Intrusion Detection System using Datamining Based Enhanced Framework

S. Suganthi Devi

Abstract-With the significant increase in the use of computers over the network and the development of applications on different platforms, the focus is on network security. The identification of multiple attacks is actually an important element of network security. The role of the IDS is to track and prevent unauthorized use or damage to network resources and systems. An intrusion detection system using Datamining Based Enhanced Framework (DEF) is presented in this paper. The model is assisted by the K-mean Clustering and Decision Tree (DT) classification techniques in which genetic algorithms (GA) for clusters, max runs and confidence can be used. The experimental results shows the promising outcome of the proposed Datamining Based Enhanced Framework (DEF).

Keywords: Datamining, Network Security, Genetic Algorithms, Intrusion Detection System (IDS)

I. INTRODUCTION

Intrusion takes place in a fraction of seconds in this modern world. Intruders use the updated command version intelligently and erase their footprint in audit files and logfiles. Successful IDS distinguish intrusive and non-intrusive documents intellectually In 1980 James Anderson implemented IDS for the first time [1,2]. Many devices have safety violations that can quickly make them vulnerable and unable to be addressed. In addition, extensive work has been carried out on intrusion detection methods, which are still considered incomplete and not a complete method for intrusion [3].

S. Suganthi Devi
Lecturer, Department of Computer Engineering, Srinivasa Subbaraya Polytechnic College, puthur, Nagappatinam, Tamilnadu, India. [Deputed from Annamalai University]
Email:suganthidevi@yahoo.com

The role of network administrators and security experts also has become a highly priority and challenge So more stable systems can not be replaced. Data is processed and sent to eliminate the noise for pre-processing [4,5]; irrelevant attributes and missing are replaced. The preprocessed information will then be analyzed and graded according to their severity. If the record is fine, no more changes are needed or the document is sent for processing [6].

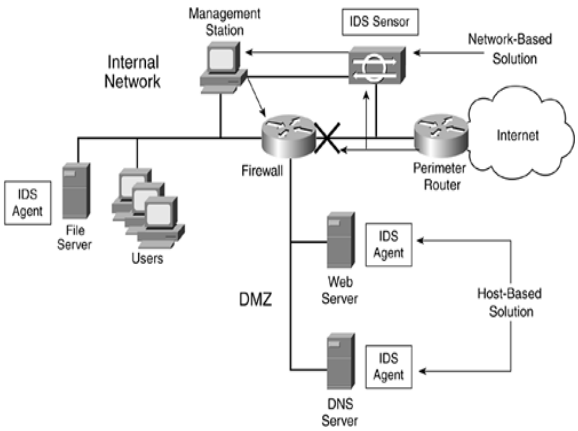

Figure 1. General Structure of Intrusion Detection System

Figure 1 shows the general structure of intrusion detection system. IDS based on data mining can accurately classify such user data and also estimate future results [6,7]. In the IT sector and culture, data mining or the discovery of information throughout data bases has gained much attention. Data mining is used to extract useful information from large volumes of noisy, volatile and dynamic data [8,9]. Intrusion detection systems (IDSs), which are now being built and strengthened by universities, research and science centres, have attracted the interest of a significant proportion of academics [10,11].

An IDS is a system for computer and network security management that tracks and identifies damaged, abusive, suspicious device or networking [12,13]. Lack of security can of led to losses of performance, defects or even temporary failures. Besides intrusion prevention tools and firewalls, IDSs often require complete protection in computer systems. Clustering algorithm groups together attributes that have similar effects. This is an unsupervised learning form. K-means is the most widely used and simplest partitioning algorithm between clustering algorithms.Which have following processes.

- K points are selected as a cluster center randomly, e.g. a centroid.
- A cluster with its centroid closest to this point shall allocate each point to a dataset.
- A centroid shall be determined for each cluster following assignment of all points to different clusters (mean of each).
- Phases 2 and 3 are iterated until no changes in the centroids are detected

The data sets shall be divided into train and test data sets in classification algorithms. These algorithms are called supervised learning algorithms [14], as all record labels are specified and the order management of the particular label type is for the benefit of the algorithm based on other records. DT classifications are

Revised Manuscript Received on November 08, 2019.
S.Suganthi Devi, lecturer in Srinivasa Subbaraya Polytechnic College, puthur, Nagappatinam, Tamilnadu

completely represented by the tree-like representation of information [15]. The range of approaches to DT generation is focused on the different collection methods and storage of DT generation parameters. In this study, an improved model is presented to enhance the new ensemble clustering method [16,17].

II. LITERATURE SURVEY

Intrusion and attack techniques are now more advanced, with large amounts of network traffic information and complex behavior overcoming conventional intrusion detection (IDS). High detection reliability, high false positives rates and high runtime benefit from the existing cloud IDSs. This paper introduces a distributed intrusion detection architecture focused on machine learning for cloud environments. The proposed system is intended to run alongside the edge network components of the service provider in the cloud. The incoming network traffic can thus interrupt network routers on the physical layer [18]. Used to make the preprocessing of traffic on a cloud router possible by the use of the Sliding Windows Algorithm (SWA) to the Naive Bayes Classification System. This anomaly detection module has a set of Hadoop and MapReduce commodity database nodes which can be used when the congestion in the network increases. Anomaly network traffic data is synchronized to a central storage database on each side of the router for each session. The next step to determine the type of each attack is to conduct a final classification stage based on the Random Forest.

Security systems can detect unusual activities or behaviors on network systems. In many unusual acts, however, conventional safety systems including firewall and anti-virus do not function well. Precise and smarter intrusion detection systems (IDSs) are required to solve this problem. Different techniques and strategies to resolve IDS shortcomings such as high false alert rates, poor reliability, and time consuming have been introduced in recent decades. A Hybrid Intrusion Identification (HII) method [19] based on DT and KNN is proposed in this paper. A functional selection method is used to extract structured information from NSL-KDD data set to improve the performance of the proposed approach.

The Intelligent Intrusion Detection system (IIDS), an essential detection measuring to protect data integrity and the network availability of attacks [20], is a requested system based on sophisticated algorithms. It's an intrusion detection combination of software and hardware.

An intrusion detection system using Datamining Based Enhanced Framework (DEF)

In this article, the design is based on k-mean algorithms and the decision-tab. It is therefore important to determine the number of Ks (cluster number) and total runs (completed runs). Against this context. The DT algorithm was also useful in data mining. To this end, after initial pre-processing of the intended data set, a genetic algorithm (GA) was used in order to improve clustering. Clusters have been assigned to cluster members as labels It revised and eventually used the dataset in the process of classification. For data identification, the modified data set containing cluster labels is then used. The GA has been used to refine the confidence parameter and therefore to improve the performance of the DT classification system.The result was presented to the proposal model, which may be utilized in marked and unscripted datasets, i.e. supervised or unsupervised approaches, in view of the implementation in this system of clustering and classification techniques. The suggested framework scheme is illustrated in figure. 2.

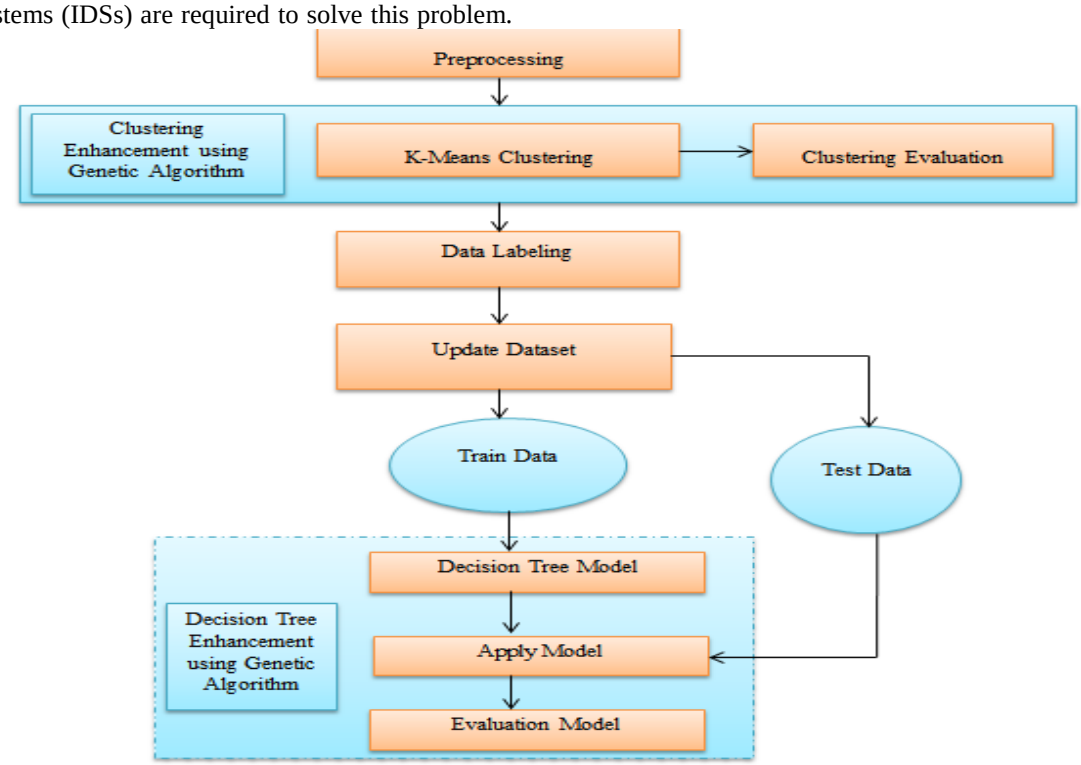

Figure 2. Datamining Based Enhanced Framework Architecture

1. Preprocessing

Upload the data set: The desired data set is uploaded in this step to be used for datamining.

Features selection: In step 1 the expected data set features are selected. My selection basis involves a general feature and a sub-set of features.

Feature roles are described in this step: ' Roles ' determines the feature identity, whether it is normal, unique, labelled, etc.

Conversion of nominal data to numerical data: nominal data in the clustering method must be translated to numerical.

Normalization: Feature values based on Z-transformation method will be normalized in this step.

2. K-Means Enhancement

As unlabeled data and unsupervised structures have been assumed in the proposed model, clustering for information clusters has been used after initial preprocessing, which is why i is an algorithm for clustering with eq(1) Euclidean distance.

$$d_F(a,b) = \sqrt{\sum_{i=1}^{k}(a_i - b_i)^2} \quad (1)$$

Where i is the number of dimensions of the problem clustering, i.e. the feature number.Therefore, a_i and b_i refer to the ith features of a and b.

By determining the number of clusters and the algorithm limit, the performance of cluster is significantly improved. That's why parameters such as t and max runs are optimised using GA. The capabilities of GA require parametric improvements.

Equation (2) demonstrates the fitness function for the cluster enhancement of k-means in the GA:

$$FF = Max\ Avg_{centroid\ \ distance} \quad (2)$$

FF=Fitness function

Where centroid distance is the distance from each other in the cluster centres. Increasing the distance between the clusters and the better value of the cluster.

3. Data Labeling & Dataset Updating

The proposed structure for supervised and unsupervised approaches in this report has been assumed to be applicable. The optimized clustering approach has been used for clustering data in the previous section. As a cluster members were tagged at this point and the data set up was subsequently revised as the architecture for classification model and system intrusion detection is subject to the specified labels for each data set. Members of the cluster are known by the names of the clusters.

4. Sub-Dataset Train & Test

After changes to the primary data collection, train and test subsets have been removed. For this reason, the updated data set of train and test subsets will increase to 70 and 30 percent respectively. The approach has been used to select various types of data, e.g. sampling without placement. The separation of data is done on the basis of eq(3-5).

D=Dataset, tr= train, ts= test

$$D = \{tr_{data} \cup ts_{data}\} \quad (3)$$

$$\{tr_{data}\} = 70\%\ of\ D \quad (4)$$

$$\{ts_{data}\} = \{D - \{tr_{data}\} | \{tr_{data} \cap ts_{data}\} = \emptyset \quad (5)$$

5. Enhancement of Decision Tree

The DT has analyzed the applied data on the system to identify data from top down. There was a feature label for each node in the DT. When a new node was formed at a certain point, a function was chosen to maximize its division power according to the records allocated to a particular sub-tree. The Gini index determined this splitting power. The Gini coefficient is a way to calculate an impurity of nodes, as eq(6) shows:

$$GINI(x) = 1 - \sum_k Q(k|x)^2 \quad (6)$$

$Q(k|x)^2$ is the sum of class k of node x record numbers to all node data. The worst division occurs when the less data in every category of the specified node is obtained in equal amounts. The more ideal and homogeneous is the larger the Gini coefficient of the node.

Confidence is one of the key parameters of the DT algorithm as the truncation error is observed. The DT algorithm GA function is shown in eq(7):

$$FF = Accuracy - Clasification\ Error \quad (7)$$

III. RESULTS AND DISCUSSION

Many devices have safety violations that can quickly make them vulnerable and unable to be addressed. In addition, extensive work has been carried out on intrusion detection methods, which are still considered incomplete and not a complete method for intrusion. The role of network administrators and security experts also has become a highly priority and challenge So more stable systems can not be replaced. So better performance is need for each intrusion detection methods. The proposed Datamining Based Enhanced Framework (DEF) provides better performance compared to other existing methods. Figure 3 shows the performance ratio of Datamining Based Enhanced Framework (DEF) .

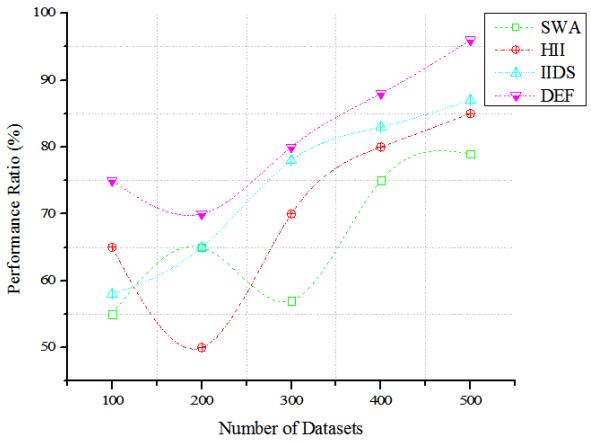

Figure 3. Performance Ratio

When specifying the number of classes and the maximum number of algorithms, the performance of clusters is significantly increased. Therefore, parameters such as i and Max runs are optimised with the GA. With the GA capabilities, parametric changes are possible. The proposed Datamining Based Enhanced Framework (DEF) provides high accuracy compared to other existing methods. Figure 4 shows the Accuracy of Datamining Based Enhanced Framework (DEF) .

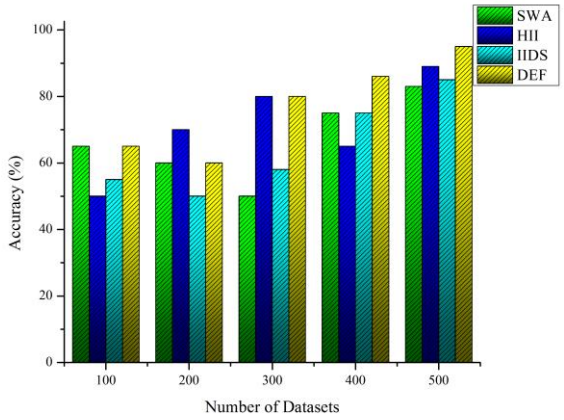

Figure 4. Accuracy

Different techniques and strategies to resolve IDS shortcomings such as high false alert rates, poor reliability, and time consuming have been introduced in recent decades. The proposed Datamining Based Enhanced Framework (DEF) provides high precision rate compared to other existing methods. Figure 4 shows the precision ratio of Datamining Based Enhanced Framework (DEF) **.**

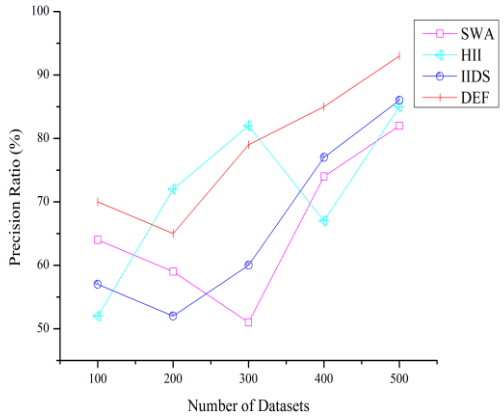

Figure 5. Precision Ratio

An IDS is a system for computer and network security management that tracks and identifies damaged, abusive, suspicious device or networking. Lack of security can of led to losses of performance, defects or even temporary failures. The proposed Datamining Based Enhanced Framework (DEF) provides high recall rate compared to other existing methods. Figure 6 shows the recall ratio of Datamining Based Enhanced Framework (DEF) **.**

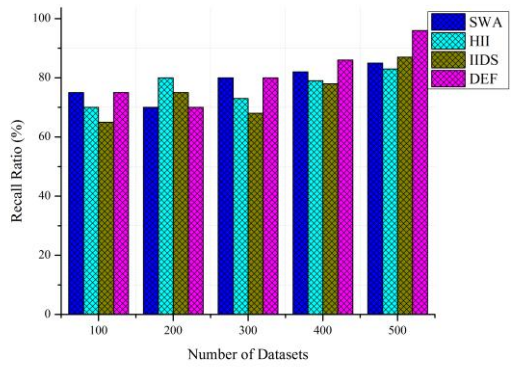

Figure 6. Recall Ratio

IV. CONCLUSION

In this research a method focused on optimizing Kmean's clusters and DT classification algorithms is proposed to detect network intruders in a greater precise manner. For this feature the GA and the higher accuracy of the used classifier optimize total runs and trust. One of the most important characteristics is the applicability of this method for both supervised and unsupervised approaches. The idea is to use the suggested model to test its generalization for future research in the field of network intrusion monitoring in different datasets.

REFERENCES

1. Liao, H. J., Lin, C. H. R., Lin, Y. C., & Tung, K. Y. (2013). Intrusion detection system: A comprehensive review. Journal of Network and Computer Applications, 36(1), 16-24.
2. Ko, C. W., & Rho, J. (2004). U.S. Patent No. 6,789,202. Washington, DC: U.S. Patent and Trademark Office.
3. Moffie, M., Kaeli, D., Cohen, A., Aslam, J., Alshawabkeh, M., Dy, J., & Azmandian, F. (2014). U.S. Patent No. 8,719,936. Washington, DC: U.S. Patent and Trademark Office.
4. Lin, C. C., & Wang, M. S. (2008). Genetic-clustering algorithm for intrusion detection system. International Journal of Information and Computer Security, 2(2), 218-234.
5. Fu, T. C., & Lui, C. L. (2007). Agent-oriented network intrusion detection system using data mining approaches. International Journal of Agent-Oriented Software Engineering, 1(2), 158-174.
6. Zhang, J., Huang, M. L., & Hoang, D. (2013). Visual analytics for intrusion detection in spam emails. International Journal of Grid and Utility Computing, 4(2-3), 178-186.
7. MuhammedShafi. P,Selvakumar.S*, Mohamed Shakeel.P, "An Efficient Optimal Fuzzy C Means (OFCM) Algorithm with Particle Swarm Optimization (PSO) To Analyze and Predict Crime Data", Journal of Advanced Research in Dynamic and Control Systems, Issue: 06,2018, Pages: 699-707
8. Xiang, C., Yong, P. C., & Meng, L. S. (2008). Design of multiple-level hybrid classifier for intrusion detection system using Bayesian clustering and decision trees. Pattern Recognition Letters, 29(7), 918-924.
9. Porouhan, P., & Premchaiswadi, W. (2017). Pattern mining and process modelling of collaborative interaction data in an online multi-tabletop learning environment. International Journal of Knowledge Engineering and Data Mining, 4(2), 114-144.
10. Baskar, S., Periyanayagi, S., Shakeel, P. M., & Dhulipala, V. S. (2019). An Energy persistent Range-dependent Regulated Transmission Communication Model for Vehicular Network Applications. Computer Networks.https://doi.org/10.1016/j.comnet.2019.01.027
11. Aljawarneh, S., Aldwairi, M., & Yassein, M. B. (2018). Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model. Journal of Computational Science, 25, 152-160.
12. Shakeel, P. M., Arunkumar, N., & Abdulhay, E. (2018). Automated multimodal background detection and shadow removal process using robust principal fuzzy gradient partial equation methods in intelligent transportation systems. International Journal of Heavy Vehicle Systems, 25(3-4), 271-285
13. Yin, C., Zhu, Y., Fei, J., & He, X. (2017). A deep learning approach for intrusion detection using recurrent neural networks. Ieee Access, 5, 21954-21961.
14. Aloqaily, M., Otoum, S., Al Ridhawi, I., & Jararweh, Y. (2019). An intrusion detection system for connected vehicles in smart cities. Ad Hoc Networks, 90, 101842.
15. Le, A., Loo, J., Lasebae, A., Aiash, M., & Luo, Y. (2012). 6LoWPAN: a study on QoS security threats and countermeasures using intrusion detection system approach. International Journal of Communication Systems, 25(9), 1189-1212.
16. Ramdane, C., & Chikhi, S. (2014). A new negative selection algorithm for adaptive network intrusion detection system. International Journal of Information Security and Privacy (IJISP), 8(4), 1-25.

17. Kavitha, B., Karthikeyan, S., & Maybell, P. S. (2012). An ensemble design of intrusion detection system for handling uncertainty using Neutrosophic Logic Classifier. Knowledge-Based Systems, 28, 88-96.
18. HyungGeun, O. H., Seung-Hyun, P. A. E. K., Cheolho, L. E. E., & Lee, D. (2012). U.S. Patent No. 8,181,248. Washington, DC: U.S. Patent and Trademark Office.
19. Idhammad, M., Afdel, K., & Belouch, M. (2018). Distributed intrusion detection system for cloud environments based on data mining techniques. Procedia Computer Science, 127, 35-41.
20. Foroushani, Z. A., & Li, Y. (2018, February). Intrusion Detection System by Using Hybrid Algorithm of Data Mining Technique. In Proceedings of the 2018 7th International Conference on Software and Computer Applications (pp. 119-123). ACM.

AUTHORS PROFILE

S.SUGANTHI DEVI, B.E. (CSE) from E.G.S. Pillay Engineering College, Nagapattinam, from Bharathidhasan University, Tamilnadu, India. M.E. (CSE) and Ph.D. (CSE) from Annamalai University, Tamilnadu, India. She has published 11 papers in conferences and journals. She was working as Assistant professor, Department of Computer Science and Engineering in Annamalai University (2007 – 2017). On 2017, she was deputed from Annamalai University. So, Now she is working as a lecturer in Srinivasa Subbaraya Polytechnic College, puthur, Nagappatinam, Tamilnadu