ANIDS: Anomaly Network Intrusion Detection System Using Hierarchical Clustering Technique

Sunil M. Sangve and Ravindra C. Thool

Abstract The Intrusion detection system (IDS) is an important tool to detect the unauthorized use of computer network and to provide the security for information. The IDS consists of two types signature-based (S-IDS) and anomaly-based (A-IDS) detection system. S-IDS detect only known attacks whereas A-IDSs are capable to detect unknown attacks. In this paper, we are focusing on A-IDS. The proposed system is Anomaly network intrusion detection system (ANIDS). The ANIDS is implemented using metaheuristic method, genetic algorithm and clustering techniques. The two different clustering techniques are used i.e. *K*-mean clustering and hierarchical clustering to check the performance of system in terms of false positive rate (FPR) and detector generation time (DGT). The system includes modules like input dataset, preprocessing on input dataset, clustering and selection of sample training dataset, testing dataset, and performance analysis using training and testing dataset. The experimental results are calculated based on large–scale dataset, i.e., NSL-KDD for detector generation time and false positive rate (FPR). Our proposed technique gives better result for false positive rate and detector generation time as compared to *K*-means clustering technique.

Keywords Intrusion detection system (IDS) · Anomaly network intrusion detection system (ANIDS) · NSL-KDD

S.M. Sangve (✉)
ZCOER, Computer Engineering, SP Pune University, Pune, Maharashtara, India
e-mail: sunilsangve@gmail.com

R.C. Thool
SGGIET Computer Science and Engineering, Nanded, Maharashtara, India
e-mail: rcthool@yahoo.com

S.C. Satapathy et al. (eds.), *Proceedings of the International Conference on Data Engineering and Communication Technology*, Advances in Intelligent Systems and Computing 468, DOI 10.1007/978-981-10-1675-2_14

1 Introduction

Now-a-days, dependency on networked computers increased and is still increasing, along with this, growing expertise in such networked system requires brilliant and adaptive threat detection. Because of this computer network security becomes a major issue. Thus, in computer security, confidentiality, integrity, and availability (CIA) plays a very vital role [1]. To identify improper or unauthorized modifications, the integrity mechanism has divided into two classes: prevention and detection [2]. Therefore, for detection of attacks, intrusion detection system (IDS) is used and for prevention of unauthorized user, intrusion detection and prevention system (IDPS) is used.

The problem of identifying the intrusion in the system is resolved by checking violation of privilege levels in the system, misuse of the system and unauthorized use. The heterogeneous computer network gives additional burden to detect the intrusions [3]. Originally, the concept of intrusion detection was proposed by Anderson in 1980 [4]. Basically, there are two types of IDS Host-based Intrusion Detection System (H-IDS) and Network-based Intrusion Detection System (N-IDS). The H-IDS detects the intrusions on the single system but N-IDS detect attacks on multiple systems by connecting the systems with each other by a network. In this paper, we are focusing on network-based intrusion detection system. The N-IDS consists of two types signature-based N-IDS and anomaly-based N-IDS whereas Signature-based N-IDS used to detect only known attack and anomaly-based N-IDS used to detect unknown attack [5].

This paper represents the network anomaly detection using metaheuristic method including genetic algorithm and clustering techniques. The metaheuristic method defined by Osman and Laporte in 1996, is an iterative generation process which gives guidance to subordinate heuristic by combining concepts for exploring and exploiting the search space, as well as learning strategies are used to find efficiently near optimal solutions [6]. Blum and Roli [7], gives fundamental properties of metaheuristic: (1) to guide search process, metaheuristic strategies are used. (2) Explore the search space to find near optimal solution.

2 Related Work

The anomaly detection methods are classified into several types. One of the methods among them is statistic-based method. It identifies the intrusion by using the predefined threshold, standard deviation, mean, and the probabilities [8]. Another category is rule-based methods. It uses the if-then and if-else rules, in order to construct the model of detection for some previously known intrusions [9]. Additionally, the State-Based approach is also there. It makes the use of Finite state machine, which is derived from the network topologies to determine the attacks [10]. Negative selection algorithm (NSA) is one of the artificial immune system (AIS) algorithms which

motivated by immune system microorganism development and tolerance toward oneself in human immune system [11]. It builds a model of non-self information by producing examples that didn't match existing ordinary (self) designs, then utilizing this model to match non-ordinary examples to recognize anomalies. NSA detectors are structured with different geometric shapes, for example, hyper-rectangles, hyper-circles, hyper-ellipsoids or various hyper-shapes. Anna Sperotto et al. [12], proposed the automatic approach for anomaly network intrusion detection using SSH (secure shell is the encrypted protocol which allows to operate remotely over an unsecured network) traffic. They suggested the procedure which selects the system parameter automatically and increases the system performance. Alexander et al. [13], have considered the problem of online anomaly detection in computer network traffic effectively. This is done using changepoint detection method.

3 Implementation Details

The main idea is based on combination of multi-start metaheuristic algorithm, genetic algorithm, and hierarchical clustering technique. The number of detectors is very important to detect anomaly. In ANIDS, we are using hierarchical clustering technique to reduce FPR and DGT and compare the results with existing *k*-mean clustering. The clustering techniques are used to select multiple initial points using multi-start method. Using multi-start method, the radius of hyper sphere detector is obtained. This radius is optimized using genetic algorithm. The rule reduction is used to remove redundant detectors to reduce detector generation time. The detector generation process is repeated to increase the detection quality. As shown in Fig. 1, we use the hierarchical clustering and *K*-mean clustering to divide the large training dataset into number of clusters. The Anomaly Network Intrusion Detection System consists of following modules:

3.1 Input Dataset

The input dataset is NSL-KDD dataset [14]. It contains Normal, Probe, U2R, R2L, and DoS attacks. The total 41 columns headers are added that contain information such as duration, protocol type, service, src_bytes, dst_bytes, flag, land, wrong fragment, etc.

3.2 Data Preprocessing

Preprocessing is applied on input dataset (I). To remove unnecessary data or words which are not useful for extracting the features, data preprocessing is used. The

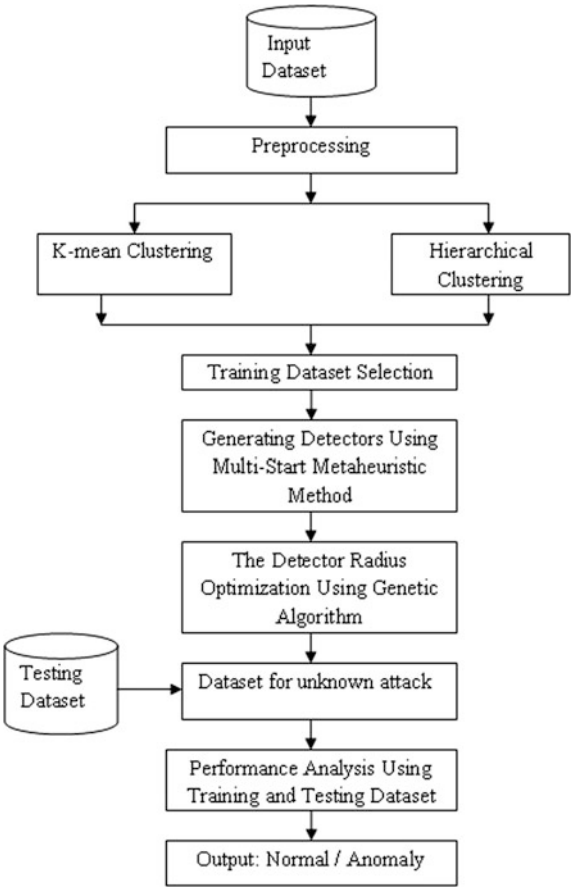

Fig. 1 Anomaly network intrusion detection system (ANIDS)

main benefit of data preprocessing is that the time required for processing will also decrease. The following example describes how preprocessing applied on input dataset:

Let us consider the one sample from I,

{0, tcp, ftp_data, SF, 491, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 2, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00, 0.00, 150, 25, 0.17, 0.03, 0.17, 0.00, 0.00, 0.00, 0.05, 0.00, normal}

When we apply the preprocessing on the above single sample, the words like tcp, ftp_data, SF (start flag) are removed to decrease the processing time. The preprocessed sample consists of numeric. The last word in sample denotes the class normal or anomaly. Therefore, the obtained vector contains two important features, i.e., pattern in numeric form and class name 'normal'.

{491, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 2, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00, 0.00, 150, 25, 0.17, 0.03, 0.17, 0.00, 0.00, 0.00, 0.05, 0.00}

3.3 Clustering

3.3.1 Hierarchical Clustering Algorithm

Given a set of N items to be clustered, and an $N * N$ matrix (distance or similarity), the main idea about hierarchical clustering is defined by Johnson [15].

3.3.2 Pseudo Code

The pseudocode for hierarchical clustering is given in [16] which are given below: First, we compute the $N * N$ similarity matrix. The algorithm executes $N - 1$ steps to merge the most similar cluster.

Hierarchical Clustering (d_1, d_2, d_3, d_4,….d_N)

1. For $n \leftarrow 1$ to N
2. Do for I $\leftarrow 1$ to N
3. Do $C[n]$ $[i] \leftarrow$ SIM (d_n, d_i)
4. $I[n] \leftarrow 1$ (used to keeps track of active clusters)
5. $A \leftarrow [\cdot]$ (assembles clustering as a sequence of merges
6. For $k \leftarrow 1$ to $N - 1$
7. Do $(I, m) \leftarrow$ arg max{$<i, m>$: $i \neq m \wedge I[i] = 1 \wedge$ I $[m] = 1\}$ $C[i]$ $[m]$
8. A. Append ($<i, m>$) (store merge)
9. For $j \leftarrow 1$ to N
10. Do $C[i]$ $[j] \leftarrow$ SIM (i, m, j)
11. $C[j]$ $[m] \leftarrow$ SIM (i, m, j)
12. I $[m] \leftarrow 0$ (Deactivate cluster)
13. Return A;

SIM (i, m, j)—used to compute similarity of cluster j with i and m cluster which are merged together. It is function of $C[i]$ $[j]$ and $C[j]$ $[m]$. The time complexity of Hierarchical Clustering is $\Theta(N3)$, here we scan $N * N$ matrix C with largest similarity in each of $N - 1$.

3.4 Selection of Dataset and Detector Generation Using Multi-start Metaheuristic Method

After applying the clustering algorithm, we select some training dataset samples from given dataset. The multiple initial start points are selected from clustering as input to generate the detectors. The detector shape is hyper sphere. Thus, we calculate the radius of hyper sphere to identify the anomaly by using following rules [17]:

The detector radius $R = \{r \in R | 0 < r \leq \text{hpu}\}$ where hpu is the hyper-sphere radius upper bound. Thus,

$U_j = \max(x_{ij})$ where $i = 1, 2, 3 \ldots m$, $L_j = \min(x_{ij})$ where $i = 1, 2, 3 \ldots, m$

UB—upper bound and LB—lower bound are used for solution space.

UB = (u_1, u_2, $u_3 \ldots, u_n$, hpu), LB = (I_1, I_2, $I_3 \ldots$ 0), the detectors $D = \{d_1, d_2, d_3 \ldots d_{\text{isp}}\}$

The solution space obtained by multi-start framework is calculated as: $D_i = (u_{i1}, u_{i2}, u_{i3} \ldots u_{in}, r_i)$ where hyper-sphere center is at $D_{\text{center}} = (u_{i1}, u_{i2}, u_{i3} \ldots u_{in})$ and hyper sphere radius is r_i.

The objective function to control the detector generation process is:

$$F(D_i) = N_{\text{abnormal}}(d_i) - N_{\text{normal}}(d_i) \tag{1}$$

where, N_{abnormal} is the number of abnormal samples covered by detector d_i and N_{normal} is the number of normal samples covered by detector d_i.

Anomaly Detection is done from the generated detectors and rule is, If (dist (D_{center}, x) $\leq r$) then {normal} else {abnormal} where r is the detector hyper-sphere radius and (dist (D_{center}, x)) is the Euclidean distance between detector hyper sphere center D_{center} and test samples x.

3.5 Testing Dataset

The testing dataset is the additional dataset to detect unknown pattern. It is used to test training patterns in the training dataset, If the system was trained successfully, outputs produced by the system would be similar the actual outputs.

3.6 Performance Analysis Using Training and Testing Dataset

Finally performance of the system is analyzed using training and testing dataset for false positive rate, detector generation time.

4 Experimental Results

The experimental results are calculated based on NSL-KDD dataset. For experimental set up, we use Windows 7 operating system, Intel i5 processor, 4 GB RAM, 80 GB Hard disk, Net Beans IDE 8 + JDK tool. The False positive rate is calculated using formula:

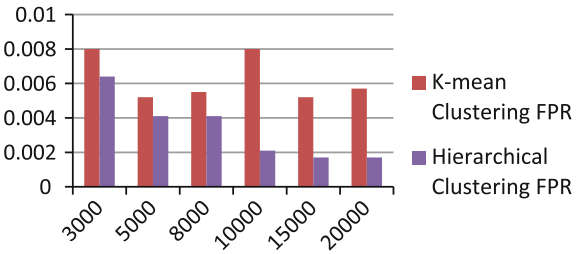

Fig. 2 False positive rate graph

The False positive rate (FPR):

$$\text{FPR} = \frac{\text{False Positive}}{\text{FP} + \text{FN}} \tag{2}$$

where, FP-False Positive, FN-false Negative.

4.1 False Positive Rate (FPR)

The minimum false positive rate is obtained for the anomaly network intrusion detection system using hierarchical clustering technique. Therefore, the results show that, Hierarchical clustering gives minimum FPR than the *K*-mean clustering. The minimum FPR is 0.0017 obtained at training dataset 15000. Fig. 2 shows comparison of false positive rate using *K*-mean and hierarchical clustering approach (Table 1).

4.2 Detector Generation

Fig. 3 shows the comparison of DGT by using *K*-mean and Hierarchical-clustering algorithm. The time required for generation of detector is less by Hierarchical-clustering. The observation is that detector generation time increases with the increase in dataset size (Table 2).

Table 1 False positive rate (FPR)

Dataset size	FPR using *K*-mean	FPR using hierarchical clustering
3000	0.008	0.0064
5000	0.0052	0.0041
8000	0.0055	0.0041
10000	0.008	0.0021
15000	0.0052	0.0017
20000	0.0057	0.0017

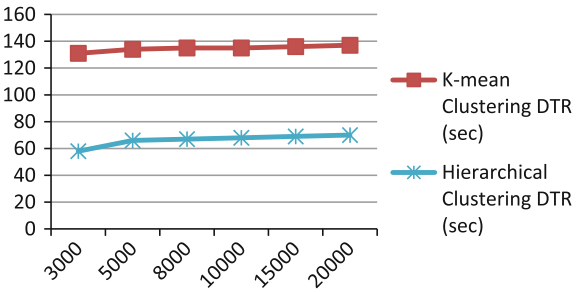

Fig. 3 Detector generation time graph

Table 2 Detector generation time

Dataset size	FPR using *K*-mean (s)	FPR using hierarchical clustering (s)
3000	131	58
5000	134	66
8000	135	67
10000	135	68
15000	136	69
20000	134	70

5 Conclusion

The anomaly-based network intrusion detection system is implemented using metaheuristic method, clustering techniques and Genetic algorithm. The two algorithms *K*-mean clustering and hierarchical clustering are used to check the performance for detection accuracy and false positive rate. The proposed approach use hierarchical clustering to reduce false positive rate. The clustering techniques are used to divide training dataset to reduce time and processing complexity. The metaheuristic method with evolutionary algorithm, i.e., genetic algorithm plays an important role to select multiple initial start points, to generate number of detectors, to calculate the radius limit of hypersphere detector and to remove redundant detectors to give final output, i.e., anomaly or normal. The experimental results are calculated using NSL-KDD dataset. Using hierarchical clustering we have obtained false positive rate and detector generation time 0.0017 and 69 s, respectively, for training dataset of size 15000. The benefit of hierarchical clustering is that it gives minimum false positive rate and detector generation time as compared to *k*-mean clustering. In future, the results will be calculated on other dataset like online to reduce false positive rate and detector generation time.

References

1. Morteza Amini, Rasool Jalili, Hamid Reza Shahriari. RT-UNNID: A practical solution to real-time network-based intrusion detection using unsupervised neural networks. Computers & security 25 (2006) 459–468.
2. Bishop M. Computer security, art and science. Addison-Wesley; 2003
3. James Brentano, Steven R Snapp et al. Architecture for Distributed Intrusion Detection. Division of computer science, University of California, 1991.
4. J.P. Anderson. Computer security threat monitoring and surveillance.Technical Report, James P. Anderson Co., Fort Washington, PA, April 1980
5. Tamer F. Ghanem, Wail S. Elkilani, Hatem. A hybrid approach for efficient anomaly detection using metaheuristic methods. Journal of advanced research, volume 6,issue 4 (2014) 609–619.
6. Osman, I.H., and Laporte, G. Metaheuristics bibliography. *Ann. Oper. Res. 63*, 513–623, 1996.
7. Blum, C., and Andrea R. Metaheuristics in Combinatorial Optimization: Overview and Conceptual Comparison. *ACM Computing Surveys*, *35(3),* 268–308, 2003.
8. Xu X. Sequential anomaly detection based on temporal difference learning: principles models and case studies. Applied Soft Computing 2010.
9. Kartit A, Saidi A, Bezzazi F, El Marraki M, Radi A. A new approach to intrusion detection system. JATIT 2012.
10. Garcia-Teodoro P, Diaz-Verdejo J, Macia -Fernandez G, Vazquez E. Anomaly-based network intrusion detection: techniques, systems and challenges. Computer Security, volume 24, Issue 1–2, (2009) 18–28.
11. Forrest S, Perelson AS, Allen L, Cherukuri R. Self- NonSelf discrimination in a computer. In: Proceedings of the 1994 IEEE symposium on security and privacy; Oakland, USA: IEEE Computer Society; 1994.
12. Anna Sperotto, Michel Mandjes, RaminSadre, Pieter-Tjerk de Boer, and AikoPras. Autonomic Parameter Tuning of Anomaly-Based IDSs: an SSH Case Study. IEEE Transactions On Network And Service Management, Vol. 9, No. 2, June 2012.
13. Alexander G. Tartakovsky, Senior Member, IEEE, Aleksey S. Polunchenko, and Grigory Sokolov. Efficient Computer Network Anomaly Detection by Changepoint Detection Methods. IEEE Journal Of Selected Topics In Signal Processing, Vol. 7, No. 1, February 2013.
14. The NSL-KDD dataset. The available World Wide Web is http://nsl.cs.unb.ca/NSL-KDD/
15. S. C. Johnson (1967). Hierarchical Clustering Schemes. *Psychometrika*, 2:241–254
16. Chapter 17, Hierarchical Clustering, DRAFT!© April 1, 2009 Cambridge University Press
17. Tamer F.Ghanem,Wail S. Elkilani, Hatem. A hybrid approach for efficient anomaly detection using metaheuristic methods. Journal of advanced research, 2014.