NVEO
Natural Volatiles & Essential Oils

Network Anomaly Detection using Data Mining Algorithms

Dr.B.Radha[1], D.Sakthivel [2]

[1] Associate Professor, Department of Information Technology, Sri Krishna Arts and Science College, Coimbatore – 641105 , Tamilnadu, India

[2] Ph.D Scholar, Department of Computer Science, Sree Saraswathi Thyagaraja College, Pollachi - 642107, Tamilnadu, India

[1]radhakbr10@gmail.coml [2]sakthivel.it13@gmail.com

Abstract

Anomaly detection is an area that is currently being explored. The purpose of this paper is to clarify a network anomaly detection framework that uses K-Means clustering and SVM classification to discover network characteristics in the NSLKDD dataset in order to reduce the false alarms rate, and further develop the positioning rate and identify the zero-day attacker. After preparing and testing the proposed data mining algorithm, the results show that the proposed method (K-Mean + SVM) has achieved a positive detection rate of (95.32%) and reduced the false alarm rate to (1, 1%), and reached (87.33%).).

Keywords: anomaly detection, SVM, k-means

Introduction

Intrusion detection is the interaction of identifying harmful examples in a large amount of information gathering, including data mining, artificial intelligence, and dataset framework technology. The purpose of the intrusion detection test guide is to save the garbled client data. The basic goal of intrusion identification is to save misrepresented customer information. There are a large number of information storage measures in different organizations, such as records, files, images, sound recordings, sound recordings, logical information, and many new information designs related to human existence. Intrusion detection is used to discover fuzzy examples, legitimate examples, and connections in a large collection of information. Likewise, it dissects and predicts suspicious exercises, and uses different types of restrictions to examine information and combines inspection, sorting, and grouping techniques. Intrusion detection framework is an interaction that identifies unauthorized use of PC or media transmission network [1] has the ability to summarize intrusion or threat associations. An ID specifically recognizes three types of PC attacks:

1. Scanning attacks
2. Denial of service (DOS) attack
3. Penetration attack

The IDS is used to transmit an alert message when there is a suspicious example of attack information. We need to plan an IDS model with high competition and low false alarm rate. Three basic strategies are used in IDS: abuse-based, anomaly-based, and hybrid-based. Signature-based programs are designed to identify known attacks using the signatures of those attacks. It has a higher recognition rate and a lower false alarm rate. The exception-based method is to store the client's normal behavior in the database and compare it with the user's current behavior. If there is a huge difference, it means something is wrong or strange. It has a high false alarm rate. Hybrid-based technology is the combined use of signature-based and exception-based programs. Whenever IDS finds a security hazard in the frame, it creates an alert to show that there is a disruption. IDS was first proposed by James Anderson in 1980, and it has become the most necessary and proven business for network leaders and security experts. We can plan and assemble a scanner to look for malicious examples and unapproved information. IDS can identify and terminate unauthorized attacks on organizations [2]. In this article, one way to approach the combined data mining program for clustering and characterization (K-Mean clustering and SVM classifier) is to familiarize yourself with the boundary representation of the updated intrusion detection framework and reduce the speed of False Alarms and False Alarms - Discover and accurately identify the alarm rate of new attacks, gradually distinguish intrusions, and use the implemented attack design to expand the detection rate in the training phase [8].

Related Work

D. Sakthivel, Dr. B. Radha, Security is a big issue for all networks in today's business environment. Hackers and intruders have made many successful attempts to disrupt corporate networks and recognized web services. Many methods have been developed to protect network infrastructure and communications on the Internet, including the use of firewalls, encryption, and virtual private networks. Intrusion detection is a relatively new addition to this type of technology. Intrusion detection methods have started to appear in recent years. With intrusion detection methods, you can collect and use information from known types of attacks and find out if someone is trying to attack your network or a specific host. Snort is a free and open source Network Intrusion Detection System (NIDS). NIDS is an intrusion detection system (IDS) that is used to scan data flowing on the network. There are also host-based intrusion detection systems, which are installed on a specific host and only detect attacks against that host. Although all intrusion detection methods are still new, Snort is among the best systems available today.

Rich son and so on. [7] Use One-Class SVM to detect abnormal conditions in solar photovoltaic power plants. The research aims to enable factories to operate safely, reduce operating costs, increase operating profits, and maintain reliability. The SVM method is used to measure the difference between normal and abnormal characteristics; this is compatible with the SVM function which can handle non-linear data. Yehaya et al. [8] He conducted research related to activities of daily living (ADL), which focused on a person's sleep patterns. Through SVM, unmarked data sets can be marked as 1 for normal data and 1 for abnormal data. Devi et al. [9] Adopting the concept of anomaly detection, using SVM combined with the Tomek hook method to deal with class imbalance.

has reported the use of the KDD Cup 99 dataset for intrusion detection research. Jin et al. specifically for advanced persistent threats, and proposed a deep neural network (DNN) [13], using 100 hidden units, combined with the modified linear unit activation function and the ADAM optimizer. Its method is implemented on the GPU using Tensor Flow. [14] Papamartzivanos et al. A new approach is proposed that combines the benefits of Sparse AE and the MAPEK framework to provide scalable, adaptive, and self-contained misuse IDS. They merged the data sets provided by KDD Cup 99 and NSLKDD to create a huge data set [3]. The model proposed in

[17] is an integrated intrusion detection system that combines Chi-square as a feature selection technology and a hybrid model of base classifier and integrated classifier. In the literature [18], an ensemble model is used that combines multiple classifiers to detect DDoS attacks, thus obtaining ideal results in the NSLKDD dataset. The literature [19] combines network intrusion detection with support vector machines based on the introduction of the whale algorithm for optimization, thus improving the precision of intrusion detection.

Although these detection methods combined with machine learning have achieved satisfactory experimental results, their algorithms are still subject to certain limitations. For example, the premature convergence of the genetic algorithm, the parameter selection problem in the SVM algorithm [19] and the neural network training data set problem. Also, for other algorithms, the convergence speed is too slow and it is easy to fall for local optimization. These problems must be solved using improved algorithms.

Network Anomaly Detection System

The basic principle behind IDS is to monitor whether there is an attack on the network system. The attack can be prompted by the simplest method possible, adjusting the user name, or it may be an exciting attack, including the arrangement of traversing multiple systems. Figure 3a shows the general design of IDS [16]. It has been set to midway so that all packages appear and be taught absurdity. Data is accumulated and sent to the pre-run to eliminate clutter; unnecessary and missing attributes are replaced. At that time, the pre-processed data will be reviewed and collected according to its severity. If the log is regular, you don't need to bother to make further changes, and you don't need to send it before the report time to trigger an alert. Taking into account the state of the data, an alert is generated so that the administrator can manage the situation early. Display attacks to allow data to be collected from the network.

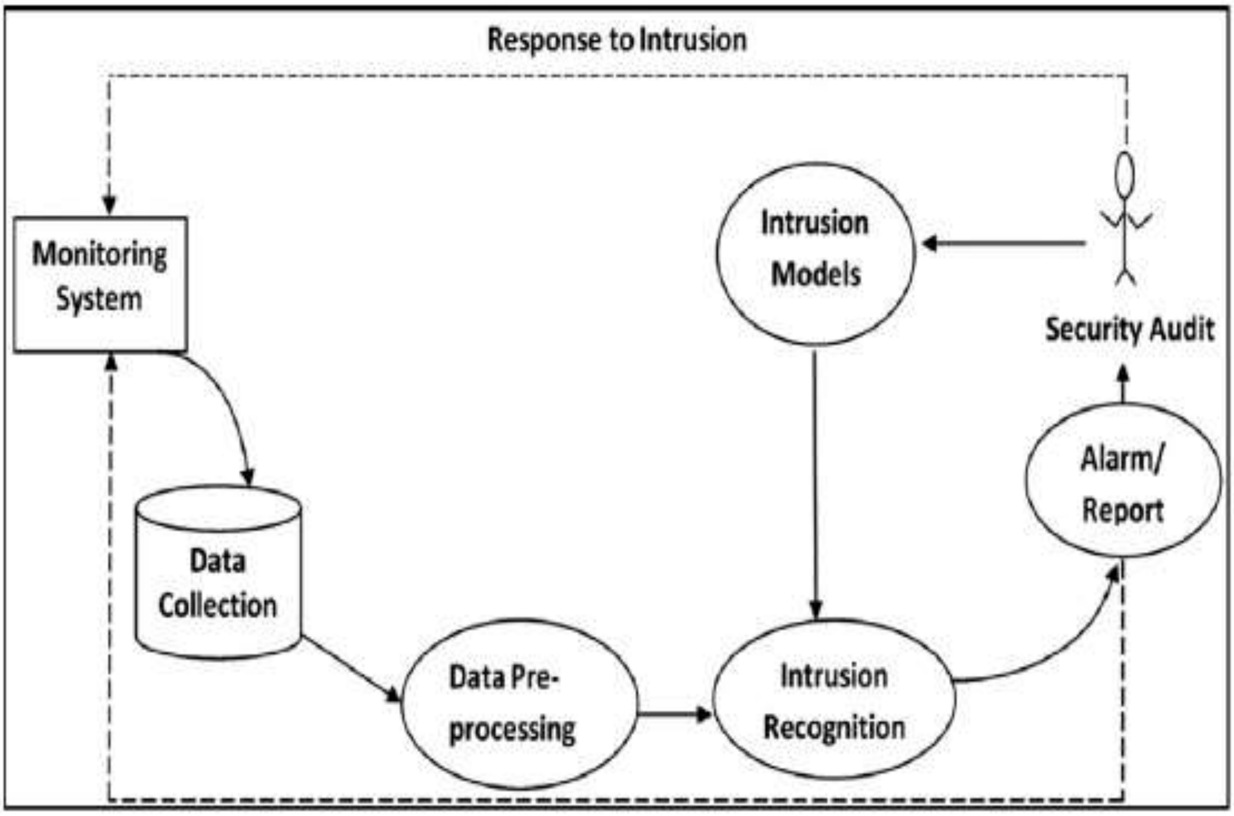

Figure3a. IDS

In this part, a new method that relies on k-Means clustering algorithm and support vector machine (SVM) anomaly detector is proposed to detect anomalies in online networks. Figure 3b. The proposed method aims to make a reasonable number of localizations. The device has high and unique test accuracy. The intrusion detection system (IDS) of the current frame of the train is fundamentally divided into three parts. The initial part is data classification, which mainly collects frame log data or network traffic information from the switch to the basic train hub. The next part is information scanning, which measures the collected information, builds detection models, and identifies interrupt practices. The third part is the response of the framework. The proposed framework can be compared. In the above three parts, the data analysis module is the center of the intrusion detection system (IDS), and the interrupt location innovation is the center of the data analysis module.

A) Preprocessing:

Preprocessing the unique NSLKDD [11] intrusion data set is an important stage, making it a suitable contribution to the ordering stage. The main goal of the preprocessing stage is to reduce suspicions and provide accurate information to the detection engine. The preprocessing stage cleanses the organization's information through aggregation and naming, and deals with missing or split data sets. Preprocessing of the data set is accomplished by successively applying the accompanying steps.

B) Feature selection:

Feature selection is the most basic stage of building an intrusion detection model, and it is also essential to improve the proficiency of information mining algorithms. Generally speaking, the information of the classifier can be found in the feature space of height measurement, but not the entire feature is related to the class to be characterized [12]. Part of the data incorporates invisible, excessive or noisy features. In the current situation, boring and unnecessary features may introduce noisy data that interferes with the learning algorithm. Reduce the number of attributes, eliminate unimportant, noisy, or repetitive layouts, and achieve application impact, such as accelerating data mining algorithms, further improving learning accuracy, and improving model comprehensibility. In this process, the arrangement of features or elements considered to be the best attributes is separated to construct an appropriate detection framework [13].

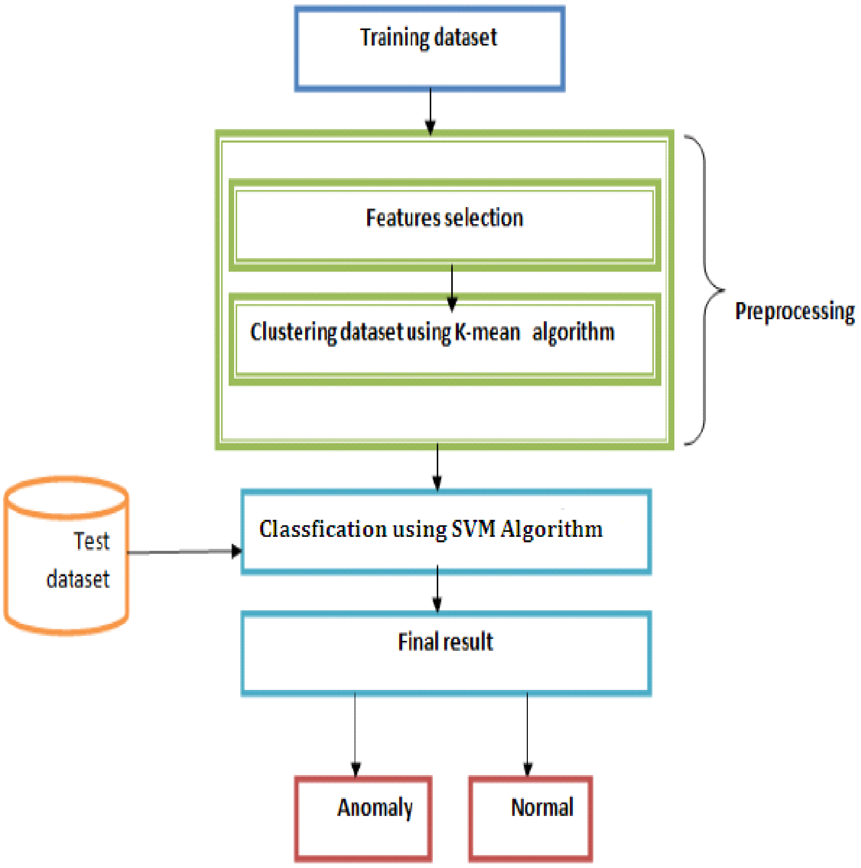

Figure 3b: Network Anomaly Detection

The purpose of selecting function is to set an acknowledgment rate and reduce the false alarm rate in network intrusion detection. Matlab is a statistical machine learning device that has been used to calculate the item determination subset of the cross method (K-Mean + SVM) to test order execution on each feature set.

The genetic search algorithm was applied to select explicit characteristics from the data set and eliminate those that are not important before the grouping and classification stage. The result is shown in Figure 3c:

C) Clustering stage:

In the clustering stage, K means The Clustering algorithm is applied, it is signaled and two clusters are produced [15]. When the algorithm is emphasized through preparation and training data, each group plans to move to another. The revitalization of the cluster changes the potential benefits of the center of mass. This change is an impression of the current cluster segmentation. When there is no movement in any cluster, the K-Means computing cluster becomes absolute.

D) Classification

This test attempts to perceive normal data and strange data. It has been found that the intrusion detection framework is basically an example identification problem, that is, a classification problem [3] [7]. It needs to be able to accurately investigate the categories in which the input data appears. Under this premise, the emission model is established as follows [9]

$$\max K = \omega_1 Accuracy + \omega_2 Spacificity \qquad (1)$$

$$\mathrm{st.T\ (n)} < t \qquad (2)$$

$$Sensitivity \geq A_3 \qquad (3)$$

Among them, the absolute instance range of the effective sorting of Precision classification instances (%);

The identifiability test of special abnormal data (%);

All positive sorts of sensitivity sorting The scope of the model (%).

Constraint (2) expresses the limitation of algorithm execution time; Constraint (3) proposes normal data transfer requirements. Can decipher the binary classification of normal data and irrelevant data. Because the support vector machine has high accuracy and strong speculation ability for the two classification problems, it can deal with the sample irregularities of normal data and abnormal data, so it relies on the support vector machine for anomaly detection and can capture irrelevant data.

Experiment

Matlab is a tool for data mining and machine learning, which has been implemented [11].

A) Data set description:

NSL KDD is a refined version, also known as KDD CUP data set replacement. It consists of many necessary attributes of the KDD CUP data set. It is open source data and can be downloaded without any problems. The advantage of using this data set is that cumbersome records are eliminated, and a sufficient number of records can be accessed for testing and training data. It consists of 41 functions, organized in nominal, binary and digital. An attribute is added as a class, it is the 42nd attribute. There are two categories, called Normal and Anomaly. The exception category can also be divided into DOS, PROBE, R2L, and U2R [14].

B) Anomaly Detection Model Based on Support Vector Machine

Support vector machine is more likely to circulate high-dimensional and non-linear information test characterization problems, and has high-order precision and a reliable speculation ability [20], so it can solve this problem and clumsiness of the test [10]. Chapter

Xi € Rd} is directly inseparable. Nonlinear programming is expected ω:

X> H; X € Rd, H € R, planning from the Euclidean space to the Hilbert space H, so that the information index can be separated directly in the H space

$$max \sum_{l=1}^{n} \alpha i - \frac{1}{2} \sum_{l=1}^{n} \sum_{l=1}^{n} \alpha i\ \alpha j y i y j \varphi(x_i)\phi(x_j) \qquad (4)$$

$$S.t \sum_{l=1}^{n} \alpha i\, yi = 0 \qquad (5)$$

$$0\ \leq \varphi i\ \leq C, i = 1, \dots, N \qquad (6)$$

As shown in the previous question, the capacity of the part can be characterized as:

$$\mathrm{K}(\,\mathrm{x}, \mathrm{z}) = \ \varphi(x).\varphi(z) \qquad (7)$$

The bit work characterizes the similarity of the two pieces of information after the scheduling change, and subsequently the characterization is dominated by SVM.

The radial basis function (RBF) part, also known as the Gaussian bit part or square exponent (SE), is generally used in different types of learning calculations, such as Gaussian process regression (GPR) [24]. It is usually characterized as the drone ability from a given focal point in space to the mean Euclidean distance, which is

$$\mathrm{K}(\,\mathrm{x}, \mathrm{z}) = \exp\left(-\gamma \left|\left|x - z\right|\right|^{2}\right)\ \ \gamma > 0 \qquad (8)$$

SVM relies on the anomaly detection model. The specific basic steps are as follows:

Phase 1: Collect information from the continuous organization of the first train, use the information index NSL KDD for fact processing, and obtain Estimates of different significant information. Then, by then, the preparations for standardization and coding have been completed.

Phase 2: Use a simple arbitrary non-replacement check strategy to build a preparation set and a test set from the first information collection.

Phase 3: Set the workload of the part and the limit of each limit in the preparation period, and use the limit vector g as the improvement limit.

Phase 4: Build and solve the secondary planning problem and obtain the layout:

The anomaly detection model dependent on SVM the particular foundation steps are as per the following:

$$\alpha^{*} \ = (\alpha 1^{*}\ , \qquad \alpha 2^{*} \quad , \dots, \alpha n^{*})^{\mathrm{T}} \qquad (9)$$

Phase 5: Solve p* according to support vector x*I and α* i

$$\rho* \ = \sum_{j=1}^{n} \alpha \mathrm{j}^{*} K(xj, xi) \qquad (10)$$

Phase 6: Construct a classification decision function:

$$f(x) = sign\ (\sum_{j=1}^{n} \alpha \mathrm{i}^{*} K(x1, x) - \ \rho*\ ) \qquad (11)$$

Phase 7: Use the obtained model to predict the test set, enter the characterization results and different limits of welfare capacity, and return the obtained welfare value to the limit rationalization measure.

Considering that accuracy and particularity are the main indicators for evaluating the results of the arrangement, the results are evaluated from these two perspectives. Fitness ability_=Yes:

$$K = \omega 1.A_{Test} + \omega 2.B_{Test} \qquad (12$$

Among them, the precision of the ATEST test suite, the specificity of the BTEST test suite.

Results

The computing environment used in the analysis is the supporting framework: CPU: Intel (R) Core (TM) i56300U, Main Stream: 2.40 GHz, Memory: 4 GB, Windows 10 64-bit. Then, at this point, the Matlab replay schedule was used to test the proposed algorithm. The results are shown in Figure 5a and Figure 5c:

Use Matlab to apply the K-Mean calculations to the 22-credit NLSKDD data set (including selection), and the results are shown below: Use the grouping results from K-Mean's data set of 22 attributes.

Accurately classified cases = 72.1877%

Incorrectly classified cases = 27.8123% 4.444

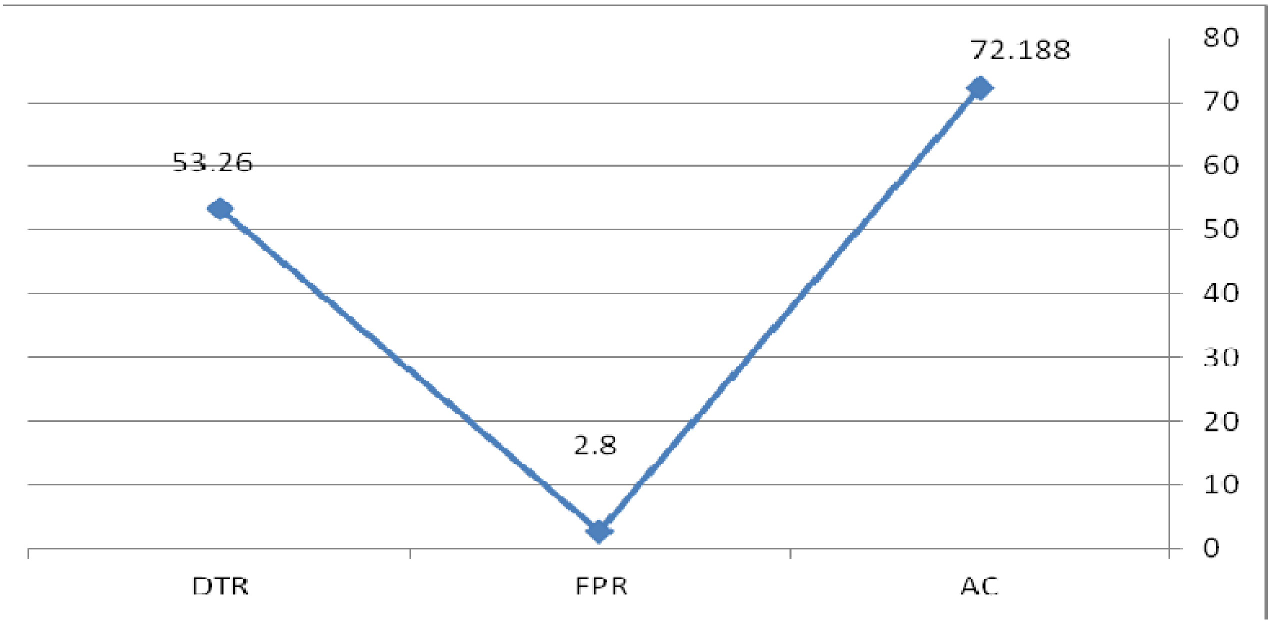

Fig. 5a. Measurement parameters of (K-mean)

According to the test results, it can be well inferred that the accuracy of the preparation group is 92.9%, the emotion degree is 99.6%, the specificity is 72.8%, and the cross-check result is 92.7%; the accuracy rate of the test set is 82.4%, and the emotion is 99.67%, clarity is 56.2814%, and the cross-check result is 87.6%. From the results, it can be well traced that the side-effects of the preparation set identification of the non-optimized support vector machine algorithm meet the requirements, and the accuracy of the test set should be improved in fact.

Use Matlab to apply a hybrid method (K-Mean + SVM) to the NLSKDD dataset, with 22 credits (including options), and the result is as follows:

Instances are effectively classified 21951 = 97.3696%

Instances are misclassified 593 = 2.6304%

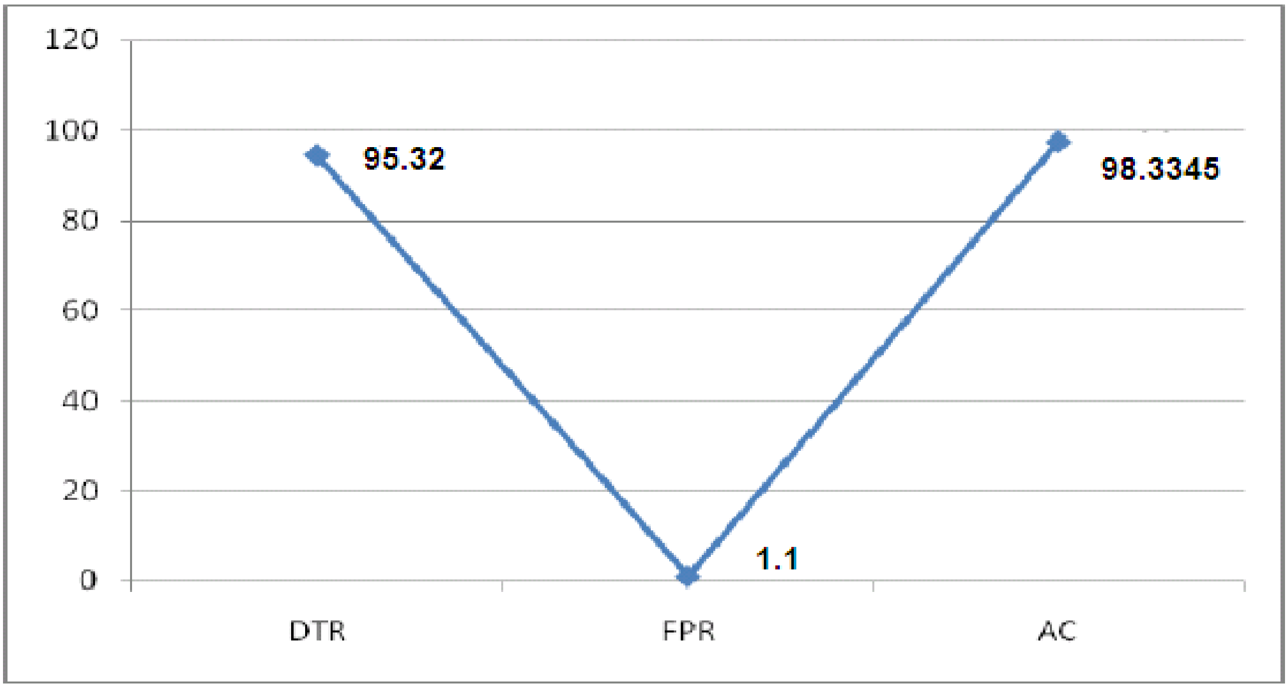

Fig. 5c. Measurement parameters of (K-mean + SVM)

Conclusion

In this article, we propose a cross-species method to process anomaly detection using K-Mean clustering and support vector machine (SVM) classification. The game plan clearly solves the problem of a large number of grade data sets. It uses feature selection in the preprocessing stage to reduce the number of data sets, applies genetic search algorithms to select explicit elements from the NLSKDD data set, and removes those that are not important before the grouping and classification stages, and It also uses K-Mean clustering to reduce Prepare the size of the dataset while staying current. Subsequently, the nature of the findings from the computational selection investigation is adjusted during the programming stage called the support vector machine (SVM). The proposed method (K-Mean + SVM) and the separate calculation of the K-Mean cluster and the SVM cluster are verified. The results show that our method outperforms other methods with a positive recognition rate (95.32%) and reduces the false alarm rate (1.1%). The accuracy rate is high (87.3345%).

With the emergence of information technology innovation, the continuous attack strategy in the dark with intrusion detection function has become more subtle. Certain types of attacks wrap the attack in application layer information, MIB database data, etc., and cannot be identified by the quality of data packets and traffic. Subsequent investigations can reveal the impact of this secret attack on train communication information, equipment, data sets, and different views, and then plan another attack and protection model based on this premise to ensure continued attacks on the train frame.

References

1. D. Sakthivel, Dr.B. Radha, "SNORT: Network and Host Monitoring Intrusion Detection System", International Journal for Research in Applied Science & Engineering Technology (IJRASET),ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887, Volume 6 Issue X, Oct 2018.

2. Nawfal Turki Obeis and Wesam Bhaya, "Review of Data Mining Techniques for Malicious Detection", Research journal of Applied Sciences 11(10):942-947, 2016.

3. D. Sakthivel, Dr.B. Radha, “Novel Study on Machine Learning Algorithms For Cloud Security”, JOURNAL OF CRITICAL REVIEWS, VOL 7, ISSUE 10, 2020.

4. J. F. Xue, Intrusion Detection Technology. Beijing, China: Posts and Telecom Press, 2016.

5. Nimmy Krishnan, Salim A “Machine Learning Based Intrusion Detection for Virtualized Infrastructures”, International CET Conference on Control, Communication, and Computing (IC4), July 05 – 07,2018.

6. Leu, F.Y., Tsai, K.L., Hsiao, Y.T. and Yang, C.T., 2015. An Internal Intrusion Detection and Protection System by using Data Mining and Forensic Techniques.

7. Fouzi Harrou, Abdelkader Dairi, Bilal Taghezouit, Ying Sun, An unsupervised monitoring procedure for detecting anomalies in photovoltaic systems using a one-class Support Vector Machine, Solar Energy, Volume 179, 2019, pp. 48-58, ISSN 0038-092X, https://doi.org/10.1016/j.solener.2018.12.045.

8. Yahaya S.W., Langensiepen C., Lotfi A.,Anomaly Detection in Activities of Daily Living Using One-Class Support Vector Machine. In: Lotfi A., Bouchachia H., Gegov A., Langensiepen C., McGinnity M. (eds) Advances in Computational Intelligence Systems. UKCI 2018. Advances in Intelligent Systems and Computing, vol 840, Springer, Cham, 2019.

9. Nada Aboueata, Sara Alrasbi, Aiman Erbad” Supervised Machine Learning Techniques for Efficient Network Intrusion Detection”, IEEE, 2019

10. Roshan Kumar, Dr. Deepak Sharma,” Signature-Anomaly based Intrusion Detection Algorithm”, Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology , 2018

11. Modi, M.U. and Jain, A. 2016, A survey of IDS classification using KDD CUP 99 dataset in WEKA

12. G. Folino and P. Sabatino, Ensemble based collaborative and distributed intrusion detection systems: A survey. Journal of Network and Computer Applications, 2016. 66: p. 1-16.

A. H. Hamamoto, L. F. Carvalho, L. D. S. Sampaio, T. Abrao, and M.L. Proenca Jr, Network anomaly detection system using genetic algorithm and fuzzy logic. Expert Systems with Applications, 2018.92: p. 390-402.

13. S. Adepu and A. P. Mathur, ``Distributed attack detection in a water treatment plant: Method and case study,'' IEEE Trans. Depend- able Secure Comput., vol. 16, no. 1, pp. 1_14, Jan./Feb.2018,

14. Hari Om and Aritra Kundu, “A Hybrid system for reducing the false alarm rate of anomaly intrusion detection system”, Recent Advances in Information Technology (RAIT), 2012 1st International Conference on 15-17 March 2012, IEEE Publisher.

15. S. Thaseen and C. A. Kumar, ``Intrusion detection model using fusion of chi-square feature selection and multi class SVM,'' J. King Saud Univ. Comput. Inf. Sci., vol. 29, no. 4, pp. 462_472, Oct. 2017, doi: 10.1016/j.jksuci.2015.12.004.

16. S. Mirjalili and A. Lewis, ``The whale optimization algorithm,'' Adv. Eng. Softw., vol. 95, no. 5, pp. 16_27, Jan. 2016, doi: 10.1016/j.advengsoft.2016.01.008.

17. H. Esquivel and T. Esquivel, ``Router-level spam filtering using tcp fingerprints: Architecture and measurement-based evaluation,'' in Proc. 6th Conf. Email Anti-Spam (CEAS). Mountain View, CA, USA: IEEE, 2009, pp. 1_10.

18. J. C. Platt, ``Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods,'' Adv. Large Margin Classifiers, vol. 10, no. 3, pp. 61_74, 1999.

Authors Profile

Dr. B. Radha has completed his PhD. Anna University Computer Science major, selection and programming resources in the grid. He received MCA and Bachelor of Science degrees from Bharathiar University. She is currently an associate professor in the Sri Krishna Department of Information Technology, Art and Science in Coimbatore. He has 16 years of teaching experience in the field of computer science. He has published 35 research articles in UGC Care, Scopus Indexed and Web of Science Indexed Journals.

He has published 5 books in the fields of computer science and information technology.

Mr. D. Sakthivel has completed his PhD. Bharathiar University has a bachelor's degree in computer science and has completed a master's degree in philosophy of using data mining technology for network information extraction from Bharathiar University. He received a master's degree in computer science and BCA from Bharathiar University. His expertise is in data mining, intrusion detection systems and cloud computing. He has 12 years of teaching experience in the field of computer science. He has published 12 research articles in indexed journals UGC Care and Scopus. He completed and earned certified online courses on cloud computing, Python data science, ethical hacking, and IP addressing on the SWAYAM NPTEL, Coursera, and Udemy online learn portals