Published in final edited form as:

Int J Netw Secur Appl. 2020 January ; 12(1): 1–18. doi:10.5121/ijnsa.2020.12101.

A SURVEY ON THE USE OF DATA CLUSTERING FOR INTRUSION DETECTION SYSTEM IN CYBERSECURITY

Binita Bohara[1], **Jay Bhuyan**[1], **Fan Wu**[1], **Junhua Ding**[2]

[1]Dept.of Computer Science, Tuskegee University, Tuskegee, AL, USA

[2]Dept.of Information Science, University of North Texas, Texas, USA

Abstract

In the present world, it is difficult to realize any computing application working on a standalone computing device without connecting it to the network. A large amount of data is transferred over the network from one device to another. As networking is expanding, security is becoming a major concern. Therefore, it has become important to maintain a high level of security to ensure that a safe and secure connection is established among the devices. An intrusion detection system (IDS) is therefore used to differentiate between the legitimate and illegitimate activities on the system. There are different techniques are used for detecting intrusions in the intrusion detection system. This paper presents the different clustering techniques that have been implemented by different researchers in their relevant articles. This survey was carried out on 30 papers and it presents what different datasets were used by different researchers and what evaluation metrics were used to evaluate the performance of IDS. This paper also highlights the pros and cons of each clustering technique used for IDS, which can be used as a basis for future work.

Keywords

Intrusion detection system; clustering technique; network security

1. INTRODUCTION

Due to the increasing growth of computer network usages, network security is becoming an important issue. With the evolution of new technology, there is a rapid increase in the incidents of hacking and intrusion [1]. There is no doubt that all computers are suffering from security vulnerabilities, that are both technically challenging and costly for manufacturers to fix. So, any malicious activity on network security vulnerabilities or computers can seriously affect the system and breach its confidentiality, availability, and integrity. Therefore, the intrusion detection system has become an integral part of network security architectures.

An intrusion detection system can the ability to locate and identify malicious or anomalous activities in the computer network by examining the network traffic in real-time [2]. The intrusion detection system is classified into two groups: misuse and anomaly detection systems. The misuse intrusion detection system is usually used for commercial purposes because of its predictability and high accuracy, while anomaly detection is favored in

research studies. The intrusion detection system can also be classified as Host-based and Network-based, depending on the location of detection. There are many number of data mining techniques available for detecting network intrusions.

Data mining is increasingly becoming a popular technique in the network security environment for finding regularities and irregularities in datasets. Data mining can be defined as the process of using efficient techniques to extract useful and unexpected patterns from huge datasets [3]. There are different supervised techniques have been used to detect intrusion. However, this approach depends on the labeled data and requires the system to be trained on the known data. The problem with this type of technique is its high dependency on training data and the inability to detect any new type of attacks. To overcome this limitation of the supervised technique, an unsupervised approach can be used for intrusion detection, which can detect the unlabeled data. Clustering is one of the commonly used unsupervised techniques for classifying huge datasets and detecting the intrusions. Clustering is a data mining technique that is used to infer the conclusion from unlabeled data and find hidden patterns in datasets [3]. In other words, clustering is a technique that is used to group similar data into one cluster and other dissimilar data into the different clusters.

The clustering approach is based on two assumptions [4]. The first assumption is that the number of normal connections is larger than abnormal connections and the second assumption is that the feature of the abnormal network is different than the normal network. Depending upon these assumptions, there are many clustering techniques can be used for detecting intrusions and attacks such as hierarchical, partitional, grid-based, and density-based clustering techniques.

Hierarchical clustering is a clustering algorithm that constructs a hierarchy of clusters and it clusters the objects based on their distance. This technique is based on the cluster proximity measure and there are three measures such as single-linkage, average-linkage, and complete linkage. Partitional clustering splits the data points into many or some number of separate partitions, where each partition is known as clusters. K-means is one of the simplest and efficient partitional clustering algorithms that is used for detecting intrusions in a computer system. Meng et al [5] used k-means algorithms on KDD99 datasets to detect unknown attacks in different settings. Density-based clustering algorithm clusters data points based on the density of the points in a region. DBSCAN is one of the best density-based algorithms. Yang Jian et al [6] used improved intrusion detection based on DBSCAN that generates clusters depending upon the density-based method.

This paper presents a literature review on clustering techniques applied in the intrusion detection system literature. Articles that used clustering technique for intrusion detection were carefully reviewed and the type of attack different clustering techniques can detect were analyzed. The different clustering techniques were compared in terms of different evaluation metrics used, datasets used, their strengths and weakness.

The remaining part of the paper is divided into different sections. Section II describes the datasets used in intrusion detection research. Section III outlines the related definition of the terms used in this paper. Section III also describes the different terms used in the clustering

technique literature. Section IV lists the related work of the clustering technique used for intrusion detection. Section V lists the research questions used in this research. Section VI illustrates the results and discussion. Section VII concludes the paper.

2. DESCRIPTION OF DATASET

KDDCup99 dataset, a publicly available dataset, is the most common dataset that is being used for evaluating intrusion detection systems [7]. This dataset was used in the KDDCup99 competition and is based on the DARPA98 IDS evaluation dataset [8]. The dataset comprises of 4 GB of compressed TCP dump data of nearly 7 weeks of network traffic. The KDDcup99 dataset comprises of training and test datasets, where the training dataset consists of nearly 5 million datasets and the test dataset consists of 3 million datasets [9]. The datasets contain 41 features and classes: normal and attack. There are 24 different types of attacks in this dataset, which belongs to four types of major attacks such as:

Denial of Service (DOS): This is an attack where attackers make the system too busy for a legitimate user to be able to use the system or resources.

User to Root Attack (U2R): U2R starts with the attacker getting access to a normal user account on a system and exploits its vulnerability so as to get root access to the system.

Remote to Local Attack (R2L): R2L is an attack where an attacker sends packets to the system over the network and exploits its vulnerability so as to gain access to the system as a normal user. Probing Attack: This is an attack where an attacker scans the system or network to collect information so as to identify the potential vulnerability of the system.

The new version of KDD99 datasets that is publicly available is called NSL-KDD99 datasets. To solve the inherent problem of KDD-99, NSL-KDD was proposed. It contains only selected records from KDD datasets. This new dataset overcame shortcomings of KDD datasets; there were no redundant records in the training set, which removed the bias in learning algorithms towards frequent datasets. Similarly, the KDD duplicate data was also removed from the test set. NSL-KDD data consists of 21 types of attacks in the training set and 37 different types of attacks in the test set. The attacks are classified into four different types: DOS, Probe, U2R, and R2L [10]. DARPA 98 dataset is also one of the popular datasets used in the intrusion detection system to evaluate detection rate and false alarm rate in network traffic. This dataset consists of four different major attack types: DOS, Probing, U2R, and R2L. However, this dataset faced lots of criticism since a model that was used to generate the traffic was too simple [10].

Both KDD-99 and NSL-KDD do not reflect the real data flow in a computer network as they are generated on virtual networks on simulation. However, Kyoto 2006+ dataset consists of real datasets collected over 3 years from November 2006 to August 2009. This dataset comprises of 14 statistical features derived from KDD along with 10 additional features for analysis and evaluation of intrusion detection system network. These datasets were captured using honeypots, darknet sensors, email server, and web crawler [10].

Another type of dataset that is used in the literature reviewed in this research for intrusion dataset is ISCX-2012 intrusion evaluation dataset. This dataset consists of 1512000 packets with 20 features and was obtained by observing network traffic for seven days. There were 75372 normal traces and 2154 attack traces in training datasets; while in case of testing datasets, there were 19202 normal traces and 37159 attack traces [11].

3. RELATED TERMINOLOGY

- Detection Rate (DR): Detection Rate (DR) can be referred to as a number of attacks detected by the system divided by the total number of intrusions in the dataset. DR can be calculated as: DR = True Positive (TP)/(Total number of intrusions in dataset) [12].
- False Alarm Rate (FAR): False Alarm Rate (FAR) is the number of normal instances classified as attacks divided by the number of normal instances in the dataset. FAR can be calculated as: FAR= False Positive (FP)/(FP+ True Negative (TN) [12]
- False Negative Rate (FNR): False Negative Rate (FNR) represents the number of intrusions that were wrongly identified as normal [12]
- True Positive Rate(TPR): True Positive Rate corresponds to IDS that correctly identifies network activity as a malicious attack [12].
- True Negative Rate(TNR): True Negative Rate occurs when normal instances are detected as normal [12].
- Accuracy: Accuracy is defined as the percentage of correctly classified instance in datasets. Accuracy can be calculated as: Accuracy=(TP+TN)/((TP+TN+FP +FN)) [12]
- Sum of Square Error (SSE): The Sum of Square Error (SSE) is the sum of square of difference between each data point and its cluster mean.

4. RELATED WORK

The intrusion detection system plays an important role in detecting any malicious activities on a computer network. Various clustering techniques have been designed and implemented for detecting the intrusions.

Leung et al [13] were able to get a high detection rate while suffering from a high false positive rate using fpMAFIA. The performance value of fpMAFIA was 0.867, which was around 3%−12% worse off than other algorithms. However, this density and grid-based clustering had the disadvantage of not evaluating each and every point, which made the anomalies clustered into a set of small clusters and their identification was just more straightforward.

Leonid Portnoy [14] conferred the algorithm that could automatically detect both known and unknown intrusion. The author used a single-linkage hierarchical clustering method to distinguish abnormal activities from normal activities. In this algorithm, at the first number

of empty clusters was formed and the distance of instances to the cluster center was checked. The calculated shortest distance was then compared with the predefined cluster width (W) and if that distance was shorter than W, then that record was assigned to that cluster. Otherwise, a new cluster was formed and other instances were assigned to that cluster. The clusters were then updated to the mean of the cluster centers. The series of updating and reassignment took place until the cluster center was no more updated. With the algorithm, the average detection rate was around 44– 55% and the false positive rate was detected 1.3–2.3. Even though this algorithm overcame the shortcoming of K-means clustering to predict the number of clusters, it still had some limitations. One of the limitations was in determining the cluster width, which has to be determined manually. Therefore, there was a chance of mislabeling the normal instance as an abnormal and vice-versa if W was not determined properly.

Meng et al. [5] used K-means algorithm clustering technique to detect unknown intrusions in a computer network. This algorithm was used on the KDD cup 1999 dataset, where a number of clusters were chosen to be 5. Using different settings, the DR with this method was always found to be above 96 and FAR was below 4 and time complexity was low. This showed that the K-means algorithm was an effective method for intrusion detection. However, with the K-means algorithm, cluster number needs to be defined correctly in order to get correct intrusion detection. In addition, it is also sensitive to the categorical attribute and the results of this algorithm are usually not steady. In other words, the K-means clustering result can vary, even with the same input parameter. Similarly, Gerhard et al. [15] also applied a K-means clustering technique for intrusion detection. The technique was used to classify normal and abnormal network traffic flow using cluster centroids as a pattern to detect intrusion.

Guan et al. [16] used Y means clustering algorithm to form a number of ‘normal’ and ‘abnormal’ clusters. This algorithm was similar to that of K-means where datasets were partitioned into k clusters ranging from 1 to n (total number of instances). The second step was to detect the empty clusters, where new clusters were created to replace empty clusters; followed by re-assigning instances to existing centers. This process was iterated until there were no empty clusters, which eventually led to the removal of outliers (splitting) to form new clusters (merging). The last step in this process was to label the clusters based on their population, meaning that if the population of the certain cluster was higher than a given threshold of 2.32 (is the standard deviation of the data), then the population was re-classified as normal which otherwise was labeled as an intrusion. H- means+ algorithm was used for clustering KDD-99 dataset with different initial values of k and showed that the decline of SSE (Sum of Square Errors) was fast when the value of k was very small. After the turning point (k = 20, DR = 79 % and FAR = 1 %) of k, the decline of SSE was very slow and the dataset was partitioned into small clusters that were closer to each other, so there was no decrease in the value of SSE. On average, this method detected 86.63% of intrusion with a false alarm rate of 1.53%. The Y-means algorithm was again trained with 12,000 unlabeled KDD-99 dataset. The test on the trained system with 10,000 labeled instances had 82.32% DR and 2.5% FAR indicated that Y-means is one of the more promising clustering methods for intrusion detection.

Zhou et al. [4] proposed a graph-based clustering algorithm as an intrusion detection algorithm to differentiate between regular and irregular connections. The graph-based algorithm can identify clusters of any shape and it only uses a parameter and does not require to define any cluster number. This algorithm used an outlier detection method, which was based on the local deviation coefficient with different values for k (5,8,11,15). Although there was not much difference between DR and FPR depending on the value of k, the results showed that DR was higher (95.3%) and FPR was lowest (2.08%) for k equals to 8.

Wei Jiang et al. [17] used the improved fuzzy clustering algorithm for intrusion detection. The results showed that their detection rate was much higher with a lower false-positive rate over KDD's 99 datasets. In this procedure, the KDD-99 dataset was randomly grouped into 5 groups with 10 thousand records in each group. The average detection rate was 78.66% and a false positive rate of 0.704. The proposed method overcomes the disadvantage of fuzzy clustering by adding weighted value to the data object's membership degree and optimizes the clustering number by introducing validity function.

M.Jianliang et al. [5] used a K-means clustering algorithm for intrusion detection to detect the unknown attack. The algorithm was used in KDD-99 dataset, where the number of clusters (k) was 5. The experiment was carried out in a different setting and in every setting, the detection rate was always greater than 96% and a false alarm rate was below 4%. Even though the K-means clustering technique is used to detect unseen attacks and partition large data it has disadvantages of degeneracy and cluster dependency.

Li Xue–Yong and Gao Guo [18] proposed improved density-based clustering algorithm IIDBC for intrusion detection to improve the drawbacks of the density-based clustering algorithm (DBSCAN) by using the rational method in merging clusters and calculating the distance. The result showed that IIDBC improved the performance of DBSCAN, increased the detection rate and decreased the false-negative rate. The average detection rate was around 92.33/

Z.Muda et al. [19] used a new intrusion detection algorithm by combining clustering and classification techniques. This algorithm was a hybrid learning approach based on a combination of K-means clustering and OneR classification and used KDDcup 99 dataset. The main objectives of this algorithm were to separate the potential attack from normal instances into different clusters. This hybrid intrusion detection algorithm has the accuracy and detection rate above 99% and a false alarm rate below 2.75%. It was found that the performance of the hybrid classifiers was higher as compared to the single classifier. This algorithm was capable of classifying most of the instances correctly; however, it could not classify U2R and R2l attack.

Chandrashekhar et al. [20] proposed a new approach based on K-means, fuzzy neural networks and SVM classifiers as an intrusion detection technique. This technique uses KDDcup99 datasets to perform the experiment. The proposed technique achieved an accuracy of 97.78% and was effective for low-frequent attacks such as U2R and R2L.

Ravi Ranjan and G.Sahoo [12] presented a new clustering technique for detecting anomaly intrusion and attacks. They used the K-medoids method for clustering and the proposed

algorithm achieved high detection rate and overcame the K-means algorithm defects. This approach has advantages over the existing algorithm such as dependency on initial centroids, cluster number, and irrelevant clusters. The detection rate for this proposed approach was 91.2% and the false alarm rate is 3.2%. However, the detection rate for the proposed algorithm was low for probing attack (70.51%) and user to root attack (70.13%).

Chitrakar et al. [21] used a hybrid approach for anomaly-based intrusion detection by combining Naïve Bayes classification and k-Medoids clustering. This hybrid approach showed better performance as compared to k-means and Naive Bayes hybrid algorithm. The algorithm used Kyoto 2006+ dataset and the algorithm provided 4% improvement in terms of accuracy and detection rate and the false alarm rate was reduced by 1%.

Zhiengje et al. [22] used particle swarm organization with the combination of K-means clustering method for intrusion detection technique on KDD Cup 1999 data set. The best performance for this method was when the false detection rate was 2.8% and the detection rate was 82%. The invasion detected by this technique can be broadly categorized as Probing, DoS, U2R, and R2L. This method helped to detect known attacks up to 75.82% and unknown attacks to 60.8%. This method has relatively low detection for DoS, which was due to mislabeling of abnormal data as normal.

Similarly, Lizhong et al. [23] also used K-means clustering with particle swarm organizations for anomaly intrusion detection. The algorithm was used in KDD-99 datasets and the experimental results on PSO-KM showed the detection rate as 86% and false positive rate as 2.8%. The experiment shows that the accuracy was very good for attack type U2R (78%) and DOS (94%). However, the accuracy was very low for R2L at 22%.

Fatma et al. [24] used two-stage detection technique in order to improve the detection rate on DARPA dataset. The detection technique for the first stage was a self-organizing map (SOM) with K-means algorithm and neural GAS with fuzzy c-means algorithm. The second stage included SOM with K-means algorithm, support vector machine and decision trees. The results showed that for the first stage: 69%of the alerts were false attacks and only 31% were real threats, and for the second stage: 88.77% were false alerts and only 21.23% were real threats. In the first stage, the neural GAS technique provided better results since the main objective of the first stage was to cluster low-level alerts into meaningful partitions. In the second stage, the SVM algorithm provided good result to reduce the rate of false attacks.

Zhong et al. [25] used Kyoto 2006+ and KDD Cup 1999 dataset to evaluate the intrusion detection by grid-based clustering technique. The results showed that the performance was insensitive to the variation of the convergence criterion of clustering, attack, and normal condition in labeling. With the variation on stop condition of cell split, the FPR was 3.25 and DR was 58.51% for 0.001 value of and FPR and DR value were 5.14% and 62.29%, respectively for equivalent to 500.

Horng et al. [26] used KDD Cup 1999 training set for detecting different attacks, which were broadly classified as DoS, U2R, R2L, and Probe. An SVM-based intrusion detection system was combined with hierarchical clustering algorithm for IDS, which had an accuracy of 95.72 with only 0.73 false-positive rate. The detection for DoS and Probe were 99.5 and

97.5 respectively, which were comparatively higher than ESC-IDS [27], KDD'99 winner [28], KDD'99 runner-up [29], Multi-classifier [30] and Association rule [31].

Lin et al. [32] used KDD-Cup 99 dataset for intrusion detection, where the attacks were classified as normal, probing, denial of service (DoS), remote to local (R2L), and user to root (U2R). They introduced cluster center and k-nearest neighbor approach known as CANN. In this algorithm, different evaluation metrics such as accuracy, detection, and false alarms were considered to evaluate the performance of intrusion detection and they were calculated using a confusion matrix. The experiment was carried out on 6-dimensional and 19-dimensional KDD dataset [32]. The results showed that the CANN algorithm has an accuracy of 99.76%, detection rate 99.9% and false alarm rate of about 0.003 for 6-dimensional datasets. Even though the performance was shown best, CANN totally misclassified U2R and R2L as normal whereas in case of 19-dimensional KDD dataset, CANN was able to correctly classify U2R and R2L as attacks. However, the accuracy rate was very low (3.846 and 57.016).

Yongguo et al. [33] proposed a new detection algorithm known as the IDBGC algorithm, Intrusion Detection Based On Genetic Clustering. The algorithm is a combination of two stages, nearest neighbor clustering and genetic optimization. The average detection rate for this algorithm was observed around 60% and the false-positive rate was only 0.4%. The results showed that the IDBGC algorithm is feasible and effective in detecting unknown intrusions. More than 50% of DoS, R2L, and U2R unknown attacks were detected. However, for PROBE attack, the detection rate was less than 50%, which is relatively very low.

Sanjay et al. [34] proposed K-means clustering via naïve bayes classification algorithm for detecting anomaly-based network intrusion. It was observed that the above algorithm performed better on KDD cup 99 datasets compared to naive bayes classification in terms of detection rate. The performance of the algorithms was evaluated based on the accuracy, detection rate, and false alarm rate. The detection rate was observed as 99% for K-means clustering via naïve bayes algorithm whereas the false alarm rate was 4. The algorithm was shown to be efficient in detecting network intrusion; however, this approach had a high false-positive rate.

Amuthan et al. [35] proposed the anomaly network detection method based on K-means clustering and C4.5 decision tree algorithm. K-means clustering was used first to partition the training data into k clusters and then decision tree, C4.5 decision tree, was built on each cluster. The experiment was carried out on KDD99 datasets and the performance was evaluated using metrics such as true positive rate or detection rate, false-positive rate, precision, accuracy, and F-measure in percentage. The true positive rate obtained was 99.6%, the false-positive rate was 0.1%, precision was 95.6%, accuracy was 95.8% and F-measure was 94.0. The proposed algorithm gave a notable detection rate.

Reda et al. [36] proposed a hybrid network intrusion detection system based on random forests and weighted K-means. The proposed algorithm was evaluated on different percentages of KDD-99 datasets and the result showed that it achieved high detection rate as 99% and very low false-positive rate as 12.6 in anomaly detection method whereas in case of

misuse detection rate, the detection rate was 92.73% and relatively very good false positive rate as 0.54%.

Chih-Fong et al. [37] introduced a triangle area based nearest neighbor (TANN) approach to detect the intrusions in a network. The TANN approach to intrusion detection is a combination of K-means clustering and k-NN classifier. The approach was used in KDD99 dataset with 10-fold cross-validation and the result showed that TANN can effectively detect the intrusion detection with high accuracy and detection rates as compared to support vector machines, k-NN, and the hybrid classifier K-means and k-NN. The accuracy rate for the proposed algorithm was 99.01%, detection rate was very high at 99.27, and false-positive rate was 2.99%.

Li Tian et al. [38] introduced improved K-means based on the k-medoids cyclic method and the improved triangle trilateral relations theorem to detect network intrusion detection. This cluster algorithm for network intrusion detection used KDDcup 99 datasets. The dataset was divided into 5 subsets and the average detection rate was obtained as 89.5%, whereas the false alarm rate was 4.896%, which is relatively very high.

Witcha et al. [39] proposed a fuzzy rough c-means clustering algorithm for detecting new intrusions so as to improve the detection rate and reduce false-positive rate in intrusion detection systems. The fuzzy clustering algorithm was used to predict normal and suspicious behavior. The performance of the algorithm was measured based on the detection rate, accuracy, false alarm rate, and correlation. The detection rate for this algorithm was 91.45, accuracy was 82.46, false alarm rate was 24.8, and correlation was 0.556.

Warusia et al. [11] also integrated K-means clustering and naïve bayes classification (KMC +NBC) for anomaly intrusion detection for the intrusion detection system. The algorithm was assessed using ISCX 2012 datasets. The dataset was classified as training and testing datasets and KMC+NBC algorithm proved to enhance the accuracy and detection rate and reduced false alarm rate of NBC. The accuracy was 99.8%, the detection rate was 95.4% and the false alarm rate of the above algorithm was 0.13 in terms of training data; whereas, in testing data, the accuracy was 99%, the detection rate was 98.8% and false alarm rate was 2.2%.

Vipin Kumar et al. [40] used NSL-KDD datasets to cluster the data into normal and four different types of attacks i.e., DoS, R2L, U2R, and probe. The experiment was carried out using WEKA software with clustering technique. Simple K-means algorithm was used to cluster the data into four different groups of attacks. The algorithm was proven to be very useful in detecting a large amount of unlabeled data, i.e., detecting new types of attacks.

Abhaya et al. [41] proposed an efficient intrusion detection method for detecting normal and abnormal instances based on the fuzzy c-means clustering technique and support vector machine. They used KDD99 datasets for this approach and compared the performance with K-means, SVM, K-means and naïve bayes, and FCM and naïve bayes. In this approach, the datasets are transferred to the clustering technique and it is transferred to the classification technique and finally, the performance is evaluated using accuracy. The accuracy of FCM

+SVM was 99.74% as compared to NB+FCM, SVM+k-means, NB+k-means which accounts to 95.63%, 99.37% and 95.32%.

Hari Om and Aritra Kundu [42] proposed hybrid intrusion detection by combining K-means and two classifiers, naïve bayes and k-nearest neighbor, to detect anomaly. This algorithm used KDD-cup99 datasets to detect the intrusion and classify them into four categories: DoS, U2R,R2l, and probe. The main objective of the proposed algorithm was to reduce the false alarm rate. The detection rate of the proposed approach was 99.35%, the false alarm rate was 1.394%, and accuracy was 98.20%.

Partha et al. [43] evaluated the data mining technique based on the fuzzy c-means clustering and K-means clustering technique over the NSL-KDD datasets to detect the four types of attacks. It was observed that only 45.95% of attacks were detected by fuzzy c-means clustering technique, whereas the detection rate is 44.72% using K-means clustering technique.

5. RESEARCH QUESTIONS

The main objective of the research is to analyze the clustering techniques used in the literature of the intrusion detection system over 10 years. To achieve this objective, the following questions were established:

1. What datasets have been used in IDS?
2. What clustering technique has been used in the intrusion detection system research?
3. What are the evaluation metrics used to measure the performance of clustering technique?
4. What are the strengths and weakness of clustering techniques that have been used in IDS?

6. RESULT AND DISCUSSION

The result for this research was analyzed using the research question that was initially posted. There were four questions that were to be answered, which we analyzed separately as following:

1. RQ1: What was the dataset that has been used in intrusion detection system?

There are different datasets being used by different learning algorithm such as classification, clustering, regression, etc. to detect the intrusions. In this survey, 30 papers were reviewed for clustering technique. The most common type of dataset that is being used for clustering technique is KDDcup99 datasets. The number of papers using KDD-99 dataset had peaked in year 2011.

Among 30 papers, there were 24 that used KDD-99 dataset, where Zhong et al. used KDD99 data along with Kyoto 2006+ datasets accounting to 80% of whole datasets. KDD-99 dataset is derived from DARPA dataset, this is one of the datasets that is used by researchers for

clustering algorithm. In addition, only 2 papers used NSL-KDD dataset and 2 papers used Kyoto 2006+ that account for 7% of the datasets used in this research as stated in figure 2. Similarly, only one paper each reviewed in this research used DARPA 98 datasets and ISCX 2012 datasets.

2). Q2: What clustering technique has been used in the intrusion detection system research?

One of the major questions about this research is to find out what is the most common clustering technique used in intrusion detection system research. Out of 30 papers studied, the most common clustering technique used for intrusion detection was K-means (it was used 14 times). This clustering technique was used as a single technique for three times while for others, it was used as a hybrid technique such as hierarchial, One each R, fuzzy and SVM, Naïve Bayes, PSO, fuzzy C-means, SOM and neutral gas, and C 4.5 decision tree. Besides these graph-based, II DBC, K-medoids, grid-based, SUM + hierarchical, CAAN, IDBGC, Random forest + weighted, and TANN were used.

3). RQ3: What are the evaluation metrics used to measure the performance of the clustering technique?

The clustering technique used in intrusion detection for computer networks can be assessed and evaluated using various types of evaluation metrics. The detection rate, false alarm rate, false-positive rate, false-negative rate were the most common evaluation metrics used to measure the performance of clustering technique in the clustering detection literature. All the 30 papers reviewed used these metrics for detecting the performance of the clustering technique in the intrusion detection system. However, Leung et al. [13] and Zhiengje et al. [22] also used ROC curve while evaluating the techniques; and similarly, Guan et al. [16] also considered sum of square error along with detection rate and false-positive rate. Chandrashekar et al. [20] also looked at specificity and sensitivity for the datasets in addition to TPR, FPR, DR, FPR and FNR. The performance of the clustering technique is considered high if the detection rate, accuracy is high and the false alarm rate is low. These evaluation matrices were also used to compare the algorithms with the already existed algorithm.

4) RQ4) What are the strengths and weakness of clustering techniques that have been used in IDS?

After reviewing 30 papers in this research, we also evaluated the strengths and weakness of the clustering techniques used. The common techniques used by different research were combined and analyzed for strengths and weakness. The results were then incorporated as shown in table 2. From table 2, it can seen that each and every clustering technique has its own pros and cons in analyzing intrusion detection.

7. CONCLUSION

The intrusion detection system is a way of analyzing the network traffic so that unwanted packets or any malicious activities on the system are detected and prevented. In this paper, 30 papers of clustering techniques were studied and evaluated for intrusion detection. It was found that 27 different types of clustering techniques were applied in order to detect

intrusions. The most common clustering technique used for intrusion among the 30 papers was K-means clustering technique.

Apart from using a single clustering technique, there was a hybrid clustering technique that was combined with different classification techniques for detecting intrusions in a system. These clustering techniques were used in different datasets. However the most common datasets were found to be KDD-99 datasets. The KDDCup99 datasets comprise of millions of connections with different features and different types of attacks. Similarly, the performance of this technique was evaluated by different researchers using different evaluation metrics. Detection rate and False

Alarm rate was commonly used as evaluation metrics by most of the researchers to check the effectiveness of their procedures for intrusion detection. Similarly, in this research, the strength and weakness of different technologies used for clustering attacks were also analyzed.

As future work, we are researching on the effectiveness of user-oriented clustering to enhance the effectiveness of an Intrusion Detection System. In this method, clusters of activities in the log files are formed based on an expert feedback of whether or not the activities are an intrusion. The cluster is defined based on the activities it contains. A future activity is defined as intrusion or not based on its similarity to a cluster.

ACKNOWLEDGEMENTS

This research was supported in part by National Science Foundation grants #1761735, #1723586, #1663350. National Institutes of Health grant NIH TU CBR/RCMI #U54MD007585, and a grant from Rockwell Collins, USA

AUTHORS

Binita Bohara is a graduate student of Information System and Security Management at Tuskegee University. She completed her Accounting degree from Australia. She is currently working on a research on clustering techniques for intrusion detection for network security under professor Dr. Jay Bhuyan. Currently, she is also working as a teaching assistant for undergraduate students.

Jay Bhuyan is a professor in the Department of Computer Science at Tuskegee University. His research and teaching interests include Telecom Software Architecture and Development, Software and Network Security, Big Data Analytics, and Machine Learning.

He received a Ph.D. in Computer Science from the University of Louisiana at Lafayette. He has over 25 years of full-time and part-time teaching experience as well as over 15 years of research and development experience in the Telecom industry. He is a member of ACM, IEEE and IEEE Computer, IEEE Communications Societies.

Fan Wu is a professor of Computer Science and the director of Tuskegee University Center of Academic Excellence in Information Assurance Education. He received his Ph.D. degree in Computer Science from Worcester Polytechnic Institute (WPI) in 2008. His teaching and research interests lie primarily in the area of Information Assurance, Software Security, Mobile Security, High Performance Computing and Artificial Intelligence. He has led several cybersecurity related projects funded by NSF, DHS, and DoD. He has served as the editor-in-chief of the International Journal of Mobile Devices, Wearable Technology, and Flexible Electronics (IJMDWTFE). He has involved students in research and published a number of research papers in cybersecurity, software security, mobile security, and information assurance.

Junhua Ding is a Professor of Data Science and the Director of Data Science program in the department of Information Science at the University of North Texas. His research interests include: data analytics, machine learning, software security and software engineering. He received a Ph.D., an M.S., and a B.S. in Computer Science from Florida International University, Nanjing University, and China University of Geo- sciences, respectively.

REFERENCES

[1]. Revathi S and Malathi A, “A detailed analysis on nsl-kdd dataset using various machine learning techniques for intrusion detection,” International Journal of Engineering Research & Technology (IJERT), vol. 2, no. 12, pp. 1848–1853, 2013.

[2]. Salo F, Injadat M, Nassif AB, Shami A, and Essex A, “Data mining techniques in intrusion detection systems: A systematic literature review,” IEEE Access, vol. 6, pp. 56 046–56 058, 2018.

[3]. Siddiqui MK and Naahid S, “Analysis of kdd cup 99 dataset using clustering based data mining,” International Journal of Database Theory and Application, vol. 6, no. 5, pp. 23–34, 2013.

[4]. Mingqiang Z, Hui H, and Qian W, “A graph-based clustering algorithm for anomaly intrusion detection,” in 2012 7th International Conference on Computer Science & Education (ICCSE). IEEE, 2012, pp. 1311–1314.

[5]. Jianliang M, Haikun S, and Ling B, “The application on intrusion detection based on k-means cluster algorithm,” in 2009 International Forum on Information Technology and Applications, vol. 1. IEEE, 2009, pp. 150–152.

[6]. Jian Y, "An improved intrusion detection algorithm based on dbscan [j]," Microcomputer Information, vol. 25, no. 13, 2009.

[7]. Cup K, "Dataset," available at the following website http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, vol. 72, 1999.

[8]. Lippmann RP, Fried DJ, Graf I, Haines JW, Kendall KR, McClung D, Weber D, Webster SE, Wyschogrod D, Cunningham RK et al., "Evaluating intrusion detection systems: The 1998 darpa off-line intrusion detection evaluation," in Proceedings DARPA Information Survivability Conference and Exposition. DISCEX'00, vol. 2. IEEE, 2000, pp. 12–26.

[9]. Tavallaee M, Bagheri E, Lu W, and 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, pp. 1–6.

[10]. Protić DD, "Review of kdd cup'99, nsl-kdd and kyoto 2006+ datasets," Vojnotehnički glasnik, vol. 66, no. 3, pp. 580–596, 2018.

[11]. Yassin W, Udzir NI, Muda Z, Sulaiman MN et al., "Anomaly-based intrusion detection through k-means clustering and naives bayes classification," in Proc. 4th Int. Conf. Comput. Informatics, ICOCI, no. 49, 2013, pp. 298–303.

[12]. Ranjan R and Sahoo G, "A new clustering approach for anomaly intrusion detection," arXiv preprint arXiv:1404.2772, 2014.

[13]. Leung K and Leckie C, "Unsupervised anomaly detection in network intrusion detection using clusters," in Proceedings of the Twenty-eighth Australasian conference on Computer Science-Volume 38. Australian Computer Society, Inc., 2005, pp. 333–342.

[14]. Portnoy L, "Intrusion detection with unlabeled data using clustering," Ph.D. dissertation, Columbia University, 2000.

[15]. Münz G, Li S, and Carle G, "Traffic anomaly detection using k-means clustering," in GI/ITG Workshop MMBnet, 2007, pp. 13–14.

[16]. Guan Y, Ghorbani AA, and Belacel N, "Y-means: A clustering method for intrusion detection," in CCECE 2003-Canadian Conference on Electrical and Computer Engineering. Toward a Caring and Humane Technology (Cat. No. 03CH37436), vol. 2. IEEE, 2003, pp. 1083–1086.

[17]. Jiang W, Yao M, and Yan J, "Intrusion detection based on improved fuzzy c-means algorithm," in 2008 International Symposium on Information Science and Engineering, vol. 2. IEEE, 2008, pp. 326–329.

[18]. Xue-Yong L, Guo-hong G, and Jia-Xia S, "A new intrusion detection method based on improved dbscan," in 2010 WASE International Conference on Information Engineering, vol. 2. IEEE, 2010, pp. 117–120.

[19]. Muda Z, Yassin W, Sulaiman MN, and Udzir NI, "Intrusion detection based on k-means clustering and oner classification," in 2011 7th International Conference on Information Assurance and Security (IAS). IEEE, 2011, pp. 192–197.

[20]. Chandrasekhar A and Raghuveer K, "Intrusion detection technique by using k-means, fuzzy neural network and svm classifiers," in 2013 International Conference on Computer Communication and Informatics. IEEE, 2013, pp. 1–7.

[21]. Chitrakar R and Huang C, "Anomaly based intrusion detection using hybrid learning approach of combining k-medoids clustering and naive bayes classification," in 2012 8th International Conference on Wireless Communications, Networking and Mobile Computing. IEEE, 2012, pp. 1–5.

[22]. Li Z, Li Y, and Xu L, "Anomaly intrusion detection method based on k-means clustering algorithm with particle swarm optimization," in 2011 International Conference of Information Technology, Computer Engineering and Management Sciences, vol. 2. IEEE, 2011, pp. 157–161.

[23]. Xiao L, Shao Z, and Liu G, "K-means algorithm based on particle swarm optimization algorithm for anomaly intrusion detection," in 2006 6th World Congress on Intelligent Control and Automation, vol. 2. IEEE, 2006, pp. 5854–5858.

[24]. Fatma H and Mohamed L, "A two-stage technique to improve intrusion detection systems based on data mining algorithms," in 2013 5th International Conference on Modeling, Simulation and Applied Optimization (ICMSAO). IEEE, 2013, pp. 1–6.

[25]. Zhong Y, Yamaki H, and Takakura H, “A grid-based clustering for low-overhead anomaly intrusion detection,” in 2011 5th International Conference on Network and System Security. IEEE, 2011, pp. 17–24.

[26]. Horng S-J, Su M-Y, Chen Y-H, Kao T-W, Chen R-J, Lai J-L, and Perkasa CD, “A novel intrusion detection system based on hierarchical clustering and support vector machines,” Expert systems with Applications, vol. 38, no. 1, pp. 306–313, 2011.

[27]. Toosi AN and Kahani M, “A new approach to intrusion detection based on an evolutionary soft computing model using neuro-fuzzy classifiers,” Computer communications, vol. 30, no. 10, pp. 2201–2212, 2007.

[28]. Pfahringer B, “Winning the kdd99 classification cup: bagged boosting,” SIGKDD explorations, vol. 1, no. 2, pp. 65–66, 2000.

[29]. Levin I, “Kdd-99 classifier learning contest: Llsoft’s results overview,” SIGKDD explorations, vol. 1, no. 2, pp. 67–75, 2000.

[30]. Sabhnani M and Serpen G, “Application of machine learning algorithms to kdd intrusion detection dataset within misuse detection context.” in MLMTA, 2003, pp. 209–215.

[31]. Xuren W, Famei H, and Rongsheng X, “Modeling intrusion detection system by discovering association rule in rough set theory framework,” in 2006 International Conference on Computational Inteligence for Modelling Control and Automation and International Conference on Intelligent Agents Web Technologies and International Commerce (CIMCA’06). IEEE, 2006, pp. 24–24.

[32]. Lin W-C, Ke S-W, and Tsai C-F, “Cann: An intrusion detection system based on combining cluster centers and nearest neighbors,” Knowledge-based systems, vol. 78, pp. 13–21, 2015.

[33]. Liu Y, Chen K, Liao X, and Zhang W, “A genetic clustering method for intrusion detection,” Pattern Recognition, vol. 37, no. 5, pp. 927–942, 2004.

[34]. Sharma SK, Pandey P, Tiwari SK, and Sisodia MS, “An improved network intrusion detection technique based on k-means clustering via naïve bayes classification,” in IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM-2012). IEEE, 2012, pp. 417–422.

[35]. Muniyandi AP, Rajeswari R, and Rajaram R, “Network anomaly detection by cascading k-means clustering and c4. 5 decision tree algorithm,” Procedia Engineering, vol. 30, pp. 174–182, 2012.

[36]. Elbasiony RM, Sallam EA, Eltobely TE, and Fahmy MM, “A hybrid network intrusion detection framework based on random forests and weighted k-means,” Ain Shams Engineering Journal, vol. 4, no. 4, pp. 753–762, 2013.

[37]. Tsai C-F and Lin C-Y, “A triangle area based nearest neighbors approach to intrusion detection,” Pattern recognition, vol. 43, no. 1, pp. 222–229, 2010.

[38]. Tian L and Jianwen W, “Research on network intrusion detection system based on improved k-means clustering algorithm,” in 2009 International Forum on Computer Science-Technology and Applications, vol. 1. IEEE, 2009, pp. 76–79.

[39]. Chimphlee W, Abdullah AH, Sap MNM, Srinoy S, and Chimphlee S, “Anomaly-based intrusion detection using fuzzy rough clustering,” in 2006 International Conference on Hybrid Information Technology, vol. 1. IEEE, 2006, pp. 329–334.

[40]. Kumar V, Chauhan H, and Panwar D, “K-means clustering approach to analyze nsl-kdd intrusion detection dataset,” International Journal of Soft Computing and Engineering (IJSCE), 2013.

[41]. Kumar K et al., “An efficient network intrusion detection system based on fuzzy c-means and support vector machine,” in 2016 International Conference on Computer, Electrical & Communication Engineering (ICCECE). IEEE, 2016, pp. 1–6.

[42]. Om H and Kundu A, “A hybrid system for reducing the false alarm rate of anomaly intrusion detection system,” in 2012 1st International Conference on Recent Advances in Information Technology (RAIT). IEEE, 2012, pp. 131–136.

[43]. Bhattacharjee PS, Fujail AKM, and Begum SA, “A comparison of intrusion detection by k-means and fuzzy c-means clustering algorithm over the nsl-kdd dataset,” in 2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC). IEEE, 2017, pp. 1–6.

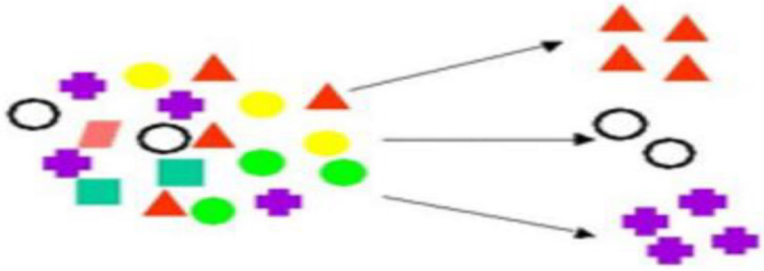

Figure 1:
A schematic of clustering process showing three clusters (one red triangle, one black circle, and one pink plus cluster)

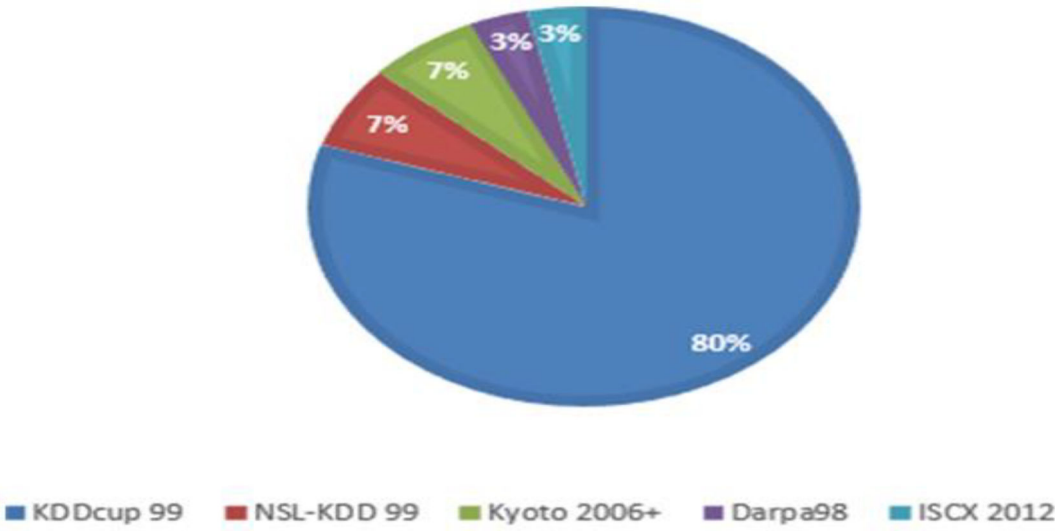

Figure 2:
A schematic view of datasets used in clustering technique

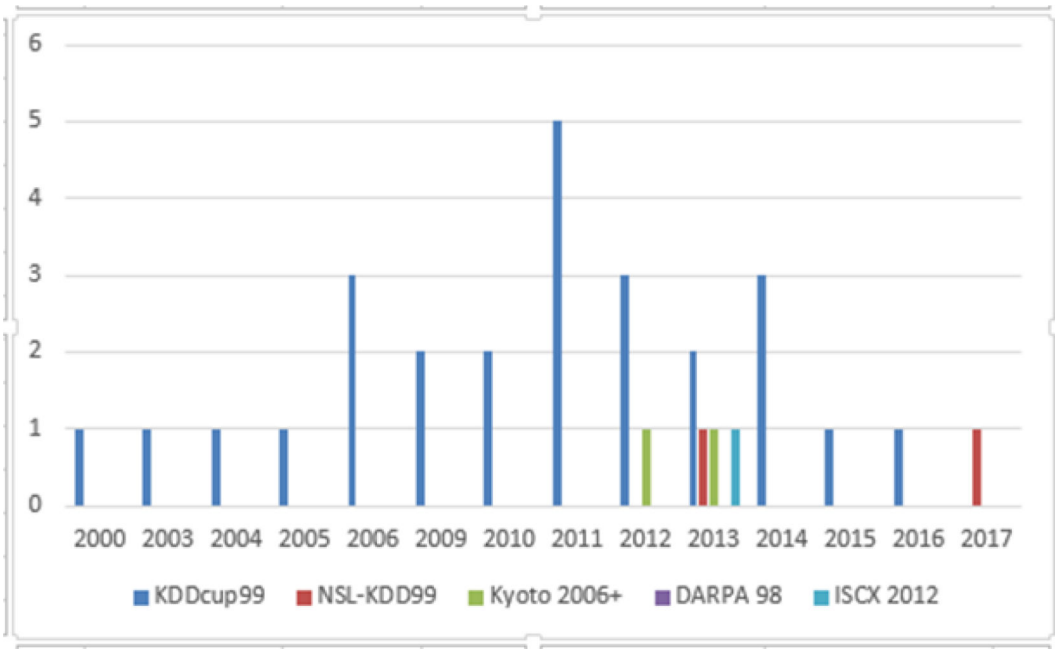

Figure 3:
A schematic view of distribution of datasets over the years

Table 1:

Different types of Attack from Kddcup 99 Datasets [1]

DOS	Probe	R2L	U2R
Back	Nmap	Spy	Buffer overflow
Land	Portsweep	Phf	Rootkit
Neptune	Ipsweep	Multihop	Loadmodule
Pod	Satan	ftp write	Perl
Smurf		Imap	
Teardrop		Warezmaster	
		Guess passwd	

Table 2:

Strengths and Weakness of Clustering Technique used in intrusion detection literature

Clustering Technique	Strength	Weakness	Ref
K-means	DR was high and FAR was below 4% and time complexity was low	Cluster number needs to be defined	Meng et al., Gerhard et al., Vipin et al
Hierarchical clustering overcame the shortcoming of K-means clustering to predict the number of clusters	Determining the cluster width manually, there was a chance of mislabeling the normal instance as an abnormal and vice-versa if W was not determined properly	Leonid et al.	
Y-means	Number of cluster dependency and dengenracy of K-means was overcome	High false positive compared to other algorithm	Guan et al.
Graph-based	Identifies clusters of any shape and it only uses a parameter and does not require to define anyclus- ter number	Computation increased as number of records in- creased	Zhou et al
k-medoids	Has advantages over the existing algorithm such as dependency on initial centroids, cluster number, and irrelevant clusters	The detection rate for the proposed algorithm was low for probing at- tack(70.51) and user to root attack(70.13)	Ravi et al.
IFCA	Introduce the function of validity for choosing number of clustering		Wei et al.
IIDBC	Detection rate was high and the performance of DBSCAN was improved	Selection of Parameters	Li-XUE et al.
Grid-based	The performance was insensitive to the variation of the convergence crite- rion of clustering, attack and normal condition	Low detection rate and high false positive	Zhong et al.
CANN	Detection rate was high for 6-dimensional KDD datasets	Misclassified U2R and R2L as normal in case of 6-dimensional KDD datasets	Lin et al.
TANN	The accuracy rate, detection rate were high		Chih-Fong et al.
Improved K-means		False alarm rate was relatively very high	Li Tian et al.
Fuzzy C-means	Achieved good performance compared to Kmeans methods		Witcha et al
Density+Grid based	High detection rate	High false positive rate	Leung et al.
K-means+One R classifica- tion	Detection rate above 99.0 and a false alarm rate be- low 2.75 and the perfor- mance of the hybrid clas- sifiers was higher as com- pared to the single classi- fier	Couldn’tclassify U2R and R2l attack	Z.muda et al.
Kmedoids+Naive Bayes	Showed better performance as compared to K-means and Naive Bayes hybrid algorithm		Chitrakar et al.
PSO+K-means	Accuracy was very good for U2R and DoS	Low detection rate and high false positive	Lizhong et al., Zhiengje et al.
SVM+HC	Detection rate was high for all four types of attack		Horng et al.
K-means+Naive Bayes	The algorithm was shown to be efficient in detecting network intrusion	This approach had a high false-positive rate	Sanjay etal Warsula et al.
K-means+C4.5	The proposed algorithm gave notable detection rate		Amuthan et al.
Random Forest +Weighted K-means		High false-positive rate	Reda et al.
Fuzzy Cmeans+Svm	The accuracy value was slightly better than ex-isting K-Medoids+SVM and it was far better than		Abhaya et al.

Clustering Technique	Strength	Weakness	Ref
	SVM alone and it was shown to be stable over other		
SOM+K-means+Fuzzy Cmean	Showed to outperform other competitor method by reducing FAR		Fatma et al
Kmeans + Fuzzy + SVM	Effective for low-frequent attacks such as U2R and R2L		Chandrashekar et al.
K-means + KNN + NB		May sometime misclassify the records	Hari Om et al.
Fuzzy Cmeans+K-means		Very low detection rate	Partha et al.