necessary as well as improving anomaly detection rates [28]. In this work, an anomaly detection study is carried out using machine learning techniques, and their hybridizations with other Artificial Intelligence techniques, such as swarm intelligence, identifying the key steps such as data pre-processing for reduction of dimensionality and detection techniques. The objective is to carry out a study of the state of the art that allows discovering open questions in the pre-processing and data processing stages with a machine learning approach.

II. LITERATURE REVIEW

A. Intrusion Detection From Misuse

When trends of real activity diverge from established patterns of misuse, identification is rule-based [24]. In the general, this method works to uncover proven attacks but is of little benefit when presented with unknown attacks [26, 30]. Any flaw in the signature concept makes the test less reliable. To be profiled and to cyber-abuse, it has four basic elements: data collection, device profiling, identification, mitigation, and reaction. one or more data points were extracted from the network, machine call source code, and/traces, etc. To meet the requirements of the rest of the framework, this data is structured to the extent that it is digestible [20].

B. Anomaly-Based Detection

Anomaly identification is based on the usual operation patterns of the device, not suspicious activity. It is derived from the idea that all things that intrude on someone's life, regardless of intent, are atypical or peculiar since anomalies are first detected, the regular properties of the observed artifacts are examined to see which are beyond the regular activity range [1, 12, and 17]. An anomaly detection model has four basic building blocks: data gathering, anomaly identification, reaction, and normalization. Information fetched and saved by the data collection component is considered standard consumer and traffic information. Special simulation is used to approximate the performance of the system's usual operations. Present practices are checked to see whether they adhere to framework norms and what proportion is declared as non-normal. In addition, the response portion issues an alert when an intruder is detected. Additionally, it is one of the best methods for finding new security flaws. Due to anomaly detection systems, though, it is expected that they have a strong false alarm rate.

On the other side, a technique-based approach tries to draw from existing data to emulate real-world actions (protocol specifications, network traffic instances, etc.). A final phase of machine learning is either an implicit or explicit model to categorize the trends, but machine learning strategies are dependent on one or the other [16].

C. NIDS Schemes Based On Machine Learning Machine Learning

Techniques are based on an established explicit or implicit model that makes it possible to categorize the analyzed patterns. A unique feature of these schemes is the need for labeled data to train the behavior model, this being a resource-demanding procedure. Many machine learning-based schemes have been applied to NIDS. Some of the most important are Bayesian Networks, Markov Models, Neural Networks, Fuzzy Logic Techniques, Genetic Algorithms, Clustering, and Outlier Detection [20, 28].

In addition to these techniques, others that help in the task of dealing with the large volumes of information contained in data sets, are known as dimensionality reduction techniques [20]. Two of these techniques are Principal Component Analysis (PCA) [19] which is based on the reduction of dimensionality from transformations applied to the data and the selection of attributes [2, 13, 19], which is the application of machine learning and search techniques, to select a subset of attributes to reduce the volume of data and increase the performance of the applied algorithms, gaining speed and from which better results are obtained classification. NIDS based on machine learning techniques has the following taxonomy.

D. Taxonomy Of Anomaly Detection Systems

To produce models based on the training data [28]. This data set consists of the attribute (characteristic) and classification data points. There are nominal and constant qualities. Attributes provide a direct influence on the effectiveness of anomaly detection strategies. As an illustration, continuous attributes usually require continuous methods to be accurate, and distant methods fail on the first cut. Many Data marks in general have one of two values: non-regular or attackable (attack).

Instead, several researchers have used various techniques, such as DDoS, U2R, probe, and ransomware they provide for further facts to be discovered with regard to forms of abnormalities. Experimental studies demonstrate that conventional learning strategies fail to reliably identify unusual situations. Sometimes, labeling performed by human experts needs high effort (G, D, M, & V, 2009)." The value of understanding traffic attributes is at the heart of machine learning techniques.

E. Attributes Of Network Traffic

One of the most important phases in the design of intrusion detection systems is the identification of the set of attributes to be used. The selection directly influences the performance of the system and the types of attacks that it will detect. But there is general confusion around what are the best network attributes, due to many causes; one of them is the lack of a universally accepted classification scheme.

Data Sets. K-means algorithm.

DONG [14] proposed the ISCX 2012 Intrusion Detection Evaluation Dataset, comparing it with the other existing data sets, taking into account a series of characteristics. It is made up of 19 attributes, including content attributes. So far there no article reports the use of this dataset in anomaly detection tasks. Many articles make use of the K-means algorithm as tagged data to test and compare intrusion detection algorithms [18]. In [29] characterize its disadvantages. But it is a publicly available, tagged, and pre-processed dataset for machine learning algorithms. This opens the field for researchers who want to test their algorithms and make valuable comparisons with other intrusion detection algorithms. Generating precise labels for data sets is a very time-consuming process, for this reason, this data set is used despite its years of creation. The data set was generated from DARPA 98. Each network connection was processed into labeled vectors of 41 attributes. These were built using data mining techniques and expert systems. The preprocessed data produced:

- 9 basic and derived header attributes of a simple connection, for each connection.
- 9 time-based header attributes, derived from multiple connections, built on a 2-second sliding window.
- 10 host-based header attributes, derived from multiple connections, built on a sliding window of 100 connections to detect scan attacks.
- 13 content-based attributes built from the content (payload) of the packages. They were designed to specifically detect U2R and R2L attacks.

As can be seen in Table 1, the K-means algorithm dataset contains around 5 million instances, each representing a TCP / IP connection that is made up of 41 quantitative and qualitative attributes. In many investigations a small portion is used that represents 10% of the original data set, it contains 494021 instances. This subset is used for training, while another subset containing 331029 instances is used for testing. Approximately 20% of both subsets represent normal traffic patterns (not attacks). The entire data set contains 39 attack types grouped into 4 categories.

The classification problem in the K-means algorithm set can be treated under two approaches:

- Binary: consists of distinguishing between attack and no attack.
- Classes multiple classes: consists of distinguishing the types of attacks.

The K-means algorithm dataset is widely used under a data stream approach to evaluate various classification algorithms.

Dataset	DoS	Probe	U2R	R2L	Normal
10 % KDD	391458	4107	52	1126	97277
Corrected KDD	229853	4166	70	16347	60593
Whole KDD	3883370	41102	52	1126	972780

Tabla 1 Características básicas de KDD Cup 99. (Bolón-Canedo *et al.*, 2011)

The percentage of attacks in both data sets is high. Most of the attacks belong to the DoS category. Despite this, the data set is very unbalanced with respect to certain categories such as U2R and R2L, of which very few examples are contained. Other research has manipulated these shortcomings and created a K-means algorithm-based dataset called NSL-KDD, from which they removed duplicate instances (78% and 75% duplicate instances in the training and test sets respectively). This

causes the data set to lose its real meaning, because instances are repeated in real environments, and the implemented methods must take this into account.

F. Attribute selection in the K-means algorithm

Attribute selection consists of determining the relevant attributes and discarding the irrelevant ones, with the aim of obtaining a subset of attributes that correctly describes the problem or process in question without affecting the performance of the algorithms. Selection has advantages such as [23]:

- Improves the performance of machine learning algorithms.
- Dimensionality reduction.
- It makes it possible to use simple models, thus gaining speed.

There are two approaches to selecting attributes [30]:

- Filter-based methods.
- Wrapper-based methods.

While wrapper methods optimize a certain algorithm as part of the selection process, filter methods rely on the general characteristic of training to select attributes regardless of the classification algorithm. However, wrapper models are time-consuming, restricting their use on large data sets. On the other hand, filter methods are less computationally expensive and have the possibility of being applied to large data sets. They can also be more generalized because they act independently of the induction algorithm.

Among the filter methods applied to the K-means algorithm are [3]:

- Correlation-based Feature Selection (CFS).
- INTERACT.
- Consistency-based.

CFS is one of the best-known and most widely used filters. INTERACT is a new approach based on the interaction between attributes and Consistency-based is one of the classic algorithms. In [10] they evaluate three methods for selecting attributes:

- Correlation-based Feature Selection.
- Information Gain and Gain Ratio.
- Feature Vitality Based Reduction Method.

G. Discretization algorithms applied to the K-means algorithm

Many filter algorithms work on discrete data. For this reason, a common practice for these algorithms is to discretize the data before making the selection. Discrete data is easier to understand, use, explain, and discretization can make learning more accurate and faster [25]. A set of algorithms only work with discrete data. Various methods to discretize appear in the literature [27], for example, Weka [65] discretizes the data using Entropy Minimization Discretization [15]. Considering that the K-means algorithm is considered a high-dimensional data set, adequate and classic discretization algorithms have been applied such as [3]:

- EMD (Entropy Minimization Discretization).
- EWD (Equal Width Discretization).
- EFD (Equal Frequency Discretization).
- PKID (It is a new approach very suitable for large data sets).

In [11] they carry out a review of the NIDS that use attributes from the K-means algorithm data set, taking into account the preprocessing techniques, algorithms used, and detected attacks. Among them are:[8, 25, 30, 28, 12,14, 17, 6].

Swarm intelligence techniques are applied to network intrusion detection tasks.

To improve NIDS performance and intrusion detection tasks, investigations have been conducted using swarm intelligence techniques to both optimize detection and improve system responses. Several articles [14, 16] have shown that hybridizing machine learning algorithms with swarm intelligence algorithms improves anomaly detection over other approaches and have proposed:

- ACO (Ant Colony Optimization) oriented to IDS (ACO to detect the origin of the attack, ACO for the induction of classification rules).
- PSO (Particle Swarm Optimization) oriented to IDS (PSO & Neural Networks, PSO & SVM (Support Vector Machine), PSO & K-Means, PSO for induction of classification rules).
- ACC (Ant Colony Clustering) oriented to IDS (ACC & SOM (Self-organizing map), ACC & SVM).

Most of the approaches that use ACO do so as a response mechanism, for example, to determine where the intrusion is from. Its use in the detection stage is less common. On the other hand, PSO is not used as a pure classification mechanism, the tendency is to hybridize it with classification algorithms and it has been shown to improve the performance of all machine learning techniques with which it has been tested. ACC has given the best classification results for most attack classes, especially for R2L-type attacks. They suggest that the study of hybridization between ACC with classification algorithms would be interesting [30].

III. METHODOLOGY

The research was carried out from the review of numerous articles related to intrusion detection under a machine learning approach, thus determining possible open questions in that area, where it could be deepened and made contributions. The research focuses on a review of the data pre-processing stages, the data set most used in this area, and the semi-supervised data processing. In the selection of the methods, aspects such as:

- The data that needed to be obtained.
- Correspondence with the theoretical design.
- Selected investigative strategy.

Advances in the research process were achieved by using scientific work methods such as:

A. General Methods

The hypothetical-deductive method to propose lines of work based on partial results; the historical-logical and dialectical method for the critical study of previous works and to use these as a point of reference and comparison of the results achieved.

B. Logical Methods

The analytical-synthetic method, by breaking down the research into separate elements and deepening the study of each one of them, to then synthesize them into the solution of the proposal; the induction-deduction method, as a way of theoretical verification during the development of the research.

C. Empirical Methods

The colloquial method for the presentation and discussion of the results; the experimental method to check the usefulness of the results obtained and the comparison with other reported methods.

IV. DISCUSSION

The application of general methods such as the hypothetical-deductive method allowed us to establish research strategies and define lines of work on the detection of intrusions in computer networks. This, together with the dialectical and historical-logical methods, allowed the study of previous works, in search of trends, using them as reference points for the determination of possible areas of research contribution in the application of machine learning techniques for the detection of intruders in the networks. In turn, logical methods such as analysis and synthesis made it easier to decompose intrusion detection into signature-based and anomaly-based detection, delving into anomaly-based detection and dividing it into machine learning-based approaches, in systems taxonomy. , as well as in the pre-processing and processing techniques used on the data; synthesizing the results obtained.

Specifically, in the preprocessing stage, empirical methods such as the experimental one were applied to verify the results of the algorithms for the selection and discretization of attributes proposed in other investigations. The colloquial method was applied for the presentation of the results in the pre-processing stages of the data for the detection of anomalies. Using this method, the results of various selection and discretization algorithms applied to the K-means algorithm data set were presented. Then, in the processing stage, which constitutes the detection of anomalies itself, better results are achieved in terms of precision. and false positive rates from hybridizing machine learning algorithms with optimization algorithms based on swarm intelligence.

When doing an unsupervised learning experiment, the researcher looks for parallels and comparisons. the "black box" solution involves a high percentage of input and output that cannot be changed by the customer This segment surveys the most common market problems that can be solved with algorithms.

Creative statement criticism clustering is used to partition points or findings into k groups where each group is completely distinguishable from the others Cluster proximity is measured by the nearest point to mean to the cluster. Many of these proximity measurements are used in the procedure, and hence it uses different algorithms to determine how near to the cluster's core data the results are.

Aside from the obvious being unable to operate until a cluster count is provided, the only downside of k-means is that the algorithm has to be launched before use. To put it another way, there is no way to calculate the optimum number of clusters for your needs. Typically, first, K values are implemented, and the most efficient one is chosen based on rule-based experience.

Despite this, k-me is an obvious disadvantage, it is one of the most efficient algorithms in market classification as well as it is in individual and consumer segmentation, image segmentation, image processing, text processing, etc.

A. Low-Density Scanning Categorical Clustering

The method used in K-means clustering is known as centroid-based, whereas those which make use of density as a differentiating factor use Q-an clustering. The following is an example of a traditional application of this method. Reachability and DBSCAN were mostly focused on two concepts: density-based clustering and density connectivity. The algorithm is allowed to differentiate and isolate areas of differing density thereby giving it the ability to classify and divide up clusters that DBSCAN runs, the parameter 'epsilon' is defined by the radius and the minimum number of points inside the search radius (m).

First, the point is discarded, which is still unaccessed data is labelled with "infinity. If it contains all the m minimal points, the cluster is formed, and therefore it is marked as "visited" in this iteration; if not, noise can be created, which can be corrected in the next iteration.

If the next datum is close to this location, then the cluster adjacent to the location ϵ was formed. Continue eliminating non-corresponding nodes until no longer density or density-corresponding nodes remain. Once out of this loop, it continues to the next data point, which causes further noise to be created. The algorithm approaches a stable state as no more of the unvisited data points remain to be visited.

B. High-Lowest Possible Leverage, Principal Component Orthogonalization

The PCA algorithm can be used to reduce the number of variables in a big dataset, by trading each of them for a less number of summary variables, but also preserving the overall dataset structure. There is almost no knowledge loss, but there is still a chance of inaccuracy. Also, PCA is mostly used in unsupervised instruction, but it has rational and sequential implementations as well. Variables are divided into sets classified as major components (PC). Orthogonal combinations are a linear combination of orthogonal variables that have weights, which are found by an eigenvector method. Where PCA is used, multi-dimensional data, such as business and bioinformatics, is concerned, it finds a lot of uses.

V. CONCLUSION

With the study carried out, it was possible to determine the design of the detection methods and the selection of the attributes of the system. Two main open questions to detect anomalies are the networks to be monitored. In addition, needs that give rise to work areas, for example, the definition of new content-based network attributes to build learning models that allow detecting new content-based attacks, are identified. The performance of new experiments that hybridize swarm intelligence algorithms with machine learning algorithms to improve detection rates in content-based attacks such as the U2R and R2L variants are issues that we are currently working on, mainly in experiments that hybridize.

REFERENCES

[1] AGRAWAL, H., C. BEHRENS AND B. DASARATHY. Learning program behavior for anomaly detection. In.: Google Patents, 2013.

[2] AHMED, P. A Hybrid-Based Feature Selection Approach for IDS. In Networks and Communications (NetCom2013). Springer, 2014, p. 195-211.

[3] BOLÓN-CANEDO, V., N. SÁNCHEZ-MAROÑO AND A. ALONSO-BETANZOS Feature selection and classification in multiple class datasets: An application to K-means algo dataset. Expert Systems with Applications, 2011, 38 (5), 5947-5957.

[4] BOLZONI, D., E. ZAMBON, S. ETALLE AND P. HARTEL Poseidon: A 2-tier anomaly-based intrusion detection system. arXiv preprint cs / 0511043, 2005.

[5] CISCO. Get the Latest Findings on Malware Threats. In., 2014.

[6] COVA, M., C. KRUEGEL AND G. VIGNA. Detection and analysis of drive-by-download attacks and malicious JavaScript code. In Proceedings of the 19th international conference on World Wide Web. ACM, 2010, p. 281-290.

[7] CHANDOLA, V., A. BANERJEE AND V. KUMAR Anomaly detection: A survey. ACM Computing Surveys (CSUR), 2009, 41 (3), 15.

[8] CHEBROLU, S., A. ABRAHAM AND JP THOMAS Feature deduction and ensemble design of intrusion detection systems. Computers & Security, 2005, 24 (4), 295-307.

[9] CHEN, C.-M., W.-Y. TSAI AND H.-C. LIN. Anomaly behavior analysis for web page inspection. In Networks and Communications, 2009. NETCOM'09. First International Conference on. IEEE, 2009, p. 358-363.

[10] CHOWDHARY, M., S. SURI AND M. BHUTANI Comparative Study of Intrusion Detection System 2014.

[11] DAVIS, JJ AND AJ CLARK Data preprocessing for anomaly based network intrusion detection: A review. Computers & Security, 2011, 30 (6), 353-375.

[12] DEORIO, A., Q. LI, M. BURGESS AND V. BERTACCO. Machine learning-based anomaly detection for post-silicon bug diagnosis. In Proceedings of the Conference on Design, Automation and Test in Europe. EDA Consortium, 2013, p. 491-496.

[13] DOKAS, P., L. ERTOZ, V. KUMAR, A. LAZAREVIC, et al., Data mining for network intrusion detection. In Proc. NSF Workshop on Next Generation Data Mining. 2002, p. 21-30.

[14] DONG, G., J. GAO, R. DU, L. TIAN, et al., Robustness of network of networks under targeted attack. Physical Review E, 2013, 87 (5), 052804.

[15] DOUGHERTY, J., R. KOHAVI AND M. SAHAMI. Supervised and unsupervised discretization of continuous features. In ICML. 1995, p. 194-202.

[16] ERTOZ, L., E. EILERTSON, A. LAZAREVIC, P.-N. TAN, et al., Detection of novel network attacks using data mining. In Proc. of Workshop on Data Mining for Computer Security. Citeseer, 2003.

[17] ESKIN, E., AO ARNOLD, M. PRERAU, L. PORTNOY, et al., Methods of unsupervised anomaly detection using a geometric framework. In.: Google Patents, 2013.

[18] ESTEVEZ-TAPIADOR, JM, P. GARCIA-TEODORO AND JE DIAZ-VERDEJO. Stochastic protocol modeling for anomaly based network intrusion detection. In Information Assurance, 2003. IWIAS 2003. Proceedings. First IEEE International Workshop on. IEEE, 2003, p. 3-12.

[19] FEINSTEIN, B., D. PECK AND I. SECUREWORKS Caffeine monkey: Automated collection, detection and analysis of malicious javascript. Black Hat USA, 2007, 2007.

[20] GARCIA-TEODORO, P., J. DIAZ-VERDEJO, G. MACIÁ-FERNÁNDEZ AND E. VÁZQUEZ Anomaly-based network intrusion detection: Techniques, systems and challenges. Computers & Security, 2009, 28 (1), 18-28.

[21] GOGOI, P., B. BORAH AND D. BHATTACHARYYA Anomaly detection analysis of intrusion data using supervised & unsupervised approach. Journal of Convergence Information Technology, 2010, 5 (1), 95-110.

[22] GUENNOUN, M., A. LBEKKOURI AND K. EL-KHATIB. Selecting the best set of features for efficient intrusion detection in 802.11 networks. In Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008. 3rd International Conference on. IEEE, 2008, p. 1-4.

[23] GUYON, I. AND A. ELISSEEFF An introduction to variable and feature selection. The Journal of Machine Learning Research, 2003, 3, 1157-1182.

[24] HEADY, R., G. LUGER, A. MACCABE AND M. SERVILLA The architecture of a network-level intrusion detection system. Edition ed.: Department of Computer Science, College of Engineering, University of New Mexico, 1990.

[25] HERNÁNDEZ-PEREIRA, E., and JA SUÁREZ-ROMERO, O. FONTENLA-ROMERO AND A. ALONSO-BETANZOS Conversion methods for symbolic features: A comparison applied to an intrusion detection problem. Expert Systems with Applications, 2009, 36 (7), 10612-10617.

[26] IDREES, F., M. RAJARAJAN AND A. MEMON. Framework for distributed and self-healing hybrid intrusion detection and prevention system. In ICT Convergence (ICTC), 2013 International Conference on. IEEE, 2013, p. 277-282.

[27] IHSAN, Z., MY IDRIS AND AH, ABDULLAH Attribute Normalization Techniques and Performance of Intrusion Classifiers: A Comparative Analysis. Life Science Journal, 2013, 10 (4).

[28] KAUR, H., G. SINGH AND J. MINHAS A Review of Machine Learning based Anomaly Detection Techniques. arXiv preprint arXiv: 1307.7286, 2013.

[29] KIANI, M., A. CLARK AND G. MOHAY. Evaluation of anomaly based character distribution models in the detection of SQL injection attacks. In Availability, Reliability and Security, 2008. ARES 08. Third International Conference on. IEEE, 2008, p. 47-55.

[30] KIM, G., S. LEE AND S. KIM A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. Expert Systems with Applications, 2014, 41 (4), 1690-1700.