INFO- 2172 UDK: 004.7.056:004.62
Primljeno/Received: 2016-03-12 Authors Review/Pregledni rad

AN INNER INTERRUPTION DISCOVERY AND DEFENSE SYSTEM BY USING DATA MINING

SUSTAV ZA OTKRIVANJE I OBRANU KORIŠTENJEM RUDARENJA PODATAKA

S.Singaravelan[1], S.Jerina Catherina Joy[1], D.Murugan[2]

Dept.of.Computer Science and Engineering, P.S.R. Engineering College, Sivakasi, Tamil Nadu, India[1]; Dept.of.Computer Science and Engineering, Manonmaniam Sundaranar University ,Tirunelveli , Tamil Nadu, India

Abstract

Network security helps to prevent the network against the intruders from performing malicious activities. The security can be provided to the networks using firewalls, anti-virus software and scanners, cryptographic systems, Secure Socket Layer (SSL) and Intrusion Detection Systems (IDS).Authentication is the commonly used technique to protect the unauthorized users from the network. But, it is easy to compromise the login passwords using brute force attacks. The IDS and firewalls concentrate on the external attacks, while the internal attacks are not taken into account. In order to solve these issues, this paper proposes an Inner Interruption Discovery and Defense System (IIDDS) at the System Call (SC) level using data mining and forensic techniques. The user's profiles are maintained and compared with the actual dataset using Hellinger distance. A hash function is applied on the incoming messages and they are summarized in the sketch dataset. The experimental results evaluate the proposed system in terms of accuracy and response time.

Sažetak

Mrežna sigurnost pomaže zaštititi mrežu od uljeza u obavljanju zlonamjernih aktivnosti. Sigurnost se može osigurati mrežama koristeći vatrozide, antivirusni softver i skenere, kriptografske sustave, *Secure Socket Layer* (SSL) i sustave za otkrivanje upada (IDS). Autentifikacija je najčešće korištena tehnika za zaštitu neovlaštenih korisnika na mreži. No, lako je kompromitirati lozinke za prijavu pomoću napada na silu. IDS i vatrozidi koncentriraju se na vanjske napade, dok se interni napadi ne uzimaju u obzir. Da bi se riješili ti problemi, u članku se predlaže unutarnje prekidanje i obrambeni sustav (IIDDS) na razini *System Call* (SC) razine pomoću rudarenja podataka i forenzičke tehnike. Profili korisnika održavaju se i uspoređuju sa stvarnim skupom podataka pomoću Hellingerove udaljenosti. Na dolazne poruke primjenjuje se *hash* funkcija i oni su sažeti u skupu skica podataka. Eksperimentalni rezultati procjenjuju predloženi sustav u smislu točnosti i vremena odziva.

INTRODUCTION

Network is a linkage of nodes using a certain topology that helps in transferring data from source to destination. Networks can be classified as data networks and synchronous networks. Data networks are connected using routers, whereas synchronous networks transfer data via switches. The data networks are prone to crucial attacks such as eavesdropping, viruses, worms, phishing, IP spoofing attacks, Denial Of Service (DOS) attacks etc. Network security is the specialized field of computer networking, which involves securing the computer network from illegal access, misuse or modification of the network resources. The security can be provided to the networks using firewalls, anti-virus software and scanners, cryptographic systems, Secure Socket Layer (SSL) and Intrusion Detection Systems (IDS). The IDS prevents the network from attackers and also notifies about the abnormal activities of the network. It can be implemented using either hardware or software. An efficient IDS must possess the following characterisitics:

- Strict prohibition of new intruders or malware
- No performance impacts on the system
- Constant monitoring
- Transparency
- Reliable and minimized false alarm rates

The IDSs are majorly classified into two types such as misuse or signature-based IDS and anomaly-based IDS. Signature-based IDS can detect only the known attacks by comparing with the signatures available in the database. It shows the low detection rate for the unknown attacks. The network anomalies are monitored using rules such as intervalrule, retransmissionrule, integrity rule, delay rule, radio transmission range and jamming rule. Anomaly-based IDS detects the attacks by analyzing the normal and anomalous behavior of the users. It will never send an alarm, if the malicious activity looks like normal traffic. There is a possibility in the generation of more false positives. The load of the network may be increased, as it requires periodic updation of normal profiles. Anomaly-based systems are capable of detecting unknown encounters without any prior knowledge. They are further classified as statistical-based, knowledge-based and machine learning-based IDS according to the nature of the process. In statistical-based IDS, an anomaly score is generated for each activity in comparison with the reference profile. The parameters such as univariate, multivariate and time series model are used in score generation. The knowledge-based IDS depend on the knowledge expert systems, finite state machines, data clustering and outlier detection. The analyzed patterns are periodically updated in the machine learning-based approach to improve the performance of intrusion detection. Signature-based approach is more flexible and accurate than the anomaly-based approach. **Figure(1)** shows the implementation of IDS in a network.

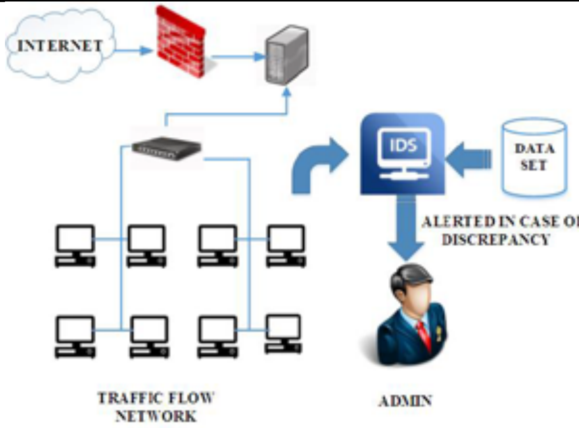

Figure 1: Implementation of IDS in a network

The traditional IDS suffer from the following drawbacks:

- It is difficult to detect the malicious activities of the internal attackers.
- The Operating System (OS) level System Calls (SCs) are not considered in the process of intrusion detection.
- They cannot detect the intrusions and malware in parallel.

In order to resolve these drawbacks, an Inner Interruption Discovery and Defense System (IIDDS) at the System Call (SC) level using data mining and forensic techniques. The main objective of this paper is to detect the internal intruders with increased accuracy and reduced response time. The user profiles are compared with actual dataset using Hellinger distance.

The remaining sections of this paper are organized as follows: Section II surveyed the traditional approaches of IDSs to detect the malware and attacks. Section III provides an overall description of the Inner Interruption Discovery and Defense System (IIDDS). Section IV shows the performance analysis of the proposed system. Section V presents the conclusion and future work of the paper.

RELATED WORK

This section reviews the existing approaches related to IDSs to provide network security. *Uddin, et al* /**1**/proposed a signature-based multi-layer Intrusion Detection System (IDS) to detect the threats with high success rate. It automatically created multiple small databases in a dynamic fashion to update signatures of new malware at regular intervals. The challenges of signature-based systems were signature creation, storage, maintenance, updation, and avoiding traffic flooding. A complementary payload-based anomaly detection system was used to detect the newly emerging attacks. Multiple IDS were deployed in different layers. The multi-layer approach optimized the detection rate and also detected uncommon attacks.

Saeed, et al /**2**/ reviewed in detail about the malware infection and the solutions to improve the intrusion detection systems. Signature-based, anomaly-based and heuristic-based intrusion detection techniques were studied. The signature-based techniques were found to be efficient in terms of known malware. The anomaly-based techniques generated a high rate of false positives or false negatives. The heuristic-based techniques were powerful in detecting obfuscated new malware, by deriving solutions from different sources without any prior knowledge.

Saeed, et al /**3**/ proposed a dynamic model checking based multi-agent architecture to detect intrusions in an efficient manner. The suggested architecture had a group of cooperative collaborating agents placed in the hierarchy. The communication rules were specified for interaction between the nodes. Temporal logic rules were applied in the agents to combine the obtained results. Dynamic model checking algorithm was used to check whether the specific formulae was satisfied or not. When compared with the previous works, the proposed architecture provided promising results. *Sai, et al* /**4**/ proposed a Naïve Bayesian based intrusion detection using dogmatic and look-ahead approaches. The behavior of malware infected nodes was observed during their opportunistic encounters. The two challenges were extended from the Bayesian intrusion detection systems to the DTN. The challenges were: (1) filtering the false evidence sequentially (2) insufficient evidence vs. evidence of the collection risk. These challenges were solved using the dogmatic filter and adaptive look ahead approach.

Thu /**5**/ proposed an integrated Intrusion Detection and Prevention System (IDPS) to detect the internal attackers using Honeypot. The suggested IDPS was a combination of anomaly detection and signature detection schemes. Honeypots were the network attached servers configured in the firewall to monitor the activities of internal users. The proposed system consumed limited resources and reduced the false alarm rates, when compared to the other systems. *Chadli, et al* /**6**/ designed a multi-agent based IDS architecture to improve the performance and security of the network. The proposed architecture is a combination of cluster-based and cooperation-based hierarchical model. A knowledge base, which contained the global ontology data, was used to improve security. When compared to the existing IDSs, the hierarchical architecture performed better in terms of performance and security. *Meng, et al* /**7**/ proposed a threshold signature scheme to provide security in the absence of the trusted party for limited resource utilization. This scheme reduced the computational overhead and resources required for computation using bilinear pairing. The generation and verification of threshold signatures eliminated the conspiracy attacks. The security was improved by the bilinear pairing based threshold key management scheme.

Sharma, et al /**8**/ presented a signature-based IDS using an identity-based scheme to achieve network protection from worm-hole attacks. The suggested scheme resisted the process of certificate distribution between the nodes in the network. The computational overhead was decreased in the identity-based signature scheme due to the elimination of signature distribution.*Casas, et al* /**9**/ presented an Unsupervised Network Intrusion Detection System (UNIDS) to detect the unknown network attacks without the use of any knowledge base. To identify the different attacks sub-space clustering and multiple evidence accumulation based unsupervised outlier detection approach was applied. The performance and detection accuracy was better in UNIDS than the traditional IDSs.*Panda, et al* /**10**/ designed an IDS using hybrid intelligence approach by combining classifiers to achieve the best detection rate. Two-class classification strategy and ten-fold cross validation method was used for final classification. The meta learning strategies were used along with grading and END to enhance the performance of the proposed system.

Akbani, et al /**11**/ proposed an Enhanced Machine Learning (EML) based reputation system to defend against many attack signatures. Reputation systems predicted future behavior of nodes by observing their past behavior. The EML based system resulted in very low bandwidth and computation overhead.*Khouzani, et al* /**12**/ formulated pontryagin maximum principle based decision problem to quantify the damage caused by a malware. Flippov-Cesari theorem was used to establish an optimal solution. The results proved that the healing rate and the recovery rate were higher than the other systems.*Houque, et al* /**13**/]presented an IDS using Genetic Algorithm (GA) to detect various intrusions in the network. Darwininan's principle was used to optimize the population of candidate solutions. The set of chromosomes obtained in the pre-calculation step was used to find the attack type. When compared to the existing IDSs, the proposed IDS resulted in high detection rate. Mechtri, et al /**14**/ proposed a distribution and cooperation based IDS architecture to overcome the weakness and flaws of the existing IDSs. The distribution was achievedby a local IDS on each network and the coopration was guaranteed by the collaboration of mobile and stationary agents. The proposed system was flexible, lightweight, and fault tolerant.*Selvaraj, et al* /**15**/ proposed a novel idea using honey pot technique and packet data analysis for intrusion detection in both network and anomaly based systems. The system was trained with a sample of malware or deep analysis of packet inspection. Honey pot had a large amount of energy and a strong field of security, which made the proposed system better than the existing IDSs.

PROPOSED METHOD

This section presents the overview of the proposed Inner Interruption Discovery and Defense System (IIDDS). The main objective of this paper is to detect the internal intruders with increased accuracy and reduced response time. **Figure(2)** describes the overall flow of the IIDDS.

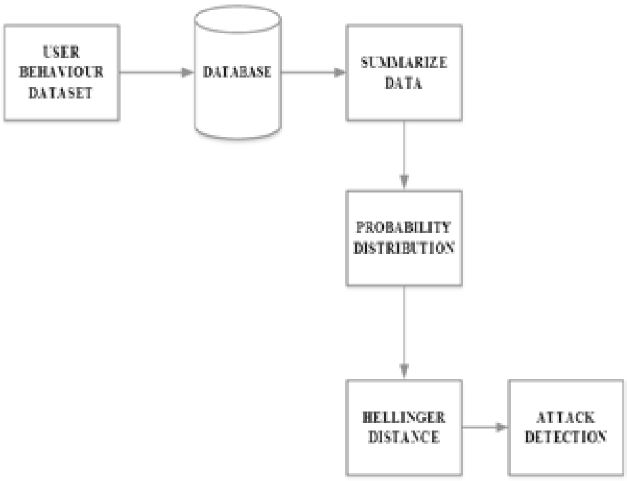

Figure 2: The overall flow of IIDDS

The proposed IIDDS consist of four stages as follows:

- Loading the user dataset
- Summarizing the sketch dataset
- Calculating Hellinger distance
- Attack detection and Prevention

LOADING THE USER DATASET

The user data which contains the user ID, process ID andSC are loaded in the dataset. The data in the dataset is available in the Comma Separated Value (CSV) format. It is difficult to process the CSV format so, it is converted into text file format. The raw data is preprocessed to remove the incorrect, inconsistent and redundant data.The inconsistent and noisy data are removed through data cleaning, whereas redundancy is eliminated through data reduction. The size of the dataset is reduced using clustering and aggregation to enhance the efficiency of mining and to improve the quality of the dataset. The user log files consists of the user SCs and their sequence of occurrence. The behaviours of the user are identified by the analysis of user profiles using SC patterns. The mining patterns available in data mining technique are used for finding the SC patterns.

SUMMARIZING THE SKETCH DATASET

The dataset is stored in the compact form and a hash function is applied on all the incoming SC patterns. The probability distribution, namely, sketch data distribution is established for each attribute in the dataset. The similarity scores of the user's input and behavior patterns of the other users in the system are analyzed to find the internal attacks. The hash function is computed for the large dataset to obtain a reduced size summary using data aggregation.

CALCULATING HELLINGER DISTANCE

Hellinger distance is calculated to find the distance between two probabilistic distributions. The distance of probability measure is used to avoid the flooding of packets in a network. The number of users in the network and their profiles along with SC patterns are provided as an input to find the attackers in the network. The normal dataset is converted into sketch dataset through summarization. The probability distribution is calculated for each entry in the sketch dataset. A summarized table is created in the sketch dataset to store the summarized entries based on the probability values. The Hellinger distance is calculated between all the probability values present in the table. The distances are compared and the attackers are identified based on the lowest distance.A threshold was computed based on the summarized data and the probability values are compared with threshold. If the value is larger than the threshold the users are identified as abnormal users.

Pseudocode for Hellinger Distance

```
Input: User Profiles along with SC patterns D

Output: Attackers A

Process

For each entry x € D
        Sketch(H(k,v))<-H(k,v)
End for
For each sketch entry x € D
        P(x)<-x
        T<-P(x)
End for
For each probability p € T
        Hdist(p1,p2)<-$p_i - q_i$
        A<-Min(p1,p2)
End for
```

Step 1: Input the preprocessed dataset that contains user profiles and SC patterns.

Step 2: Convert the dataset into sketch dataset.

Step3: Calculate the probability values for each entry in the sketch dataset.

Step 4: Create a table in the sketch dataset and store the entries with the probability values.

Step 5: Calculate the Hellinger distance between each probability value.

Step 6: Classify the normal and abnormal users based on the distance.

Step 7: The entries of users with minimum distance are termed as abnormal users.

Step 8: The profiles of the abnormal users are captured by the SC monitor and sent to the SC filter

Step 9: The SC filter isolates the users and prevents the system from attacks.

ATTACK DETECTION AND PREVENTION

Attack may cause a serious damage to the system and its data. It is necessary to identify different types of attacks to prevent the system from failure. Attacks are classified as passive and active attacks, where active attacks are more severe than passive attacks. The passive attacks include eavesdropping, traffic analysis, capturing passwords etc. discloses the information without the knowledge of the user. Active attacks such as viruses, worms, Trojans modify or corrupt the information by transiting malicious codes. The attacks that help to gain the unauthorized access is termed as close-in attacks. Thephising attacks normally occur in the internet, in which the e-mail users are hacked via fake websites. The authentication codes are cracked using three types of password attacks such as brute-force attack, dictionary attack and hybrid attack. The dictionary attack uses a list of words, whereas the brute-force attack attempts all the combinations.Insider attacks are the attacks attempted by the internal users of the organization and it may be either passive or active. The flooding attacks broadcast many requests that lead to DOS.

In the proposed system, the attacks are identified using the Hellinger distance.The SCs sent to the kernel are stored in the log files, while the user executes the shell commands. The similarity measures between the new SCs are compared to identify whether the user is an authorized user or not. The SC monitor receives a notification, when an abnormal user is detected. The SC filter isolates the intruder to protect the system from attacks. If a user tries the authentication code of another user, it can be detected using the usage patterns of the users.

PERFORMANCE ANALYSIS

This section presents the performance analysis of the IIDDS. The proposed system is compared with the Term Frequency-Inverse Document Frequency (TF-IDF). The performance of the proposed system is evaluated in terms of

- Detection accuracy
- Decisive rate threshold
- Response time
- Abnormal user count
- False alarm rate

DETECTION ACCURACY

Detection accuracy is defined as the ratio of the number of triggered alerts to the total number of alerts in the trace.

$$Detection\ Accuracy = \frac{No.of\ triggered\ alerts}{Total\ no.of\ alerts} \quad (1)$$

The graph in **figure(3)** depicts the comparison between detection accuracy of the existing and the proposed system. The detection accuracy of the TF-IDF based system is only 50%, whereas the detection accuracy of IIDDS is 80%. Thus, the proposed IDS achieved higher detection accuracy, when compared to the existing system.

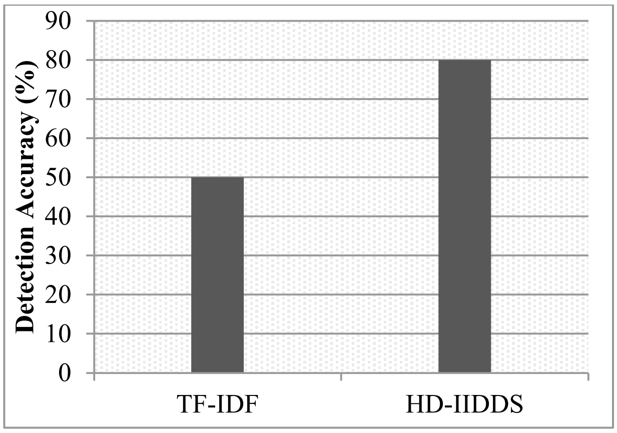

Figure 3: Detection accuracy of TF-IDF and IIDDS

DECISIVE RATE THRESHOLD

Decisive rate threshold can be defined as the similarity score between the current user profile and the profile of other users.

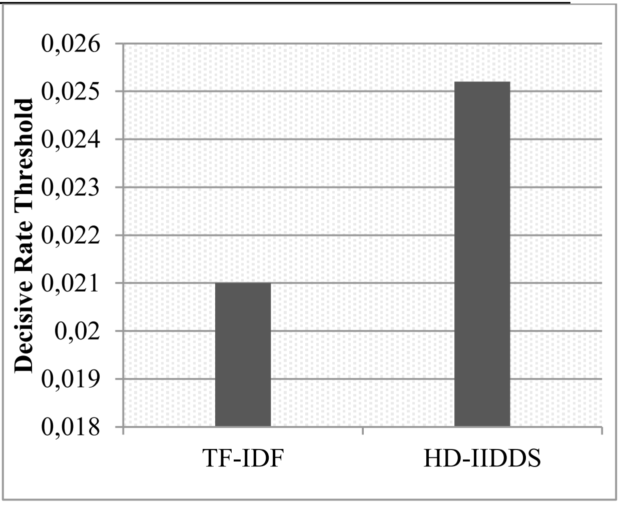

Figure 4: Decisive rate threshold of TF-IDF and IIDDS

Figure(4) illustrates the graph that is plotted for showing the decisive rate threshold of TF-IDF system and IIDDS. The decisive rate threshold of the TF-IDF is 0.021 and the decisive rate threshold of the IIDDS is 0.025. As, the decisive rate threshold is high the IIDDS detects more accurately when compared to TF-IDF based system.

RESPONSE TIME

Response time can be defined as the time taken to produce the trigger,when an attack is detected.**Figure(5)** shows the comparison between the response time of TF-IDF and IIDDS. The TF-IDF takes 42 ms for giving a response, whereas the response time of IIDDS is 25 ms. The response time of the proposed IDS is low, when compared to the existing system.

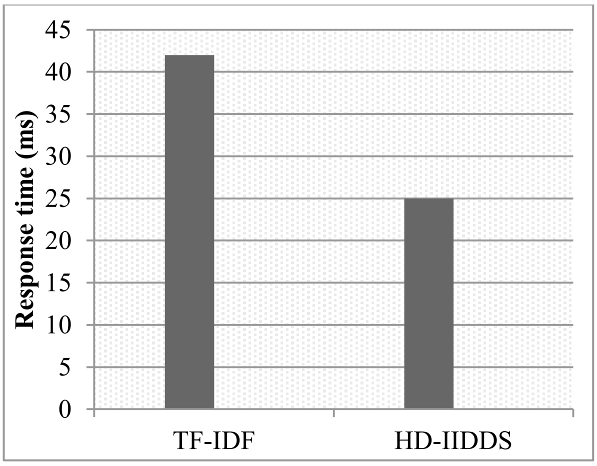

Figure 5: Response time of TF-IDF and IIDDS

ABNORMAL USER COUNT

Abnormal user count is the detection rate of the number of attackers or malicious users in the system. The graph in **figure(6)** shows comparison of the abnormal user count in both the existing and proposed system. The abnormal user count is 80 in TF-IDF but in IIDDS, there are only 50 abnormal users. The number of abnormal users is reduced in the proposed method with the help of SC filter.

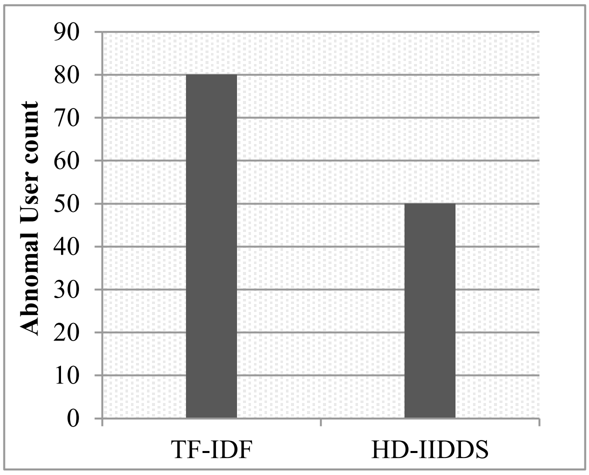

Figure 6: Abnormal user count in TF-IDF and IIDDS

FALSE ALARM RATE

False alarm rate is defined as the ratio of the number of wrongly triggered alerts to the number of total alerts.

$$False\ Alarm\ Rate = \frac{No.of\ wrongly\ triggered\ alerts}{Total\ number\ of\ alerts} \quad (2)$$

The graph in **figure(7)** describes the comparison of false alarm rate between the existing and proposed system. The false alarms generated by TF-IDF is 50%, whereas the IIDDS generated only 10% of false alarms. When compared to the existing system, the false alarm rate is reduced in the proposed system. Hence, it is clearly understood that the proposed HD based IDPS outperformed the existing systems in terms of detection accuracy, decisive rate threshold, response time, abnormal user count, and false alarm rate.

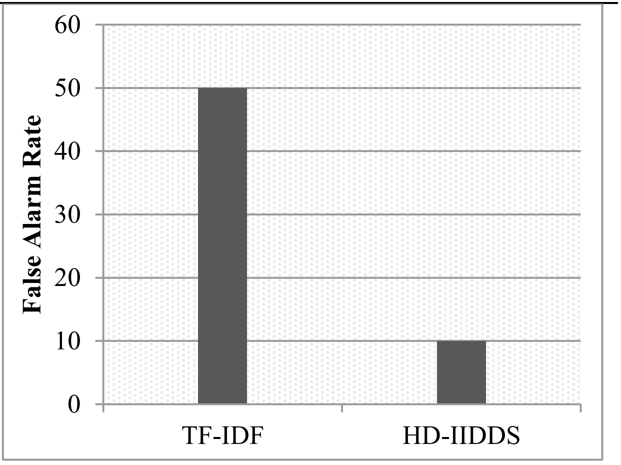

Figure 7: False alarm rate generated by TF-IDF and IIDDS

CONCLUSION AND FUTURE WORK

This proposed method in Inner Interruption Discovery and Defense System (IIDDS) to prevent the system from attacks and to find out the internal attackers based on the SC patterns. The Hellinger Distance is used to compute the probability values for the user profiles. The SC monitor and filter is used to isolate the internal attackers. The efficiency of the system is improved and the system is compared with the TF-IDF based system.The response time and the false alarm rate are reduced, whereas the detection accuracy and decisive rate threshold are increased. The results showed that the IIDDS is efficient than the TF-IDF in terms of detection accuracy, decisive rate threshold, response time, abnormal user count, and false alarm rate. The response time and the false alarm rate are reduced, whereas the detection accuracy and decisive rate threshold are increased. Further, in order to improve the detection accuracy of the system, the IIDDS will be enhanced using cosine similarity score rather than using Hellinger distance.

References

/1/ M. Uddin, A. A. Rahman, N. Uddin, J. Memon, R. A. Alsaqour, and S. Kazi, "Signature-based Multi-Layer Distributed Intrusion Detection System using Mobile Agents," *IJ Network Security,* vol. 15, pp. 97-105, 2013.

/2/ I. A. Saeed, A. Selamat, A. M. Abuagoub, and S. B. Abdulaziz, "A Survey on Malware and Malware Detection Systems," *analysis,* vol. 3, pp. 13-17, 2013.

/3/ I. A. Saeed and A. Selamat, "Multi-agent Architecture with Dynamic Model Checking for Malware Detection," *LabuanSchool of Informatics Science,* vol. 15, p. 47, 2013.

/4/ S. N. SAI and D. B. V. R. REDDY, "Detecting Malware Behavior in DTN’s Networks using Dogmatic and Look-Ahead Approaches," 2015.

/5/ A. A. Thu, "Integrated Intrusion Detection and Prevention System with Honeypot on Cloud Computing Environment," *International Journal of Computer Applications,* vol. 67, 2013.

/6/ S. Chadli, M. Saber, M. Emharraf, and A. Ziyyat, "A new model of IDS architecture based on multi-agent systems for MANET," in *Complex Systems (WCCS), 2014 Second World Conference on*, 2014, pp. 252-258.

/7/ X. Meng and Y. Li, "A novel verifiable threshold signature scheme based on bilinear pairing in mobile ad hoc network," in *Information and Automation (ICIA), 2012 International Conference on*, 2012, pp. 361-366.

/8/ D. Sharma, V. Kumar, and R. Kumar, "Prevention of Wormhole Attack Using Identity Based Signature Scheme in MANET," in *Computational Intelligence in Data Mining—Volume 2*, ed: Springer, 2016, pp. 475-485.

/9/ P. Casas, J. Mazel, and P. Owezarski, "Unsupervised network intrusion detection systems: Detecting the unknown without knowledge," *Computer Communications,* vol. 35, pp. 772-783, 2012.

/10/ M. Panda, A. Abraham, and M. R. Patra, "A hybrid intelligent approach for network intrusion detection," *Procedia Engineering,* vol. 30, pp. 1-9, 2012.

/11/ R. Akbani, T. Korkmaz, and G. Raju, "EMLTrust: an enhanced machine learning based reputation system for MANETs," *Ad Hoc Networks,* vol. 10, pp. 435-457, 2012.

/12/ M. Khouzani, S. Sarkar, and E. Altman, "Maximum damage malware attack in mobile wireless networks," *Networking, IEEE/ACM Transactions on,* vol. 20, pp. 1347-1360, 2012.

/13/ M. S. Hoque, M. Mukit, M. Bikas, and A. Naser, "An implementation of intrusion detection system using genetic algorithm," *arXiv preprint arXiv:1204.1336,* 2012.

/14/ L. Mechtri, F. D. Tolba, and S. Ghanemi, "MASID: Multi-Agent System for Intrusion Detection in MANET," in *Information Technology: New Generations (ITNG), 2012 Ninth International Conference on*, 2012, pp. 65-70.

/15/ R. Selvaraj, V. M. Kuthadi, and T. Marwala, "Honey Pot: A Major Technique for Intrusion Detection," in *Proceedings of the Second International Conference on Computer and Communication Technologies*, 2016, pp. 73-82.