Intrusion Detection System Using Data Mining A Review

Varsha Singh
Department of Computer Engineering
Mukesh Patel School of Technology
Management and Engineering,
NMIMS
Mumbai, India
Email: Singh.Varsha1612@gmail.com

Shubha Puthran
Department of Computer Engineering
Mukesh Patel School of Technology
Management and Engineering,
NMIMS
Mumbai, India
Email: Shubha.Puthran@nmims.edu

***Abstract*—Everyday huge amount of information are transferred from one network to another, the information may be exposed to attacks. The information and information system should be protected from unauthorized users. To provide and maintain the Confidentiality and Integrity of the information is a very tedious job so Intrusion Detection plays a very important role. Although various methods are used to protect the information, loopholes exist. Data mining methods is used to analyze different attack patterns in the network. Various Classification, Clustering and Classification via Clustering (CvC) algorithms are reviewed. From the review it has been concluded that CvC would be the best suitable for Intrusion detection. In Intrusion Detection field the Cyber Security and Technology Group Contributed significantly by providing KDDcup 99 dataset and to motivate the researchers by eliminating security and privacy concern. NSL_KDD, GureKDD and Kyoto 2006+ dataset is discussed with their advantages and disadvantages.**

Keywords—IDS; U2R attack; R2L attack; Classification; Decision Tree; Clustering; Classification via Clustering.

I. INTRODUCTION

Now a day people are very much addicted to internet and the attacker are becoming more smarter by Knowing different device and tricks for intruding and attacking networks. This was the main reason to encourage Intrusion Detection to detect and prevent the system and data from attackers and intruders. Intrusion is a set of a malicious activity which tries to harm the system or data. A process of gathering intrusion related knowledge occurring during system monitoring and then analyzing collected data is known as Intrusion Detection. Also to draw a conclusion whether the system is intrusive or not according to user activity, system logs etc.[1]. Intrusion Detection is categorized into two types Anomaly based detection and Misuse Based detection [29]. Anomaly based analyze the normal behavior of the system if behavior differs it raise the alarm Misuse based detection raise alarm when analyze attack matches with stored attack in database.

Intrusion Detection System is divided into three types: Host Intrusion Detection System (HIDS), Network Intrusion Detection System (NIDS), Distributed Intrusion detection System (DIDS). HIDS analyze the incoming and outgoing packet of a single system. NIDS analyze the incoming and outgoing packets of all devices in the network. DIDS analyzes the attack of multiple hosts. IDS act as defender for a network, but everyday a new attack patterns appears in a network. To detect and prevent intrusions different existing techniques, architectures and their loopholes are discussed in this paper [19].

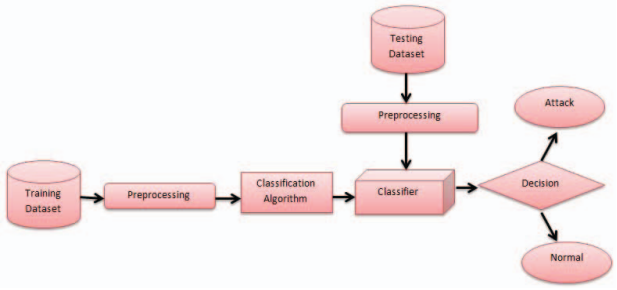

Fig.1. Classification

Intrusion Detection System using Data Mining plays a key role to analyze process and sort the data in systematic and organized manner without any mistake [2]. The learning process is gradual and induced and follows a data-centric approach. It is assumed that legitimate or illegitimate activity will have their footprints in the audit data [19]. Classification is one of the supervised learning method. It inductively learned to construct a model from the pre-classified data set [6].The working of Classification method is shown in Fig 1.

Basic Methodology of Classification Technique is divided into two stage Training and Testing stage.

A. Training Stage

By using different algorithms, the training dataset is used to construct a classifier.

978-1-5090-0467-6/16/$31.00 ©2016 IEEE.

B. Testing Stage

The classifier is used to take decision of testing tuple.

Decision tree is one of the Classification Technique. The main approach of decision tree is to select the attribute which gives the maximum information and divides the data item into classes. The detection rate of Dos, U2R, Probe and Normal attack are maximum using Decision tree Algorithm, as compared to Artificial Neural Network, Biological Neural Network and Support Vector Machine [5]. By using Decision tree, it is possible to quickly determine all the rules match an input element with maximal Comparison [6].

The rest of the paper is organized as follows. Section 2 provides review of Related work. Section 3 provides Comparative study of Hybrid methods in Section 4. Pros and cons of various dataset have been discussed in Section 5.Finally Conclusion is given in Section 6.

II. RELATED WORK

With increasing the requirement of Network security, Data mining technique can be used to find the patterns in a large datasets. It involves different methods and algorithms. Lots of research has been done using different Classifications, Clustering and Hybrid algorithm to improve the detection rate of an attack. Comparison of different classification algorithm is done to analyze the intrusion detection. This detection is done based on the speed, capability to handle large datasets and dependency on the parameters [11]. ID3 is a decision tree algorithm which uses tree like structure to take a decision, the main disadvantages of this approach is that it can select the attribute with many values because of this correct classification is not done. To overcome this problem improvement of ID3 algorithm is done which helps to classify the attack correctly[8]. C4.5 is another Decision tree algorithm which ignores the unique value attributes in the large datasets. The ID3 C4.5 algorithm is biased towards multi valued attribute, unbalanced split and size of decision tree increases, so Re Optimization technique is used to solve this problem [9]. Another method to improve the detection rate and increases the speed for execution of C4.5 is Pruning the nodes using KDDcup 99 and NSL_KDD datasets [2]. This disadvantages of pre-pruning technique is it stop before the due time and in post-pruning technique it waits to grow the tree completely and then delete the branches. To calculate the complexity of the nodes both the pruning method will take more system time. To overcome the Problem Multi-Strategy Pruning Algorithm is used to optimize the decision tree and get optimal result [7]. Support Vector machine (SVM) is another classification algorithm which can not deal with heterogeneous datasets so Radial biased function is used with SVM[15]. Only classification method can not improve the detection rate of an attack so to overcome this problem Cascading of an algorithm is used, the Cascading techniques can be of two types Cascading Supervised techniques and Cascading Supervised and unsupervised technique [16]. Cascading K-Means Clustering with C4.5 decision Tree is used for increase the detection rate of an attacks[10]. The another method of Supervised and Unsupervised technique is based on K-Means Clustering of Naïve Bayes Classification methods[18]. Combination of different Classification algorithm is also use to overcome the improvement of detection rate problem [13]. A multi layered approach is used to overcome the problem of single layer Intrusion Detection by using Genetic Algorithm (GA). In this approach each layer is correspond to each attacks (Dos.U2R.R2L.Probe). This algorithm efficiently detect R2L attack. Dos is not accurately detected by using Genetic Algorithm [20]. GA and SVM is proposed to improve the measure for detecting intrusions. Combination of two algorithms help to select the optimal feature subset. GA algorithm is useful for generation of attribute subset from the original dataset. Out of 45 attribute in KDDcup 99, 10 critical attribute are selected for highest accuracy and best result. The true positive value of this algorithm is 97.3% which is better than other selecting attribute [21].

III. CLASSIFICATION ALGORITHM

A. ID3

In early 1980, J.Ross Quinlan developed a Decision tree algorithm known as ID3(Iterative Dichotomister).It uses top down approach which start with training set tuples till class labels.ID3 is a type of Decision tree Algorithm which select the attribute as a splitting node with a highest Information gain. Information gain is one of the attribute selection measure used in ID3.The information is minimized for classifying the given tuple in resulting partition and impurity is reduced for generating the simplest tree. The main disadvantages ID3 is that it selects attribute with maximum unique value in a datasets because of this the splitting done by ID3 is not always correct, so some of the improvement have done to overcome this problem.[11][8].

B. C4.5

For generating a decision tree C4.5 is another Attribute selection measure used in C4.5 is a successor of ID3.The attribute selection measure used in C4.5 is Gain Ratio is selected as splitting attribute become huge because this the time complexity and space complexity increase .Pruning is used to control space and time problem.[2][7][4].

C. C5.0

C5.0 is a decision tree algorithm, by using boosting the performance of building tree is improved using. It combines different classifier .When training data contain lots of noise boosting method does not help [6].The detection rate of R2L,DoS,Probe and Normal are more as compared to ID3 and C4.5.Memory usage is more efficient as compared to C4.5.The decision tree is smaller than C4.5.Winnowing is used to automatically remove the attributes that may be useless. [6].

D. Random Tree

Each tree has an equal chance of being sampled, k-randomly select attribute at each node is used to construct a tree. It allows an estimation of class probability and performs no pruning [11].

E. Random Forest

It is Ensemble method to improve the classification Accuracy .Ensemble method is accurate than basic classifier. Instead of one classifier is generated using sampling by replacement method. Each of the classifier in this method is a decision tree, so collection of all classifier is known as Random forest. Attribute is randomly for all classifier at each node to determine the split. Random forest is very much useful for large datasets and also learning is fast [11].The algorithm is used for misuse, anomaly and hybrid detection also improved the accuracy of the system [24].

F. SVM

The goal of the SVM is to find the best hyper plane which maximize the margin of training sets and separate the class level [11].SVM is used for numeric prediction as well as classification. The SVM is of two types C-SVM and One-Class SVM.C-SVM is used for supervised learning and One-Class SVM is used for unsupervised learning. The SVM work as follows, the original training data is transformed into higher dimension using non-linear mapping. In the new dimensions, the best hyperplane is selected which divide one data class with other. Maximize the distance of margin so each feature feel safe on both side. The vectors which are closed to hyperplane is considered as support vectors and the whole training datasets, therefore size of the datasets is not an issue this is the main advantage of SVM [15].The main disadvantage of this algorithm is time consuming for multiclass datasets. The learning ability of SVM cannot play fully to the role of data mining, and hence the detection reliability is not high [6].

Table 1 Comparison of different classification algorithm

Author	*Approach*	*Dataset*	*Result*
Guangqun Zhai, Chunyan Liu (2010)	Improved ID3	Online data	1. Time Complexity is less as compared to original ID3. 2. False alarm rate is low
Neha G. Relan, Prof. Dharmaraj R. Patil (2015)	C4.5 & C4.5 with Pruning	KDDCup 99	1. Accuracy of C4.5 with pruning is improved as compared to C4.5
Neha G. Relan, Prof. Dharmaraj R. Patil (2015)	C4.5 & C4.5 with Pruning	NSL_KDD	1. Accuracy of C4.5 with pruning is improved as compared to C4.5
Huaibin Wang, Boting Chen (2013)	MultiStrategy Pruning	KDDCup 99	1. Improved Detection Rate of R2L, DOS and Normal data as compared to C4.5. 2. Reduce the number of Nodes without increasing the Detection Error Rate.
Shailendra Sahu, B M Mehtre (2015)	J48	Kyoto 2006+	1. Improve Precision and Recall of attack and Normal data but not correctly classified the unknown attack.
Kailash Shivshankar Elekar (2015)	Random Forest & Random Tree	KDDCup 99	1. Random Tree correctly detects R2L and U2R attacks. 2. Random Forest correctly detects DoS and Probe attacks.
Manjiri V. Kotpalliwar, Rakhi Wajgi (2015)	SVM	KDDCup 99	1. It increase the Validation and Classification Accuracy. 2.Time Complexity increases.

IV. HYBRID METHODS

Signature learning methods are not able to recognize unknown attacks and the attacks with high accuracy are unable to detect by Anomaly learning methods [18]. Some intrusion behaviour are similar to normal or other attributes. Using any particular algorithm cannot give proper result. So, to increase the accuracy and the detection rate hybrid method is used.

A. K-Means with C4.5

K-Means is a type of Clustering Algorithm, which suffers from two problems the first one is Class Dominance Problem and second one is Forced Assignments Problem. Combination of K-Means with C4.5 eliminates the K- Means problems. The reasons for using K-Means algorithm for making cluster of instance in first stage is 1. K-Means use greedy search strategy which guarantees local minimum of criterion function 2. It is data consumed method with relatively few assumptions on the distribution of underlying data [10] [12]. This hybrid algorithm works in two phases which as follows:

a) Selection Phase: It is used to find closest cluster using Euclidean distance of training datasets.

b) Classification Phase: Classify whether the given test tuple is normal or attack.

K-Means usually tacke care that each sample belongs to only one cluster. If any of the sample belong to two cluster then C4.5 classifies it with If then rule [10].

B. K-Means with Naïve Bayes

Cascading of supervised learning with unsupervised learning improves the result. K-Means is a type of partitioning method, which uses mean to represent a cluster center. Naïve Bayes classifier assumes that any effects in an attribute value of the given class are independent on the other attribute value. The hybrid method divided into two stages:

a) First Stage: Based on their attack behaviour a group is formed of similar attribute of a datasets as a pre-classification component.

b) Second Stage: The resulting cluster is classified into attack classes using Naïve Bayes. Some data can be missclassified in First stage can be correctly classified in the second stage.

Further the hybrid approach is compared with Naïve Bayes, the result obtained from hybrid method was found to be more accurate.

C. J48 with Random Tree

The accuracy of normal data is more as compared to other attacks using J48 whereas accuracy of U2R and R2L is more using Random Tree and using Random Forest detection rate is more for DoS and Probe attack. Using different algorithm for individual attack is a tedious job because of this time complexity and space complexity will increase. To overcome this problem and to improve the accuracy author proposed the combinations of the algorithms [13]. Category wise attacks are detected using combination of different algorithms such as J48 with Random Tree, Random Forest with Random Tree and J48 with Random Forest. Random Forest with Random Tree gives better detection rate and false attack detection rate for only Probe type attack and J48 with Random Forest does not improve the detection rate of any attacks as compare to other two algorithms. The advantages of the J48 with Random Tree are discussed in Table 2.

Table II Comparison of different classification algorithm

Author	*Approach*	*Dataset*	*Result*
Muniyandi, Amuthan Prabhakar, R. Rajeswari, and R. rajaram (2012)	K-Means with C4.5	KDDCup 99	1. The True Positive Rate, Precision and F-Measure increase as compared to K-Means and C4.5
Muda, Z. Yassin, W. Sulaiman, M. N. Udzir, N. I. (2011)	K-Means with Naïve Bayes	KDDCup 99	1. The detection rate of Probe, Normal, U2R, DoS increase as compare to Naïve Bayes.
Elekar, Kailas Shivshankar (2015)	J48 with Random Tree	KDDCup 99	1. The detection rate and false attack detection rate is improved for DoS, U2R, R2L, Normal.

V. DATASETS

For designing an efficient and powerful classifier, dataset plays a very important role. Dataset is a collection of raw data from different sources. Before applying different algorithm on dataset it is important to pre-process the data. Pre-processing the data can improve accuracy and reduce the training time of the classifier. Data processing consists of removing redundant data, filling missing values, selecting the important features to reduce the dimension of the dataset. Santosh kumar sahu used efficient algorithm to fill the missing value, removing redundant data, selecting important features and to normalize the dataset [4]. Lots of research has been done on different IDS datasets named as KDDcups 99, NSL_KDD, Gure_KDD and Kyoto 2006+ dataset. Attack distribution in kddcup 99, NSL_KDD is shown in table 3 [4].

A. KDDcup 99:

MIT, Lincols labs has developed quality 1999 KDD dataset for surveying, evaluating research in intrusion recognition using DARPA intrusion detection evaluation method. The KDDcups 99 data consist of 41 attributes out of which 3 are categorical and 38 are discrete numeric or continuous attributes. The attributes are classified into 4 groups: basic features, content features, time based traffic features and host based traffic features [3]. By analyzing dataset it is recognized that out of 23 levels, one level is fall under normal data category and remaining 22 are diverse attack. 22 attack is divided into 4 categories named DOS, U2R, R2L, and Probe.

a) Denial of service (DOS) :
Dos attack tries to stop the authorized user for accessing or consuming the services .

b) Remote to local (R2L):
In R2L unauthorized user attempts to get the access of the victiim's system by guessing password or breaking it.

c) User to root (U2R):
In this intruder has the right to access local machine but tries to get the access of super user(administrator) .

d) Probe:
Probe attempts to steal the data of the target machine.

KDD cups99 dataset can be obtain from 3 diverse files known as KDD full dataset, 10% KDD and corrected KDD dataset. Detail of all three dataset such as number of records, attack category and removing redundant data are given in Table 3.

B. NSL_ KDD:

NSL_KDD dataset are designed to overcome the problem of KDDcup 99. NSL_KDD remove the redundancy of a dataset.

NSL_KDD training set has 125973 tuples and testing set have 22544 tuples. The attributes in NSL_KDD is same as KDDcup 99 dataset.

C. GureKDDcup:

In GuruKDDcup the connection is obtained from kddcup99 database. But its payload is added for each tuple. By using payload of each tuples the information is removed. The GureKDDcup dataset follows same procedure to generate KDDcup 99 dataset. GureKDDcup contains 41attributes are divided into 3 groups. Intrinsic attributes are obtained from header of the packets. Content attributes are obtained from the content of the packet based on the expert person knowledge. Traffic attributes are calculated taking previous connection into account. GureKDDcup 99 dataset size is 9.3 GB and six percent of GureKDDcup dataset is 4.2 GB.

D. Kyoto 2006+:

KDDcup 99 cannot consider the recent network situations and latest attack pattern so Shailendr Sahu [14] used a new dataset called Kyoto 2006+, in which 3 years (November 2006-Aug 2009) of real traffic information is collected together obtain from different types of honeypots, Kyoto 2006+ consist of twenty four feature, 14 features are same as KDDcup 99 and rest are additional features.

Table III A detail analysis of datasets

Type	*DoS*	*U2R*	*R2L*	*Probe*	*Normal*	*Total*
KDD full	3883370	52	1126	41102	972781	4898431
After Removal of Redundant data in KDD full	247267	52	999	13860	812814	1074992
KDD 10%	391458	52	1126	4107	97278	494021
After Removal of Redundant data in 10% KDD	54598	52	999	2133	87832	145586
KDD Corrected	229269	70	16172	4925	60593	311029
After Removal of Redundant data in KDD Corrected	22984	70	2898	3426	47913	77291
NSL_KDD	45927	52	995	11656	67343	125973

VI. CONCLUSION

Various data mining methods are discussed for improving the detection rate and reduce the false alarm rates that had been implemented in the past few years. The different developments in data set selection, feature selection, classification algorithms and Classification via Clustering algorithms have been reviewed through this paper. This review will be helpful to the researchers for gaining information about the various algorithms with their strengths and limitations. The Classification Algorithm can only detect known Intrusion. From this survey, we have considered few classification algorithms to find out the suitable technique for Intrusion Detection. After the review of different classification algorithms for Intrusion detection, Decision Tree algorithm is the best classification algorithm. The reviewed Decision Tree algorithms detection rate is better when compared to other algorithms. Basic Detection Tree algorithms detection rate is 89.7%. C4.5 with pruning detection rate is 92% using KDDcup 99. Whereas the same algorithm with NSL_KDD resulted with 98.45%. The existing algorithms are reviewed for classification via clustering. The hybrid model with K-means via Naïve Bayes algorithms detection rate is 92.12%. Similarly K-means via C4.5 algorithms detection rate is 99.6%. Out of Decision Tree Algorithms, C4.5 accurately detects the Normal data where as Random Tree correctly detect R2L and U2R. DoS and Probe is correctly detect by Random Forest. To detect unknown attack Clustering Algorithm needs to be implemented, but Clustering would give false positive result. The Hybrid Algorithm (Classification via Clustering) will reduce the false positive and false negative to improve the accuracy of the detection.

ACKNOWLEDGMENT

I would like to give my sincere thanks to Dr. Dhirendra Mishra and Prof Avanish Tiwari for their unconditional and continuous support and guidance.

REFERENCES

[1] Pu, Wang, and Wang Jun-qing, Intrusion detection system with the data mining technologies , Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on. IEEE, 2011.

[2] Relan, Neha G., and Dharmaraj R. Patil, Implementation of network intrusion detection system using variant of decision tree algorithm ,Nascent Technologies in the Engineering Field (ICNTE), 2015 International Conference on IEEE, 2015.

[3] Kayacik, H. Gnes, A. Nur Zincir-Heywood, and Malcolm I. Heywood. Selecting features for intrusion detection: A feature relevance analysis on KDD 99 intrusion detection datasets, Proceedings of the third annual conference on privacy, security and trust. 2005.

[4] Sahu, Samir Kant, Saumendra Sarangi, and Sanjaya Kumar Jena,A detail analysis on intrusion detection datasets. Advance Computing Conference (IACC), 2014 IEEE International. IEEE, 2014.

[5] Prajapati, Naveen Mohan, Atish Mishra, and Praveen Bhanodia. Literature survey-IDS for DDoS attacks, IT in Business, Industry and Government (CSIBIG), 2014 Conference on. IEEE, 2014.

[6] Kumar, Manoj, M. Hanumanthappa, and TV Suresh Kumar. Intrusion detection system using decision tree algorithm,Communication Technology (ICCT), 2012 IEEE 14th International Conference on. IEEE, 2012.

[7] Wang, Huaibin, and Boting Chen. Intrusion detection system based on multi-strategy pruning algorithm of the decision tree, Grey Systems and Intelligent Services, 2013 IEEE International Conference on IEEE, 2013.

[8] Zhai, Guangqun, and Chunyan Liu. Research and improvement on ID3 algorithm in intrusion detection system., Natural Computation (ICNC), 2010 Sixth International Conference on. Vol. 6. IEEE, 2010.

[9] Thakur, Devashsish, Nisarga Markandaiah, and D. Sharan Raj. Re optimization of ID3 and C4. 5 decision tree. Computer and Communication Technology (ICCCT), 2010 International Conference on IEEE, 2010.

[10] Muniyandi, Amuthan Prabakar, R. Rajeswari, and R. Rajaram.Network anomaly detection by cascading k-Means clustering and C4. 5 decision tree algorithm, Procedia Engineering 30 (2012): 174-182.

[11] Aggarwal, Preeti, and Sudhir Kumar Sharma. An empirical comparison of classiers to analyze intrusion detection, Advanced Computing Communication Technologies (ACCT), 2015 Fifth International Conference on. IEEE, 2015.

[12] Yanjun, Zhao, and Wang Jing, Realization of intrusion detection system based on the improved data mining technology ,Computer Science Education (ICCSE), 2013 8th International Conference on. IEEE, 2013.

[13] Elekar, Kailas Shivshankar, Combination of data mining techniques for intrusion detection system, Computer, Communication and Control (IC4), 2015 International Conference on. IEEE, 2015.

[14] Sahu, Shailendra, and B. M. Mehtre, Network intrusion detection system using J48 Decision Tree, Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on. IEEE, 2015.

[15] Kotpalliwar, Manjiri V., and Rakhi Wajgi, Classication of Attacks Using Support Vector Machine (SVM) on KDDCUP'99 IDS Database,Communication Systems and Network Technologies (CSNT), 2015 Fifth International Conference on. IEEE, 2015.

[16] Agrawal, Shikha, and Jitendra Agrawal, Survey on Anomaly Detection using Data Mining Techniques ,Procedia Computer Science 60 (2015): 708-713.

[17] Kruegel, Christopher, Fredrik Valeur, and Giovanni Vigna,Intrusion detection and correlation: challenges and solutions, Vol. 14. Springer Science and Business Media, 2004.

[18] Muda, Z., Yassin, W., Sulaiman, M. N., Udzir, N. I. (2011, July), Intrusion detection based on K Means clustering and Nave Bayes classication, In Information Technology in Asia (CITA 11), 2011 7th International Conference on (pp. 1-6). IEEE.

[19] Bashir, Uzair, and Manzoor Chachoo, Intrusion detection and prevention system: Challenges an opportunities., Computing for Sustainable Global Development (INDIACom), 2014 International Conference on. IEEE, 2014.

[20] Padmadas, M., et al, Layered approach for intrusion detection systems based genetic algorithm,Computational Intelligence and Computing Research (ICCIC), 2013 IEEE International Conference on. IEEE, 2013.

[21] Aslahi-Shahri, B. M., et al. "A hybrid method consisting of GA and SVM for intrusion detection system." Neural Computing and Applications (2015): 1-8.

[22] Kotsiantis, Sotiris B, Decision trees: a recent overview,Articial Intelligence Review 39.4 (2013): 261283.

[23] Han, Jiawei, Micheline Kamber, and Jian Pei, Data mining: concepts and techniques ,Elsevier, 2011.

[24] Zhang, Jiong, Mohammad Zulkernine, and Anwar Haque, Random-forests-based network intrusion detection systems ,Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 38.5 (2008): 649-659.

[25] http://www.takakura.com/Kyoto_data/.

[26] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.,

[27] http://nsl.cs.unb.ca/NSLKDD/

[28] http://www.sc.ehu.es/acwaldap/.,

[29] Kruegel, Christopher, Fredrik Valeur, and Giovanni Vigna, Intrusion detection and correlation challenges and solutions - chapter 2 ,Springer Science and Business Media, 2004.