International Journal of Advanced Technology and Engineering Exploration, Vol 3(18)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2016.318006

Review Article

An approach for efficient intrusion detection for KDD dataset: a survey

Namita Sharma[1*] and Bhupesh Gaur[2]
M.Tech Research Scholar, Computer Science, TIT, Bhopal, India[1]
Professor and Head, Computer Science, TIT, Bhopal, India[2]

Abstract
Identifying possible attacks on the network system is a challenging task. There is several research works are processed in this direction, but the need of improvement is always remains in the research. The accuracy of detection is the major challenge today. It becomes tougher as the intrusion types and their nature are different. So considering the above event detail discussion and analysis has been presented in this paper. Analysis of different techniques like data mining, machine learning and optimization is also presented so that the significant improvement may be judges and better future suggestion can be came out of this. DOS, U2R, R2L and probe attacks are considered and the Knowledge Discovery and Data mining (KDD) database have been considered for comparison of different techniques.

Keywords
Intrusion detection, Data mining, Machine learning, Optimization, KDD, DOS, U2R, R2L and probe.

1.Introduction
As of late, numerous specialists are centered to utilize information digging ideas for Intrusion Detection [1]. This is a procedure to extricate the understood data and learning. Intrusion detection is the procedure of pernicious on the framework and system when we are currently correspondence or removing information in the constant environment [2][3]. Since its creation, interruption location has been one of the key components in accomplishing data security. It goes about as the second-line guard, which supplements the detection controls. At the point when the controls fizzled, the interruption identification frameworks ought to have the capacity to identify it constant and caution the security officers to take incite and suitable activities [3][4].

Interruption detection framework manages regulating the occurrences happening in PC framework or system situations and analyzing them for indications of conceivable occasions, which are encroachment or inevitable dangers to PC security, or standard security rehearses Intrusion detection system (IDS) have risen to identify activities which jeopardize the uprightness, privacy or accessibility of are sourced as a push to give an answer for existing security issues [5].

So in the above bearings we review a few angles in the resulting segments. We likewise examine about information mining and advancement methods, in light of the fact that it can be utilized as a part of shaping the structure which delivers a better recognition framework.

As we examine this study toward a superior system with the blend of information mining and streamlining. These techniques are valuable and has been utilized as a part of diverse methodologies like [6][7][8][9][10][11]. So the utilization of these calculations can improve an effect. The researches have extended their views in this direction by several research papers as in [12][13][14][15].

2.Literature survey
In 2010, G. Schaffrath et al. [16] provide a survey of current research in the area of flow-based intrusion detection. The survey starts with a motivation why flow-based intrusion detection is needed. The concept of flows is explained, and relevant standards are identified. The paper provides a classification of attacks and defense techniques and shows how flow-based techniques can be used to detect scans, worms, Botnets and DoS attacks.

In 2011, Zhengjie Li et al. [17] propose a K-means clustering algorithm based on particle swarm optimization (PSO-KM). The proposed algorithm has overcome falling into local minima and has relatively

*Author for correspondence

good overall converged. Experiments on data sets KDD CUP 99 has shown the effectiveness of the proposed method and also shows the method has higher detection rate and lower false detection rate.

In 2012, LI Yin–huan [18] focuses on an improved FP-Growth algorithm. According to author Preprocessing of data mining can increase efficiency on searching the common prefix of node and reduce the time complexity of building FP-tree. Based on the improved FP Growth algorithm and other data mining techniques, an intrusion detection model is carried out by authors. Their experimental results are effective and feasible.

In 2012, P. Prasenna et al. [19] suggested that in conventional network security simply relies on mathematical algorithms and low counter measures to taken to prevent intrusion detection system, although most of this approaches in terms of theoretically challenged to implement. Authors suggest that instead of generating large number of rules the evolution optimization techniques like Genetic Network Programming (GNP) can be used .The GNP is based on directed graph. They focus on the security issues related to deploy a data mining-based IDS in a real time environment. They generalize the problem of GNP with association rule mining and propose a fuzzy weighted association rule mining with GNP framework suitable for both continuous and discrete attributes.

In 2011, LI Han [20] focuses on intrusion detection based on clustering analysis. The aim is to improve the detection rate and decrease the false alarm rate. A modified dynamic K-means algorithm called MDKM to detect anomaly activities is proposed and corresponding simulation experiments are presented. Firstly, the MDKM algorithm filters the noise and isolated points on the data set. Secondly by calculating the distances between all sample data points, they obtain the high-density parameters and cluster-partition parameters, using dynamic iterative process we get the k clustering center accurately, then an anomaly detection model is presented. They used KDD CUP 1999 data set to test the performance of the model. Their results show the system has a higher detection rate and a lower false alarm rate, it achieves expectant aim.

In 2011, Z. Muda et al. [21] discuss about the problem of current anomaly detection that it unable to detect all types of attacks correctly. To overcome this problem, they propose a hybrid learning approach through combination of K-Means clustering and Naïve Bayes classification. The proposed approach will be clustering all data into the corresponding group before applying a classifier for classification purpose. An experiment is carried out to evaluate the performance of the proposed approach using KDD Cup ’99 dataset. Result show that the proposed approach performed better in term of accuracy, detection rate with reasonable false alarm rate.

In 2014, Deshmukh et al. [22] presents a Data Mining method in which various preprocessing methods are involved such as Normalization, Discretization and Feature selection. With the help of these methods the data is preprocessed and required features are selected. They used NaIve Bayes method in a supervised learning method which classifies various network events for the KDD cup'99 Dataset.

In 2014, Benaicha et al. [23] present a Genetic Algorithm (GA) approach with an improved initial population and selection operator, to efficiently detect various types of network intrusions. They used GA to optimize the search of attack scenarios in audit files, thanks to its good balance exploration / exploitation; according to the authors it provides the subset of potential attacks which are present in the audit file in a reasonable processing time. The testing phase of the Network Security Laboratory Knowledge Discovery and Data Mining (NSL-KDD99) benchmark dataset has been used to detect the misuse activities. Their approach of IDS with Genetic algorithm increases the performance of the detection rate of the Network Intrusion Detection Model and reduces the false positive rate. In 2014 Kiss et al. [24] suggest that Modern Networked Critical Infrastructures (NCI), involving cyber and physical systems, are exposed to intelligent cyber-attacks targeting the stable operation of these systems. To ensure anomaly awareness, their observed data can be used in accordance with data mining techniques to develop Intrusion Detection Systems (IDS) or Anomaly Detection Systems (ADS). They proposed a clustering based approach for detecting cyber-attacks that cause anomalies in NCI. Various clustering techniques are explored to choose the most suitable for clustering the time-series data features, thus classifying the states and potential cyber-attacks to the physical system. The Hadoop implementation of MapReduce paradigm is used to provide a suitable processing environment for large datasets.

In 2014, Thaseen et al. [25] proposed a novel method of integrating principal component analysis (PCA) and support vector machine (SVM) by optimizing the kernel parameters using automatic parameter selection technique. Their approach reduces the training and testing time to identify intrusions thereby improving the accuracy. Their proposed method was tested on KDD data set. The datasets were carefully divided into training and testing considering the minority attacks such as U2R and R2L to be present in the testing set to identify the occurrence of unknown attack. Their results indicate that the proposed method is successful in identifying intrusions. Their experimental results show that the classification accuracy of the proposed method outperforms other classification techniques using SVM as the classifier and other dimensionality reduction or feature selection techniques.

In 2014, Wagh et al. [26] suggested Network security is a very important aspect of internet enabled systems in the present world scenario. According to the authors due to intricate chain of computers the opportunities for intrusions and attacks have increased. Therefore it is need of the hour to find the best ways possible to protect our systems. So the authors suggest intrusion detection system are playing vital role for computer security. The most effective method used to solve problem of IDS is machine learning. Thy observed that the rising field of semi supervised learning offers a assured way for complementary research. So they proposed a semi-supervised method to reduce false alarm rate and to improve detection rate for IDS.

In 2014, Masarat et al. [27] introduced a novel multistep framework based on machine learning techniques to create an efficient classifier. In first step, the feature selection method will implement based on gain ratio of features by the authors. Their method can improve the performance of classifiers which are created based on these features. In classifiers combination step, we will present a novel fuzzy ensemble method. So, classifiers with more performance and lower cost have more effect to create the final classifier

3.Problem domain

After discussing several research works we can come with some problem area in the traditional approaches which are following:

1) Need of Hybrid Intrusion Detection System, which is better at detecting R2L and U2R attacks [20].
2) The IDS approach can be enhanced by providing more security to mobile agents[18].
3) Step Propagation is missing.
4) Optimization based classification is missing.
5) Neuro-Fuzzy Combination can be used as the distributed classifier.
6) All types of attacks is not well detected.
7) Maintain long log files for detection.
8) Triangle Area Nearest Neighbor (TANN) and K-Means with K-Nearest Neighbor (KMKNN) approach for better intrusion detection. This approach showed a reasonable detection rate compare to our approach. Unfortunately, a potential drawback of this technique is the rate of false alarms [20][28].
9) In [29] Evolutionary Soft Computing based Intrusion Detection System (ESC-IDS) which focuses to detect and classify intrusion has proposed. This approach has serious shortcomings in its low accuracy rate as well as the tendency to produce high false alarm compare to [20].
10) A probability of less detection in U2R and R2L Detection technique so there is the need of a detection technique which improves in the hybridization of above two.

4.Analysis

After studying and observing several research works we compare the resulting discussions by their techniques, so that we identify the weak attack detection area. The results comparison of different methods is shown in *Table 1*

Table 1 Analysis

S.No	Approach	Accuracy (%)	Precision (%)	Recall(%)
1	NB Training[20]	DOS94.3	U2R80	R2L65.6
2	KM + NB Training[20]	DOS99.5	U2R40	R2L61.6
3	NB Testing[20]	DOS82.5	U2R80	R2L90.3
4	Km + NB Testing[20]	DOS99.6	U2R80	R2L83.2
5	Rule based[19]	92.14	87.24	84.43

S.No	Approach	Accuracy (%)	Precision (%)	Recall(%)
6	FSVM[21]	97.14	96.11	94.10
7	Rule based[21]	89.90	84.32	83.23
8	FSVM[21]	95.23	92.14	91.21
9	Rule based[21]	91.34	86.14	85.11
10	FSVM[21]	95.12	93.21	91.13
11	Rule based[21]	92.22	88.21	87.66
12	FSVM[21]	97.13	94.52	93.67
13	Fuzzy Ensemble[27]	93.00	NA	NA
14	Random Forest [30]	92.93 %	NA	NA
15	JRip [31]	92.30 %	NA	NA
16	SVM [32]	92.18 %	NA	NA

5.Conclusion and future work

This paper provides a direction in the face of Intrusion detection improvement. Our study suggests the following gaps which can be considered as the future directions:

1) A data Mining technique which dynamic rule association can be fruitful.
2) Combination K-Means with Optimization can increase the pattern recognition.
3) Particle warm optimization can help in better pattern detection.
4) The detection approach can be better at detecting R2L and U2R attacks more efficiently as well as anomaly detection approach, which is better at detecting attacks at the absence of match signatures as provided in the misuse rule files[20].
5) Hybridization of Association and Optimization can provide better detection.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] Jianliang M, Haikun S, Ling B. The application on intrusion detection based on k-means cluster algorithm. In international forum on information technology and applications 2009 (pp. 150-2.) IEEE.
[2] Lee HY, Wang NJ. The implementation and investigation of securing web applications upon multi-platform for a single sign-on functionality. International Journal of Advanced Computer Research. 2016; 6(23): 39-46.
[3] Tian L, Jianwen W. Research on network intrusion detection system based on improved k-means clustering algorithm. In international forum on computer science-technology and applications 2009 (pp. 76-9). IEEE.
[4] Devaraju S, Ramakrishnan S. Performance analysis of intrusion detection system using various neural network classifiers. In international conference on recent trends in information technology (ICRTIT) 2011 (pp. 1033-8). IEEE.
[5] Ishida M, Takakura H, Okabe Y. High-performance intrusion detection using optigrid clustering and grid-based labelling. In international symposium on applications and the internet (SAINT) 2011 (pp. 11-9). IEEE.
[6] Brugger ST. Data mining methods for network intrusion detection. University of California at Davis. 2004.
[7] Farhaoui Y. How to secure web servers by the intrusion prevention system (IPS)? International Journal of Advanced Computer Research. 2016; 6(23):65-71.
[8] Nalavade K, Meshram BB. Mining association rules to evade network intrusion in network audit data. International Journal of Advanced Computer Research. 2014; 4(15):560-7.
[9] Lee W, Stolfo SJ. Data mining approaches for intrusion detection. In Usenix security. 1998.
[10] Naoum R, Aziz S, Alabsi F. An enhancement of the replacement steady state genetic algorithm for intrusion detection. International Journal of Advanced Computer Research. 2014; 4(15):487-93.
[11] Lee W, Stolfo SJ, Mok KW. A data mining framework for building intrusion detection models. In proceedings of the IEEE Symposium on security and privacy 1999 (pp. 120-32). IEEE.
[12] Kumari S, Shrivastava M. A study paper on IDS attack classification using various data mining techniques. International Journal of Advanced Computer Research. 2012; 2(5):195-200.
[13] Venkatesan R, Ganesan R, Selvakumar AA. A comprehensive study in data mining frameworks for intrusion detection. International Journal of Advanced Computer Research.2012; 2(7):29-34.

[14] Kaushik M, Ojha G. Attack penetration system for SQL Injection. International Journal of Advanced Computer Research. 2014; 4(15):724-32.
[15] Patel R, Bakhshi D, Arjariya T. Random particle swarm optimization (RPSO) based intrusion detection system. International Journal of Advanced Technology and Engineering Exploration (IJATEE). 2015; 2(5): 60-6.
[16] Sperotto A, Schaffrath G, Sadre R, Morariu C, Pras A, Stiller B. An overview of IP flow-based intrusion detection. IEEE communications surveys & tutorials. 2010; 12(3):343-56.
[17] Li Z, Li Y, Xu L. Anomaly intrusion detection method based on k-means clustering algorithm with particle swarm optimization. In international conference on information technology, computer engineering and management sciences (ICM) 2011 (pp. 157-61). IEEE.
[18] Yin-huan LI. Design of intrusion detection model based on data mining technology. In international conference on industrial control and electronics engineering 2012.
[19] Prasenna P, RaghavRamana AV, Krishnakumar R, Devanbu A. Network programming and mining classifier for intrusion detection using probability classification. In international conference on pattern recognition, informatics and medical engineering (PRIME) 2012 (pp. 204-9). IEEE.
[20] Han LI. Using a dynamic k-means algorithm to detect anomaly activities. In seventh international conference on computational intelligence and security (CIS) 2011 (pp.1049-52). IEEE.
[21] Muda Z, Yassin W, Sulaiman MN, Udzir NI. Intrusion detection based on K-Means clustering and Naïve Bayes classification. In international conference on information technology in Asia (CITA 11) 2011 (pp. 1-6). IEEE.
[22] Deshmukh DH, Ghorpade T, Padiya P. Intrusion detection system by improved pre-processing methods and Naïve Bayes classifier using NSL-KDD 99 dataset. In international conference on electronics and communication systems (ICECS) 2014 (pp. 1-7). IEEE.
[23] Benaicha SE, Saoudi L, Guermeche B, Eddine S, Lounis O. Intrusion detection system using genetic algorithm. In science and information conference (SAI) 2014 (pp. 564-8). IEEE.
[24] Kiss I, Genge B, Haller P, Sebestyen G. Data clustering-based anomaly detection in industrial control systems. In international conference on intelligent computer communication and processing (ICCP) 2014 (pp. 275-81). IEEE.
[25] Thaseen IS, Kumar CA. Intrusion detection model using fusion of PCA and optimized SVM. In international conference on contemporary computing and informatics (IC3I) 2014 (pp. 879-84). IEEE.
[26] Wagh SK, Kolhe SR. Effective intrusion detection system using semi-supervised learning. In international conference on data mining and intelligent computing (ICDMIC) 2014 (pp. 1-5). IEEE.
[27] Masarat S, Taheri H, Sharifian S. A novel framework based on fuzzy ensemble of classifiers for intrusion detection systems. In international eConference on computer and knowledge engineering (ICCKE) 2014 (pp. 165-70). IEEE.
[28] Tsai CF, Lin CY. A triangle area based nearest neighbours approach to intrusion detection. Pattern recognition. 2010; 43(1):222-9.
[29] Xiang C, Yong PC, Meng LS. Design of multiple-level hybrid classifier for intrusion detection system using Bayesian clustering and decision trees. Pattern Recognition Letters. 2008; 29(7):918-24.
[30] Zhang J, Zulkernine M, Haque A. Random-forests-based network intrusion detection systems. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews. 2008; 38(5):649-59.
[31] Nguyen HA, Choi D. Application of data mining to network intrusion detection: classifier selection model. In challenges for next generation network operations and service management 2008 (pp. 399-408). Springer Berlin Heidelberg.
[32] Ambwani T. Multi class support vector machine implementation to intrusion detection. In proceedings of the international joint conference on neural networks 2003(pp. 2300-05). IEEE.