Enhancing The Efficiency of Detecting Intrusions Using HDEGMM Algorithm

Mahipal Singh Yadav[1], Dr. Nirupma Tiwari[2]
[1]Dept of computer Science & Engineering
[2]Asst. Professor, Dept of computer Science & Engineering
[1, 2] ShriRam College of Engineering & Management,Banmore

Abstract- *Network safety was one of the greatest computer network management problems, and security threats became the most widely publicized. Intrusion prevention has been a significant area of network security over recent years. When each attack class is handled as a different problem and controlled with advanced algorithms, IDSs yield better performance. A variety of surveys indicate that the intrusion into the network in the last few years has been gradually growing and that it has contributed to personal data theft. Network interruption is an unauthorized computer network operation. An efficient intrusion detection system must be in operation. In this paper, we know how to detect intrusions using the Gaussian model of mixture optimization Hybrid Differential Evolution optimization Gaussian mixture model (HDEGMM) . This paper is contrasted to an IGKM system for intrusion detection using a KDD-99 dataset HDEGMM algorithm. The experiment reveals that HDEGMM algorithms have better protection for intrusion detection than IGKM.*

Keywords- Intrusion detection system, IGKM and HDEGMM,data mining,KDD Cupp 99.

I. INTRODUCTION

The IDSs (Intrusion detection systems) are devices attached to the security wall to prevent the activity of the malicious system. Systems for intrusion detection. Most of this is because they can detect the most complex spectrum of attacks compared to other IDSs.Network IDSs analyze ongoing and incoming network attacks. Currently, commercial IDSs are mainly used to detect attacks on networks or host computers using a rules database called signatures. Intrusion detection systemsare device or network intrusion monitored.
The activity described by Christopher Kruegel et al. as a sequence of activities carrying by a malicious adversary leading to failure of the target system is an Intrusions is unauthorized and anomalous activity.[1] An IDS is an essential method for network administrators because it is not easy to examine a large number of travel packets second without a computer.The field is still open for further studies on the accuracy of detection, particularly after more than 30 years of intensive research on intrusion detection systems. Moreover, in versions of established attacks or new ones, the device is sometimes used without being detected.

The IDS goals layout the IDS policy requirements.

Potential goals include:

- Enforcement of connection policies
- Prevention of attacks
- Enforcement of use policies
- Collection of evidence

- Detection of attacks
- Detection of policy violations

IDSs are used in particular for the identification, assessment, reporting, and reporting of unauthorized or unapproved network operations to deter potential disruption. The IDS can be split into 2 groups, network-based or host-based based, based on the data sources they use. NIDS (Network Intrusion Detection Systems) test network detection packets[2]. The audit trails or system calls generated by each server are examined. TCP dumping data into connections containing network session context information.

As network traffic volume increases, several sensors are used by many NIDSs and distributed computers to increase computing speed. NIDS can detect IP based attacks, like multi-computer Denial-of-Service attacks.The host-based IDS finds these attacks difficult to identify when it tracks information obtained from the computer device only. As more systems communicate across networks, NIDS is gaining prominence. Also, IDSs can be classified using detection methods [3]. Two forms of identification occur basically: misuse detection and anomaly detection.The main implication of the two methods is the assumption that the detection of misuse detectionthe intrusions is based on the characteristics and anomaly detectionof known attacks.[4].

1.1 Misuse detection

The identification of misuse intrudes in respect of known attack characteristics. This method looks for patterns and signatures of documented network attacks. Known attack signatures are normally supplied in an updated database. Any behavior consistent with known attack patterns or vulnerabilities is considered invasive. The misuse detection System block diagram is shown in figure1.

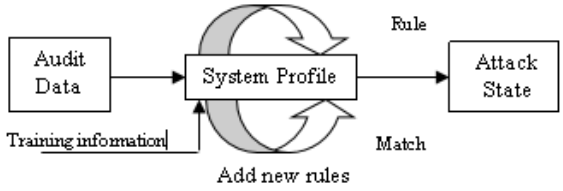


Figure 1 Misuse based detection system

1.2 Anomaly detection

The methodology is focused on traffic irregularities identification. The divergence from the standard profile is calculated from the tracked traffic. Several different variations of this technique based on the metrics used for measuring the variance in traffic profile have been suggested. The anomaly detection device block diagram shows in fig.2

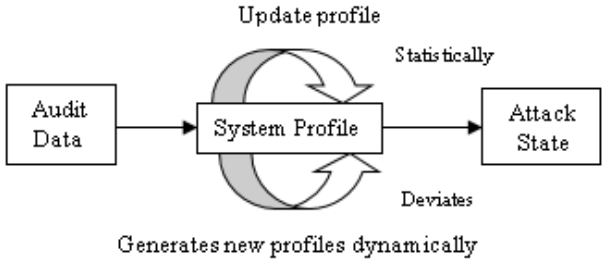


Fig.2 Anomaly-based detection system

The document below is structured. The overview of the literature concerned is given in section 2. Section 3 discusses the proposed K mean an algorithm or network intrusion detection. The experimental findings are discussed in Section 4. This concludes this paper in section 5.

II. LITERATURE REVIEW

A comparative study of IDS techniques and methods have been discussed in this section.

In [5] The author proposed a malicious website technique. The self-designed JAVA program uses static content Web pages and regular expressions to create signatures. The shape of Honeypot's website is finally completed as it is used to search websites. The Microsoft OS consists of four modules: source code analysis and attack detection proxy behavior. Recording. The static research of this operating system often reveals low precision for automated & successful identification of malicious nodes.

In [6] The proposed IDS method produces a lot of unimportant, false, and redundant alerts in observing network attack.Therefore, this system's disadvantage. The online technique is used for the data set ShahidRajaee Port Complex and with dataset DARPA 1999. When the outcome of this method is obtained, the number of alerts is decreased to 94.32 percent. This system also has a high alarm rate and high detection rate.the approach is not ideal for online research, so a new method has to be developed to minimize increase detection rateandincorrect warning rate.

In [7] The writer has suggested a SQL injection attack detection approach that is a technique for stealing confidential data or back-end database information such as a credit card number. The use of query transformation and text similarity is advised to identify different forms of SQL injection attacks by SQL Injection Detection (IDS-SQLiDDS). For testing built using MySQL & PHP, five honeypot web applications are used. This is aimed at identifying all types of SQL attacks.

In [8] The writer examined the (advanced persistent threat)APT using various types of attack methods in the initial stage of access to the unwanted system. The 'Packet Stage' IDS is extended with its design approach to enhance the results. This model is made up of event search-patterns (P),hypothesis (H),classes (C),rules (R).The system model is obtained by combining log information from distributed networks, and the network node is also extracted without log lines knowledge.The loglines in this model form distinguish and detect various meaningful subsets. After applying this model, the SCADA dataset is used for the experiment and the outcome is a positive 1. False-positive is 0.

In [9] The author proposed to use a cross-site-scripting attack (XSS) method to inject javascript functions to exploit known vulnerabilities in the web application. There's been different types of XSS attacks or operating in 2 kinds that monitor web application's cross-site vulnerabilities. For this method three steps are used, namely sanitizing, coding, and matching ordinary expressions. For the avoidance of malicious insertion, all Html tags are sanitized by the user. The Javascript code is specified as per potential standard malicious expressions. For true or non-valid tests, the standard predefined expressions match any user input.

In [10] Proposed a malicious JavaScript detection tool. This suggested approach uses linear regressions and three layers of stacked auto-encoders (SDA). In comparison, the test results are contrasted with other classifiers with strong positive and second-best false positive.

In [11] The deep-learning approach was to construct an efficient and versatile NIDS. The technique called Self-taught learning (STL) allows sparse autoencoder & Soft Max regression to be combined. The data set of the NSL-KDD is used to apply and evaluate the approach proposed. Promising classification precision for both 5 class and binary classification is achieved at a promising level. The overall F score of 75.76% is extracted in its 5 category classification.Unsupervised learning to learn the flow of natural networks. RNN, Deep learning,and car encoder principles are utilized in this process. The exactness is not fully contained and the exactness for the proposed process is not so exact. A concept for tracking network flow data has also been suggested. An accuracy of 75.75% with six specific features is claimed, however, an assessment through the NSL-KDD dataset is presented.

In [12] The state-of-the-art survey of deep learning technologies was proposed for the NIDS paradigm for health monitoring., Conventional methods are compared with four popular methods of deep learning like (recurrent neural network)RNN&CNN (convolution neural network), auto-encoders, orRBM (restricted Boltzmann machine). Test results show that traditional approaches are lacking and deep learning methods are extremely accurate.

In [13] The suggested (deep neural network)DNN combined toRectified Linear Unit function&ADAM-optimizer proposed tasks besides advanced persistent threats,100 hidden units. KDD data is used to classify and accurately. Both LSTM (long-term memory) and RNN models are needed for the potential use of 99 percent.

In [14] The survey of NIDS methods was mentioned andcomprehensive taxonomy constructed by low & deep learning was established. The most important findings from this work are aggregated. Table 1 offers a specific comparison of NIDS techniques.

III. PROPOSED METHODOLOGY

In the existing work, a clustering-based hybrid approach has been used in which an optimal number of clusters will be generated or later clustering is applied. For identifying optimal clusters and K-means are used as clustering methods, a genetic algorithm has been used.

We implement the feature selection first with the information gain technique in the suggested methodology. We subsequently applied differential development to find the maximum number of clusters and then clustering by GMM methodology.

The population-based metaheuristic search algorithm,difference evolution (DE), optimizes the problem by successfully enhancing the candidate solution. The method creates system architecture by retaining a population of candidate solutions (individual) or by combining existing solutions in a particular phase. The next iteration of the algorithm retains candidates with better objective values, such that as a population participant the new goal value of an individual is improved or new objective value is discarded. The process will continue until that completing criterion is accomplished.

Initialization

The initial value in [X j^L, x j^U], is typically randomly chosen uniformly for any parameter j at the lower of the X j^L and upper of the X j^U.

Mutation

Three vectors (X pour (r1,G) X (r2,G) X_ (r3,G)) are chosen at random to differentiate the indices I r1, r2, and r3. The weights of the two vectors are applied to the third vector by adding a donor vector V (i, G+1):

$$V_{i,G+1} = X_{r1,G} + F.((X_{r2,G} - (X_{r3,G}),\quad r_1 \neq r_2 \neq r_3 \neq i$$

where *F* is a constant from (0, 2)

Crossover

Three parents have been chosen and the infant is one of them disturbed. With target vector (Xi,G) elements and donor vector elements, the Donorvector (Xi,G) is developed. Donor vector components are like probability in the test vector with CR: $randj,i \sim U(0, 1)$, Irand is integer random (1, 2, ...D) where D is the dimension of solution EX. the number of control parameters. Irand is in charge of Vi,G+1 ≠ Xi,G.

Selection

The comparison with measure Vi, G+1, the Xi, G goal is accepted to the next generation with the better fitness rating. The following equation can be used to represent the selection operation in DE:

where $i \in [1, N_P]$.

GAUSSIAN MIXTURE MODEL

The K clusters are available (The assumption here is that are established and it is K for sake of simplicity).For each k, it is calculated therefore μ and ∑ to be. If there's only one distribution, the maximum-like process may have been calculated. Since these clusters do therefore have K and all such distributions' probability densities are known as the linear function of densities [16], i.e. π

$$p(X) = \sum_{k=1}^{k} \pi_k G(X|\pi_R, \sum k$$

where $\pi_k$is a k-th distribution mixing coefficient.
To estimate the parameters by log-like technique, compute p(X|μ, $\sum$, π)
Lnp(X|μ, $\sum$, π)
= $\sum_{i=1}^{N} p(X_i)$
= $\sum_{i=1}^{N} ln \sum_{k=1}^{K} \pi_k G(X_i|\mu_k, \sum k)$

IV. IDS DESIGN

Fig.3 shows the system design for the intrusion detection. The flow diagram shows that he has taken steps in carrying out this analysis. The IDS can be seen as follows:

A. Dataset

In this IDS, the cup dataset KDD-99 is the dataset. A dataset contains 42 characteristics showing various data points features in the dataset.This compilation of data comprises 4.8 million cases. The dataset contains two, R2L, U2R, and poking intrusions. The above-mentioned forms of intrusion can be additionally listed as 22 types. Reference [3] is given in-depth. This dataset is used in larger datasets to discoverthe form of the IGKM algorithm. Theminor part of the KDD-99 dataset, with 1000 examples, is also used for this article. The essence of the IGKM algorithm is seen in smaller datasets for this dataset.

B. Feature Selection

The reason behind the selection of important and significant functions is the consistency of the structural alert correlation and to represent the attack steps from the alert pattern (SAC). The two-tier ranking, i.e. the function ranking and the additional feature is described in this section. The classification function uses a filtering approach with the Gain algorithm (IG) algorithm.

The step aims to classify subsets of features in a decreasing order based on high data entropy. The additional function process, meanwhile, is focused on the work in which the detection of relations between alerts involves the study of attributes of alerts, and it may not be enough to extract specific attributes to figure out entirely the relation between these alerts. The goal of this step is therefore to widen the connection between alerts with a higher level of classification than the initial ranks.

C. Training phase

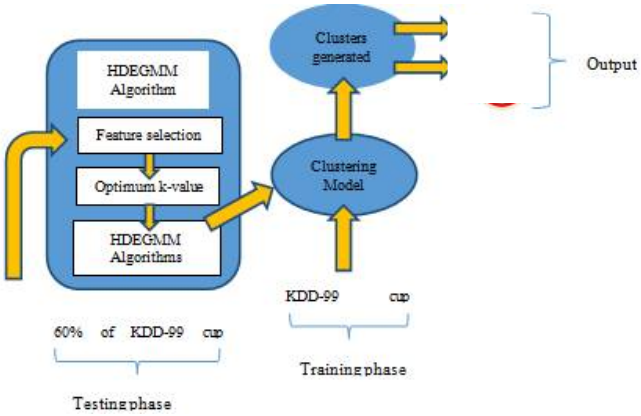
The process and during the training period consistsof providing known inputs to the algorithm. The reduced data set attribute is used with the IGKM algorithm. The data set is grouped, with optimal value for the type of clusters to be made. In the training stage, the IGKM algorithm is conditioned by 60 % of the KDD-99 data set or clusters. There are ten generations to reach an optimized cluster.

D. Testing phase

The function is provided unknown inputs during the test process and users confirm unless the result is accurate and not.Random input values are supplied as input during that process from the remaining 40% of the KDD-99 dataset. The system then uses clusters that were generated in the workout to track the type of attack.

E. Classifier

The IDSusesa classifier to validate if the algorithm's performance is correct. To verify the correct result, the classifier uses ID mapping. This is reduced to a date set of seven attributes during the reduction of the attribute, such as the ID number. The ID number shows the corresponding instance in the official KDD-99 dataset. The results ID number is used to cross-reference and search for accuracy.



CFigure 3 Flow diagram of IDS that uses HDEGMMtechniques.

V. RESULT DISCUSSIONS AND ILLUSTRATIONS

The series of positive factors (TP) is employed to evaluate the output: The number of positive references

Tuples are accurately labeled by classification.

False-positive (FP): theRefers to incorrect labeling of several negative classifiers.
False Negative (FN): These are the good days that were mislabeled negative.
Precision: The ratio of true positive to false positive.

$Precision = (TP/FP)$

Recall The proportion of true positive to several false positive or false negatives.
$Recall = TP/(FP + FN)$

Accuracy (ACC): That's the total accuracy of the classifier.

$ACC = (Precision/Recall)$

Table 1 IGKM and HDEGMM algorithm for KDD-99 datasets Comparison of accuracy

Algorithm	Precision	Recall	Accuracy
IGKM	0.831394	0.831394	0.831081
HDEGMM	0.994038	0.994038	0.992732

The comparisons between existing research studies and research programs I e, IGKM and HDEGMM can be seen in Table 1. The analysis indicates that the proposed study has improved precision and recall quality in comparison to that of the previous study.

In figure 4 or fig 5, the graph demonstrates the fitness values of both research works. The graph reveals that the HDEGMM is higher than the IGKM fitness value.

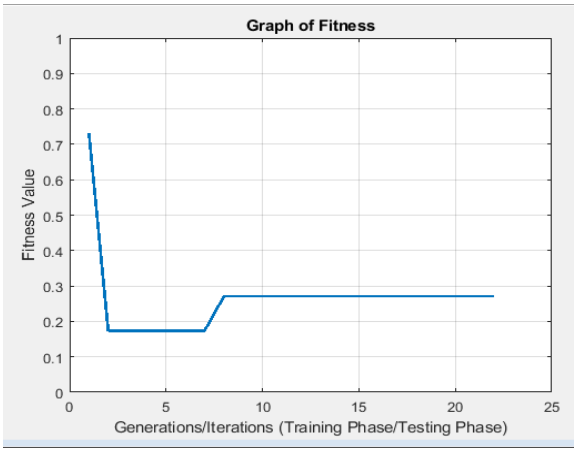


Figure 4 Graph of fitness of KDD-99 dataset of IGKM

The fitness function used by algorithms defines optimal k value before clustering during the training and testing phase.

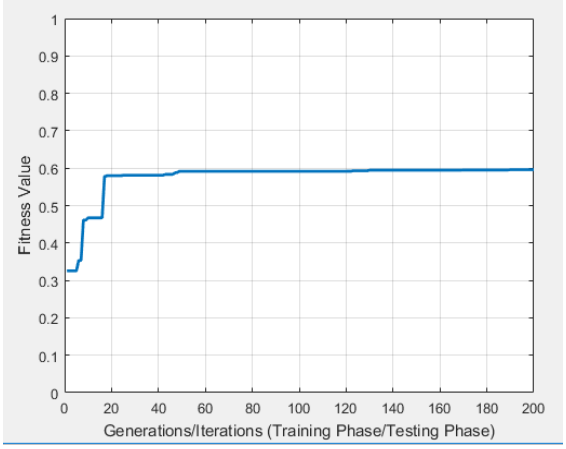


Figure 5 Graph of fitness of KDD-99 dataset of HDEGMM

The time complexity of both research works is seen in figures 6 and 7. Time complexity is a computer science term that quantifies how long a set of code and algorithms take to a methodand run depending on the input amount.

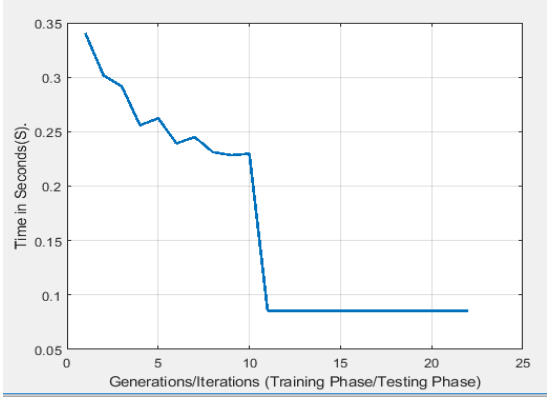


Figure 6 Time complexity of the IGKM

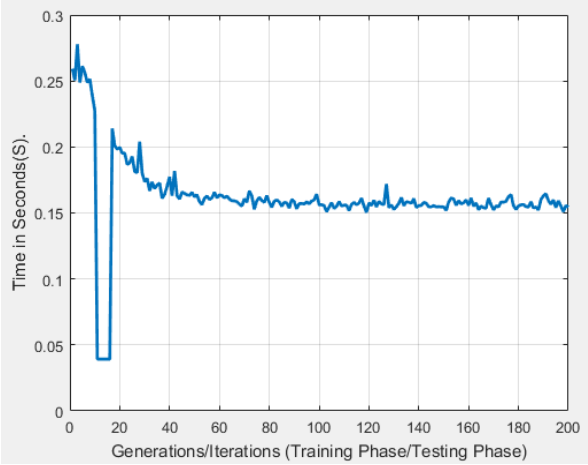


Figure 7 Time complexity of the HDEGMM

The time complexity of a proposed which is better than the IGKM can be seen in Figure 7.

VI. CONCLUSION

Intrusion crimes are on the rise every day. Therefore, compared to IDS utilizing standard clustering algorithms, an optimal intrusion detection method must be found. We have developed IDS using the HDEGMM algorithm IDS in this paper. The optimal value of k is determined by using the fitness function to effectively detect the attack by optimized clusters. Through this paper's tests, we can infer that a method of IDS that uses IGKM algorithms is less specifically the dataset in use but that in contrast with the intrusion detection system used by IGKM, the intrusion detection system of HDEGMM uses a clustering algorithm that shows comparatively greater precision.

REFERENCES

[1] Zhang Z, Shen H. Application of online-training SVMs for real-time intrusion detection with different considerations. Computer Communications. 2005; 28(12):1428–42.

[2] Shyu ML, Chen S, Sarinnapakorn K,Chang L. A novel anomaly detection scheme based on principal component classifier. Proceedings of the IEEE Foundations and New Directions of Data Mining Workshop, in conjunction with the Third IEEE International Conference on Data Mining (ICDM03), 2003. p. 172–79.

[3] Denning DE. An Intrusion-Detection Model', IEEE Transactions on Software Engineering. 2006; SE-13(2):222–32.

[4] Lee W, Stolfo SJ. A framework for constructing features and models for intrusion detection systems. ACM Transactions on Information and System Security. 2000; 3(4):227–61.

[5] Landgrebe TCW, Pavel P, Duin RPW, Bradley AP. Precision-Recall Operating characteristic (PROC) curves in imprecise environments. Proceedings of 18th International Conference on Pattern Recognition, ICPR2006, Hong Kong. 2006; 4. p.123–27.

[6] Wang W, Guan XH,Zhong X. Processing of massive audit data streams for real tim anomaly intrusion detection. Computer communications. 2008; 31(1):58–72.

[7] Garcia-Teodoro P, Diaz-Verdejo J, Macia-Fernandez G, Vazquez E. Anomaly-based network intrusion detection: techniques, systems and challenges. Computer Security. 2009; 28(1-2):18–28.

[8] MIT Lincoln Labs. DARPA intrusion detection evaluation [Online]. 2014 Nov. Available from: http://www.ll.mit.edu/mission/communications/ist/corpora/ideval/index.html

[9] Lippmann RP, Fried DJ, Graf I, Haines JW. Evaluating intrusion detection systems: The 1998 DARPA off-line intrusiondetection evaluation. Proceedings of the 2000 DARPAInformation Survivability Conference and Exposition(DISCEX'00), Hilton Head, SC. 2000; 2. p. 12–26.

[10] Tavallaee M, Bagheri E, Lu W, Ghorbani AA. A Detailedanalysis of the KDD CUP 99 Dataset. Proceedings of theIEEE Symposium on Computational Intelligence in Securityand Defense Applications. 2009; 1–6

[11] Tsai C-F, Hsu Y-F, Lin C-Y,Lin W-Y. Intrusion detection by machine learning: A Review. Expert Systems with Applications. 2009; 36(10):1994–2000.

[12] Witten IH, Frank E, Hall MA. Data Mining- Practical Machine Learning Tools and Techniques. Morgan Kaufmann: San Francisco, CA, 2011.

[13] Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS, Zhou Z, Steinbach M, Hand DJ, Steinberg D. Top Ten Data

Mining Algorithms. Knowledge and Information Systems Journal, SpringerVerlag London. 2007; 14(1):1–37.

[14] Gaffney JE, Ulvila JW. Evaluation of intrusion detectors: A decision theory approach. Proceedings of the IEEE symposium on Security and Privacy, S&P'01, Oakland, CA, USA. 2001; 50–61

[15] Apte C, Weiss S. Data mining with decision trees and decision rules. Future Generation Computer Systems. 1997; 13(2-3):197–210.

[16] AbdJalil K, Kamarudin MH, Masrek MN. Comparison of Machine Learning algorithms performance in detecting network intrusion. 2010 International Conference on Networking and Information Technology (ICNIT), Manila, IEEE. 2010. p. 221–26.