Volume 9, Issue 5
October 2020

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

A Survey on Feature Selection Techniques in Intrusion Detection

Waleed Mohammed Al-Geshari , Iftikhar Ahmad and Madini O. Alassafi

King Abdul-Aziz University, Department of Information Technology,

Faculty of computer science and information technology) Jeddah, Saudi Arabia

E-mail: walgshari@stu.kau.edu.sa

ABSTRACT

Recently many people have relied on computer devices. It has become difficult to prevent our system from the attackers. Cyber-attacks can cause large damage to systems and heavy loss of data, it is necessary to find a solution to detect and prevent such unacceptable and illegal operations. One major difficulty in detecting attacks is the need to analyze a large amount of information to detect threats. It could face intrusion detection of large data challenges. Selecting features can help mitigate the large data challenges that intrusion detection faces by focusing only on the most relevant data. In this paper we provide an overview of intrusion Detection system(IDs) and different feature selection techniques with their accuracy and the used dataset. The previous research of the feature selection has been very focused on the KDD dataset of the IDS. The researchers need access to additional, high quality, publicly available data sets. We give and overview of the dataset that have been used in FS for IDS.

Keywords: *Feature Selection, Intrusion detection, Dataset*

1. INTRODUCTION

In recent years, the Internet has come to be a necessary part of daily activities for people, such as working online, sale and buy, pay the bills, exchange money, booking tickets and, on these days they rely on internet 100% for studies because corona Covid- 19 virus and other services. In addition, the organizations of the government help to keep their information protected, available and reliable by store over the network.

Most recently researchers have been used different techniques in order to improves the detection rate. But there are some drawbacks of their techniques. Some of them it takes time to determine the optimal in a reign of convergence, and sometimes unable to find a suitable solution [1]. So researchers looking forward in order to fine a method or technique that can give better efficient in the performance.

In this paper, we make an overview of some techniques that proposed in field of the feature selection research by exploration of surviving contributions. Further, we provide an overview of intrusion Detection system(IDs) and different feature selection techniques with their accuracy and the used dataset.

In section 2, we introduce the IDS and explain its types. Feature selection (FS) and its categories with related work in section 3. In section 4, Different of public available datasets are described. In section 5, Conclusion and future work are described.

2. Intrusion Detection System

In the recent days, the world is depending more and more on sensors and connected actuators that look alike to set the life and IDS both are the systems that designed for monitoring and analyzing the network communication thus monitoring enhances the detection of intrusions.

The goal of IDSs is to transact with weak systems. They are not designed to replace but complement the security technique that is already existing. Therefore, community or users are allowed to keep their systems with efficient systems by sensing security trespasses that are not detected by other security tools. Additionally, IDSs provides helpful information about breaches by taking elastic and suitable measures to preventing them. Thus, it works as quality control for security design and Managements.

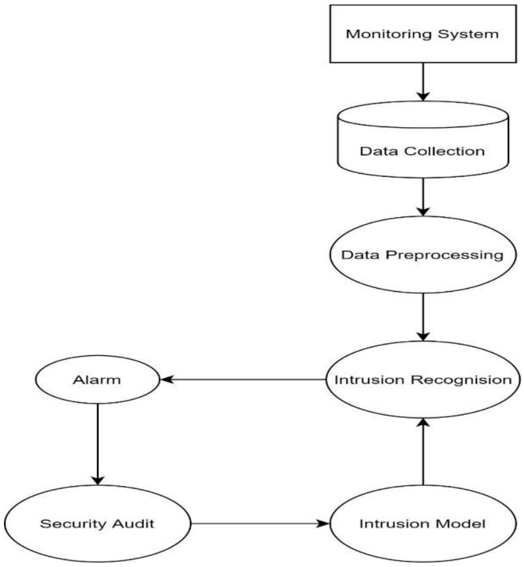

Figure 1: Intrusion Detection Architecture

the IDSs are defined Based on NIST that is the special publication on intrusion detection systems, [2] as “software or hardware systems that automate the process of monitoring the events occurring in a computer system or network, analyzing them for signs of security problems". Figure 2.1 shows the general structure of the intrusion detection system.

IDSs work in such a way that suspected behaviors is specified on three bases in computer systems and networks: Anomaly-based detection, Misuse-based detection, and State-full protocol analysis. These types [3-8], can be integrated and used individually as well.

2.1. Anomaly-based Detection

IDSs create a model of signals for the monitored system's normal behavior and then find perversions using nothing but an observation. The historical data constructs the reference models that are gathered over a long and normal duration of the operation. Thus, audited data is collected by IDSs while using different measures for identifying activities to deviate from the normal line. Thus, the prediction of different behavior from the norm alters this method.

2.2 Misuse-based Detection

IDSs in this type of detection look alike to sense breaches on the footing of predetermined designs of the known attacks. A particular sequence of jobs or a patterns of suspect events is included through which attacks are qualified, that is why system activities are analyzed by IDSs for finding events that seem to match other events according to such defined patterns. Misuse-based detection is also known as signatures-based detection.

2.3 State full Protocol Analysis

In this type, IDSs function to identifying events' deviation through the comparison of accepted qualifier' predetermined profiles of protocol activity that are normal for each and every state of the protocol. Indeed, it is different from the detection that is anomaly-based that commonly relies on universal profiles which are vendor-developed determining the working of certain protocols. There is another kind of detections that are referred to as IDSs which are hybrid in nature. It is simply designed by combining anomaly-based and misuse-based on a single system. In IDSs, this technique of detection focuses on the multiple advantages, approaches thus, overcome several issues by the production of IDS which is stronger.

3 Features selection Techniques

Feature selection has become the focus of many research areas in recent years. In particular, the wrapper filter method fine-tunes the population of GA solutions by adding or deleting features based on univariate feature ranking information [9].

feature selection that defines which features are distinctive The performance of the classification depends very much on defining the features. Therefore, the selection of a discriminatory feature is very important and a very big challenge which is the focus of this research work.

To detect such attacks, the feature subset selection is needed.it is a big challenging task in order to determine the more sensitive features, to ensure that the dataset does not contain irrelevant/redundant features, and the performance of the classifiers is influenced if there are a large volume of features.

Figure 2 show the general procedure of the feature selection algorithms, at the first it initializes the population then inside loop to select subset features then evolutionary machinists based on selection, crossover, and mutation. Finally end the loop.

```
BEGIN
        Initialize: initial population (feature subset);
        While
         do
                all feature subsets in the population
                to evaluate;
                evolutionary machinists based on
                selection, crossover, and mutation;
        End While
 End
```

Figure 2: General Procedure of FS

The feature selection algorithms aim to be divided into three categories as follows: wrapper, filter and hybrid method.

3.1 Wrap Method

The strategy of the wrapper is used as learning procedure in assessing the subcategory of chosen features. Thus, this resulting algorithm used in learning is utilized as an alternate path for the inquiry and search. The target work which is a condition with specific requirements is required in an optimized form. This is used to estimate.

T. A. Alamiedy et al [10]. subdivision of partial feature based on their perceptive recession. This is called s the rate of detection. This resultantly delivers two technique of feedback namely: the evaluation and the search. From the researcher's viewpoint, careful parameter locales are used that are then applying the components of evaluation. It uses an algorithm of learning calculation to assess the non-valuable highlights and in this way carries better component subsets.

Alomari et al. [11] anticipated a wrapper based component choice approach employing the Bees Algorithm (BA) as a search technique for subclass generation, and also applying SVM as the classifier. The analyses applied four random subsets collected from KDD-Container 99. Every subset holds around 4000 records. The performance of the expected method is measured by means of the standard IDS guesstimates. The result shows that the accuracy achieved (99%) and the feature set reduced to (8) features, while the false alarm rate was (0.004).

Manosij Ghosh et al [12] present wrapper-filter FS algorithm, RMA Method for features selection. It was in domain biomedicine. And the authors used the classifiers, SVM, MLP, KNN. To evaluate their work, they found that with MLP the Accuracy on MLL Dataset 68.57%.

3.2 Filter Method

This method filters insignificant features that have a little option in the analysis of data. The selection of features is independent of any machine learning algorithm. The features that are carefully selected are resultantly assessed in accordance to the characteristics of the data.

An intrusion detection system that utilizes a filter-based feature selection algorithm was projected via Ambusaidi et al. [13]. They expected a common information-based algorithm that systematically indicates the ideal element for grouping. This mutual information-based feature selection procedure can deal with linearly and nonlinearly dependent on data features. Its efficiency is assessed in the instances of network detection. An IDS, labelled Least Square Support Vector Machine based IDS (LSSVM-IDS) was collected using the features chosen by the suggested feature selection procedure. The performance of LSSVM-IDS was assessed utilizing three intrusion detection evaluation datasets, to be exact KDD Ampule 99, NSL-KDD and Kyoto 2006 dataset. The assessment obtained established that the nose selection procedure contributed more serious features for LSSVM-IDS to accomplish improved precision and lesser computational cost contrasted and the best in class techniques.

3.3 Hybrid Method

The hybrid method is a combination of the wrapper and filter approaches to achieve better performance. It implicitly or explicitly uses the FS algorithm. Examples of this method are the decision tree and the naive Bayes classifier, among others.

There are different of features selection techniques proposed by several researchers, in the following paragraphs a reviews of some of these techniques.

Eid, H.F et al [14]. presents a method for feature selection by using a linear correlation for building NID model. The proposed method introduces a capable way of examining feature redundancy. It contains two steps; the first step is to select a feature subset based by the analysis of Pearson correlation coefficients. While, the second step the new dataset of features is selected from within the first step features subset. The experiments applied on NSL-KDD dataset, and the linear correlation based selection of the feature shows improvement of the accuracy.

Luo, B et al [15]. Prepare a method to evaluate the distance between datasets in 5 classification problem, and based on the 4 star angels of star image the generate the numerical features for network data from KDDcup99. They proposed and efficient IDs with less features and it achieved high accuracy.

Balakrishn , S.. et al [16]. Proposed a new feature selection algorithm which is an Optimal Feature Selection algorithm based on Information Gain Ratio. This method used the KDD Cup dataset to selects the optimum number of features. In addition, they used two classifications which are Rule Based and Support Vector Machine Classification have been used for effective of the data set. They found the proposed method is effectively decreases the false alarm rate and it is very efficient in detecting DoS attacks.

Bostani, H., and Sheikhan.M [17] provides wrapper-based feature selection method for accomplishment of global search. It known as MI-BGSA, used BGSA.This approach found the features seeing the least joblessness to the selected features also the most significance to the target class. The investigational results on the NSL-KDD dataset displayed that the suggested method can reduce the feature space dramatically.

Thaseen, I.S.,and Kumar, C.A. Presents three different methods in this field, the first method is a hybrid feature selection based on Chi-square and multi class SVM [18].In the second method they proposed also a hybrid method for feature selection based on Chi-square and modified BN [19], and the third method used the Consistency based feature selection , SVM, and LPBoost [20]. All of these methods used the NSL_KDD Dataset to select the feature from it. The different methods evaluated and showed a good results compared with others technique.

Aljawarneh, S., et al [21]. Used the vote algorithm with Information Gain to design a hybrid mode for feature selection in intrusion detection system. This method used the NSL-KDD standard data. The number of the subset features is one and the number of the feature in the subset is eight features. The authors estimate the scope threshold degree in the intrusion detection system by using different classifier; Meta Pagging, REPTree, DT, RandomTree, AdaBoostM1, BN, and DecisionStump. This method showd excellent result with accuracy 99.81.

In study [22] the authors used Genetic Fuzzy method which can entertainment as a genetic feature selection method for outcome the optimal features grouping. Their experiments on KDD CUP 99 dataset, it showed an efficient detection rate for traffic attack and low false alarm for normal.

Table 1 Shows the some of the recently work that have been done in field of the feature selection in different domains, it shows the type of feature selection method, the selected evaluator., in which domain the work has been done, and the accuracy for each of them. It shows approaches among the feature selection type Filter, wrap, and hybrid.

Table 1: Comparative Analysis Between the Recent Related Work

Ref	Author	Year	No. of Featuers	FS	Classifier	Dataset	Domain	Accuracy
[14]	Eid, H.F., et al	2013	17	Pearson correlation coefficient	C4.5	KDDcup99	IDS	91.1
[15]	Luo, B et al	2018	4,16	Visualized feature generation	SVM and Generation technique.	KDDcup99	IDS	94.35
[16]	Balakrishnan, S. et al.	2014	10	Information Gain Ratio	SVM	KDDcup99	IDS	93.34
[17]	Bostani, H., and Sheikhan.M,	2017	5	Gravitational Search Algorithm Binary (BGSA) and Mutual Information with SVM.	SVM	NSL-KDD	IDS	88.36
[18]	Thaseen, I.S., and Kumar, C.A	2017	31	multi class SVM Chi-square and	multi-class SVM	NSL-KDD	IDS	98
[19]	Thaseen, I.S.,and Kumar, C.A	2016	22	Chi-square and modified NB.	Naïve Bayes	NSL-KDD	IDS	96.8
[20]	Thaseen, I.S., Kumar, C.A	2016	10	Consistency based feature selection , SVM, and LPBoost	Fusion	NSL-KDD	IDs	96.2
[21]	Raman, M.G. et al.	2017	35	Hypergraph-Genetic algorithm and SVM.		NSL-KDD	IDs	96.72
[22]	Aljawarneh, S., et al	2017	8	Vote algorithm with Information Gain	DT, Meta Pagging, Random Tree, REPTree, AdaBoostM1, DecisionStump, and BN.	NSL-KDD	IDs	99.81
[23]	Tsang. C. et al.	2007	25	Genetic and fuzzy rule	multi-objective technique	KDDcup99	IDs	92.76

4 Dataset

In this review paper we explore 10 different datasets illustrate in Table 2. Here adscription for each of them:

Table 2: Public Datasets

#	Dataset	Year	Ref
1	DARPA	1998	[24]
2	KDD Cup 1999	1999	[25]
3	NSL-KDD	2009	[26]
4	UNSW-NB15	2015	[27]
5	DEFCON	2000	[28]
6	CAIDAs	2017	[29]
7	LBNL	2016	[30]
8	UMASS	2018	[31]
9	CIC DoS	2017	[32]
10	CICDS2017	2017	[33]

4.1 DARPA

This dataset is available on 1998, it is based on the audit logs and network traffic. The training data created by network attack in seven weeks, while the testing data created in two weeks from network attacks. Based to Sharafaldin et al. [24], DARPA dataset does not characterize as real-world network traffic.

4.2 KDD Cup

This dataset is created based on DARPA'98 IDS estimation program and it a network traffic for seven weeks. It contains 4,900,000 vectors. The cyber-attacks are categorized into the following groups: User to Root attack (U2R), Denial of Service attack (DoS), Remote to Local attack (R2L), and Probing attack. The KDD Cup 1999 dataset contains 41 features, which are categorized into the following three classes: basic features, content features, and traffic features [25].

4.3 NSL-KDD

This one is proposed by Tavallaee et al. [26], it is suggested to resolve some of the problems of the KDD'99 datasets. It enhanced the KDD dataset in following improvements: no redundancy, there is no duplicate inputs, the number of selected inputs is organized, and (4) the number of inputs is reasonable. There are many papers on intrusion detection use both datasets together in the evolution of the performance, and they find that the finest results are found in the NSL-KDD.

4.4 UNSW-NB15

This dataset is Generated by four tools, IXIA PerfectStorm, Tcpdump, Argus, and Bro-IDS. These tools are used to create some types of attacks, including DoS, Exploits, Shellcode, Reconnaissance, Generic, and Worms [27].

4.5 DEFCON

It is created by two versions, counting, DEFCON8 (2000) and DEFCON-10 (2002). Attacks in DEFCON-8 contains ports scan and buffer overflow, while in DEFCON-10 it holds probing and non-probing attacks. Both of them are castoff by Nehinbe Ojo Joshua [28].

4.6 CAIDAs

It projected by the Center of Applied Internet Data Analysis, which comprises diverse datasets, containing, CAIDA DDOS, CAIDA Internet traces 2016, and RSDoS Attack Metadata (2018-09). The RSDoS Attack Metadata (2018-09) contains the unsystematically fooled denial-of-service attacks inferred from the backscatter packets organized by the UCSD Network Telescope[29] .

4.7 LBNL

Gathered by using the uPMU at the Lawrence Berkeley National Laboratory electrical network. The uPMU is micro-phasor measurement units, it products 12 streams of 120 Hz high precision values with timestamps accurate to 100 ns [84].

4.8 UMASS

This dataset has two different datasets, containing, strong flow correlation attacks and simple timing attack on OneSwarm. The first one is proposed by Nasr et al. [30], they charity several Tor clients to surf the top 50,000 Alexa websites over Tor. The second dataset is proposed by Bissias et al. [31], where the attacks adhere to the limitations of unnatural generally applicable criminal procedure.

4.9 CIC DoS

It comprises of data captured from July 3, 2017, to July 7, 2017. The CICIDS2017 dataset is proposed Jazi et al. [32] by application layer DoS attacks are normally perceived in high-volume or low-volume variations.

4.10 CICDS2017

This dataset contains data captured in the same period of the CIC DoS dataset captured. The CICIDS2017 is proposed by Sharafaldin et al. [33], by implements attacks contain Brute Force SSH, DoS, Heartbleed, Web Attack, Infiltration, Botnet and DDoS, and Brute Force FTP.

3. CONCLUSION

In this research, we provide a survey paper about feature selection techniques in intrusion detection system. We explore a background of IDS, and survey different future selection technique and data set they use, each method used different classifier in order to validate the and improve the performance accuracy. Different FS technique described and different of public data set are described. We interested to extend this work, in future work we excite to investigate different Methods and evaluate the performance with different evaluators in this feild.

REFRENCES

[1] Zuech R, Khoshgoftaar TM, Wald R "Intrusion Detection and Big Heterogeneous Data: a Survey." Journal of Big Data 2015 2(3):1–41 doi:10.1186/s40537-015-0013-4 http://www.journalofbigdata.com/content/2/1/3.

[2] Bace, R., & Mell, P. (2001). NIST special publication on intrusion detection systems. BOOZ-ALLEN AND HAMILTON INC MCLEAN VA.

[3] Al-Dhafian B., " Towards Optimal Classification Technique for Intrusion Detection " M.Sc. diss., King Saud University, 2016.

[4] Aydın, M. A., Zaim, A. H., & Ceylan, K. G. (2009). A hybrid intrusion detection system design for computer network security. Computers & Electrical Engineering, 35(3), 517-526

[5] Patel, A., Taghavi, M., Bakhtiyari, K., & JúNior, J. C. (2013). An intrusion detection and prevention system in cloud computing: A systematic review. Journal of network and computer applications, 36(1), 25-41.

[6] Dewa, Z., & Maglaras, L. A. (2016). Data mining and intrusion detection systems. International Journal of Advanced Computer Science and Applications, 7(1), 62-71.

[7] Mittal, S. (2014). Data Mining Approach IDS K-Mean using Weka Environment. International Journal of Advanced Research in Computer Scienceand Software Engineering, 4(8), 482-488.

[8] Patel, J., & Panchal, K. (2015). Effective intrusion detection system using data mining technique. Journal of Emerging Technologies and Innovative Research, 2(6), 1869-1878.

[9] R. Kohavi and G. H. John, "Wrappers for feature subset selection," Artificial intelligence, vol. 97, pp. 273-324, 1997.

[10] [10] Alamiedy TA, Anbar M, Al-Ani AK et al (2019) Review on feature selection algorithms for anomaly-based intrusion detection system. Adv Intell Syst Comput 843:605–619.

[11] Alomari, O., Othman, Z.A.: Bees algorithm for feature selection in network anomaly detection. J. Appl. Sci. Res. 8, 1748–1756 (2012).

[12] M Ghosh, S Begum, R Sarkar, D Chakraborty, U Maulik - Expert Systems with Applications, 2019.

[13] M. A. Ambusaidi, X. He, Z. Tan, P. Nanda, L. F. Lu, and U. T. Nagar, "A novel feature selection approach for intrusion detection data classification," Proc. - 2014 IEEE 13th Int. Conf. Trust. Secur. Priv. Comput. Commun. Trust. 2014, pp. 82–89, 2015.

[14] Eid, H.F., Hassanien, A.E., Kim, T.-h., Banerjee. S, "Linear correlation-based feature selection for network intrusion detection model," in Proc. of Advances in Security of Information and Communication Networks. pp. 240-248. Springer, 2013.

[15] Luo, B., Xia. J, "A novel intrusion detection system based on feature generation with visualization strategy," Expert Systems with Applications, Vol. 41, No. (9), PP. 4139-4147, 2014.

[16] Balakrishnan, S., Venkatalakshmi, K., Kannan. A, "Intrusion detection system using Feature selection and Classification technique," International Journal of Computer Science and Application (IJCSA) Vol. 3, No. (4), November 2014, 2014.

[17] Bostani, H., Sheikhan.M, "Hybrid of binary gravitational search algorithm and mutual information for feature selection in intrusion detection systems," Soft computing, Vol. 21, No. (9), PP. 2307-2324, 2017.

[18] Thaseen, I.S., Kumar, C.A, " Intrusion detection model using fusion of chi-square feature selection and multi class SVM," Journal of King Saud University-Computer and Information Sciences, Vol. 29, No. (4), PP. 462-472, 2017.

[19] Thaseen, I.S., Kumar, C.A, "Intrusion Detection Model Using Chi Square Feature Selection and Modified Naïve Bayes Classifier," in Proc. of Proceedings of the 3rd International Symposium on Big Data and Cloud Computing Challenges (ISBCC–16') 2016, pp. 81-91.

[20] Thaseen, I.S., Kumar, C.A, "An integrated intrusion detection model using consistency based feature selection and LPBoost," in Proc. of Green Engineering and Technologies (IC-GET), 2016 Online International Conference on 2016, pp. 1-6. IEEE, 2016

[21] Raman, M.G., Somu, N., Kirthivasan, K., Liscano, R., Sriram, V.S, "An efficient intrusion detection system based on hypergraph-Genetic algorithm for parameter optimization and feature selection in support vector machine," Knowledge-Based Systems, Vol. 134, PP. 1-12, 2017.

[22] Aljawarneh, S., Aldwairi, M., Yassein, M. B, "Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model," Journal of Computational Science, 2017.

[23] Tsang. C.-H, Kwong. S, Wang. H, "Genetic-fuzzy rule mining approach and evaluation of feature selection techniques for anomaly intrusion detection," Pattern Recognition, Vol. 40, No. (9), PP. 2373-2391, 2007.

[24] Sharafaldin I , Habibi Lashkari A , Ghorbani AA . A detailed analysis of the CI- CIDS2017 data set. In: Mori P, Furnell S, Camp O, editors. Information Sys- tems Security and Privacy. Cham: Springer International Publishing; 2019. p. 172–88. ISBN 978-3-030-25109-3 .

[25] Kdd Cup 1999. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html . last accessed 30 May 2019.

[26] Tavallaee M , Bagheri E , Lu W , Ghorbani AA . A detailed analysis of the kdd cup 99 data set. In: 2009

IEEE Symposium on Computational Intelligence for Security and Defense Applications. IEEE; 2009. p. 1–6.

[27] Unsw-nb15 Dataset. https://www.unsw.adfa.edu.au/unsw-canberra-cyber/ cybersecurity/ADFA- NB15- Datasets/ . last accessed 30 May 2019.

[28] Nehinbe JO . A simple method for improving intrusion detections in corpo- rate networks. In: International Conference on Information Security and Dig- ital Forensics. Springer; 2009. p. 111–22 .

[29] Center for Applied Internet Data Analysis. https://www.caida.org/data/ overview/ . last accessed 30 May 2019.

[30] Kasongo SM , Sun Y . A deep learning method with filter based fea- ture engineering for wireless intrusion detection system. IEEE Access 2019;7:38597–607 .

[31] Bissias G , Levine BN , Liberatore M , Prusty S . Forensic identification of anonymous sources in oneswarm. IEEE Trans. Depend. Secure Comput. 2015;14(6):620–32 .

[32] Jazi HH , Gonzalez H , Stakhanova N , Ghorbani AA . Detecting http-based appli- cation layer dos attacks on web servers in the presence of sampling. Comput. Netw. 2017;121:25–36 .

[33] Sharafaldin I , Lashkari AH , Ghorbani AA . Toward generating a new intru- sion detection dataset and intrusion traffic characterization.. In: ICISSP; 2018. p. 108–16 .

AUTHOR PROFILES

W. Al-Geshari received the Bachelor degree in Information Technology from King Abdulaziz University, Jeddah, Saudi Arabia in 2012 and currently, is a Master degree student in King Abdulaziz University, Jeddah, Saudi Arabia.