A SURVEY ON DIVERSE DATA MINING TECHNIQUES IN INTRUSION DETECTION SYSTEM

Dr. Ramalingam Sugumar, Professor & Head, Department Of Computer Science,
Christhu Raj College, Trichy (Affiliated to Bharathidasan University, Trichy)

K. Sivasakthi, Research Scholar, Department Of Computer Science,
Christhu Raj College, Trichy (Affiliated to Bharathidasan University, Trichy)

Abstract: As the internet and technology have gotten universal, there has been an impressive ascent in the quantity of intrusion event. It is exceptionally basic to build up a security strategy around these frameworks. The way toward recognizing unapproved action against computer or organizations is known as Intrusion Detection. It is hard to support computer frameworks refreshed as the quantity of penetrates expands step by step. IDS screens and recognizes uncertain conditions of such frameworks. Interruption location frameworks are worked for recognizing unapproved endeavours to get to or control the computer organizations. IDS gather network information to recognize various types of malware and assaults against administrations and applications. In this paper discuss about data mining based intrusion detection system.

Keywords: Intrusion Detection System (IDS), Classification and Clustering.

Introduction: Intrusion discovery is the way toward checking and examining events that happen in a computer or arranged computer framework. Discovery is completed by dissecting the conduct of clients that contention with the expected utilization of the framework. Any client utilizing a computer will be at some danger of intrusion, despite the fact that the computer isn't associated with the Internet [1]. In the event that the computer is left unattended, any interloper can endeavour to access and attempt to abuse the framework. The issue is significantly more if the computer is associated with an organization, especially the Internet. Any client from around the globe can arrive at the computer distantly. A gatecrasher may endeavour to get to significant private or secret data or dispatch a type of assault to carry the framework to a stop or stop to work viably. An interruption to a computer framework shouldn't be executed physically by an individual. It might be executed distantly and naturally with built programming [2].

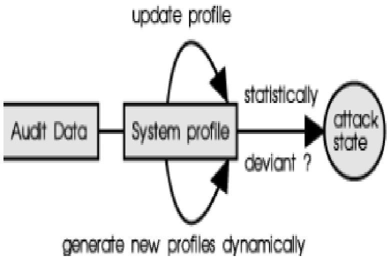
Types of IDS: Intrusion detection systems are of the following categories

a) **Signature Detection:** Signature detection likewise alluded to as misuse detection that distinguishes the attack based on the gradual information from existing attacks. This method can distinguish the attacks which have signatures. The information data set must be refreshed every now and again to keep up state-of-the-art data [3]. Nonexclusive marks can be made to recognize more varieties of a similar attack however it additionally requires decent information on attacks so as to distinguish malevolent attacks yet permit authentic traffic.

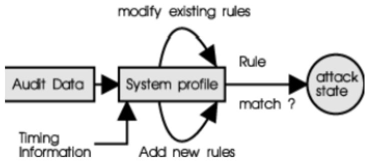
Misuse-Detection

b) **Anomaly Detection:** Anomaly detections separate among genuine and ill-conceived users through powerful ID of deviations from framework use. In this method, there is a quest for confirmation of attacks dependent on gathered information. There is a high likelihood that an intrusion is in progress when there is a strange high CPU load in mix with different measurements. There is a preferred position in recognizing new attacks in this strategy. Nonetheless, the model requires regular up-degree and changes in accordance with mirror the conduct of a typical user and minimizes the false positive rate [4].

Anomaly-Detection



c) **Host-Based IDS (HIDS):** Host-based IDS are sent locally with respect to each host computer and screen the host on which it is fixed. The progressions to various factors on the host computer are consistently observed by HIDS. These incorporate framework measures, library records, CPU usage, document access, review strategies, client records, and occasion logs. An alarm is sent to the manager when the edge surpasses or dubious honesty changes [21]. In this way HIDS can recognize unusual conduct on a framework that might be undermined yet the overseer ought to dissect the yield of HIDS at standard spans and limit the bogus positive cautions.



d) **Network-Based IDS (NIDS):** NIDS is installed on network and test the caught parcels through the network fragment. The IP parcels are dissected for distinguishing the mark assaults. Not at all like HIDS, NIDS can screen a whole organization division and can be conveyed quickly. As HIDS and NIDS are correlative, a large portion of the IDS merchants have fabricated three-tier structures that join both HIDS and NIDS sensors with a brought together comfort [22]. Distant observing of this incorporated worker should be possible for examining the logs, reports, deal with the sensor arrangement, and adjust the interruption location strategy.

Data Mining Based IDS: Data mining has become a noteworthy factor in the advancement of the intrusion detection system. Different data mining systems like association rule mining, clustering, classification, and outlier detections are used for examining network information to acquire data identified with intrusions.

Clustering: The method of marking information and doling out it into bunches is known as clustering. New instance are bunched in related groups by these methods. The two types of clustering are:

- ➢ Pair-wise Clustering
- ➢ Central Clustering

New information cases are appointed based on a separation measure in pair-wise clustering. Central clustering is additionally alluded to as centroid clustering models in each grouping dependent on the centroid. Centroid based grouping calculations are more productive as far as intricacy in examination with pair-wise clustering methods [5].

The basic steps involved in clustering are

- Distinguish the greatest group i.e., the group which has the biggest number of tests and arranges it as normal.
- The rest of the groups are sifted through as per their good ways from the biggest group in climbing request.
- The primary 'K' groups are chosen with the end goal that the all out samples in the groups signify ¼ S, and dole out them as typical where 'S' is the level of ordinary samples.
- Distinguish the rest of the groups as intrusion groups.

Classification: Classification is fundamentally the same as clustering as the methodology targets partitioning information tests into particular areas called classes [6]. Classification necessitates that the expert realizes how classes are characterized early. Each case in dataset is used for building up the classifier that prior has an incentive for the element which is utilized to portray classes. Classifications calculations use the training set for building up the model. The examples are then arranged by the model as would be normal or attacked. The classification consist the following steps.

- Develop a training data set.
- Identify classes and attributes.
- Identify the features required for classification.
- Classify the unknown sample using model.

Some important categories of classification algorithms are:

- Decision tree
- Rule based methods
- Naïve Bayes and Bayesian networks
- Support Vector Machines

Decision Tree: Decision tree is a prescient model that gets sample for perception and predicts the class labels. The leaves indicate class names and the node specify features. They make rules which are understandable for people. These principles help to look for records in the data base. The trees can choose the best attributes by utilizing the property of information gain which gauges the way of isolating the preparation cases into the particular objective group [7]. The element that compares to the most elevated information gain is chosen. The information gain is estimated by entropy.
Given a set 'S' of 'c' results

Entropy(S,I)= S –p (I) $\log_2$ p (I)

Where p (I) is the fraction of 'S' belongs to class I. S represents the entire sample set.

Rule-Based Method: Rule-based strategies are among the underlying techniques used for misuse detection. These procedures convert intrusive circumstances into the standard set for examination against review information. The standard coordinating procedure alarms an interruption at whatever point there is any aberrance [8]. A portion of the standard based strategies are Multicast Intrusion Detection and Alerting System (MIDAS), Intrusion Detection Expert System (IDES), and Next-age Intrusion Detection Expert System (NIDES) (Sebring et al. 1988, Lunt, 1988, Anderson et al. 1995). The restrictions of rule-based methods are:

- Identification of the connection

between rules is a difficult activity.

- Verification of the principles for precision is trying because of the communication between rules.
- Most of the standard bases are old.

Naïve Bayes: Naive Bayes classifiers are utilized for displaying normal and abnormal event (Panda and Patra, 2007). Naïve Bayes depends on the theorem of Bayes and is a supervised learning classifier (Bishop, 2006).

$$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)}$$

Where,
P(X/Y) is the posterior probability of label for the attribute.
P(X) is the preceding probability.
P(Y) is the posterior's probability.

Consequently the probability of event X as to information Y can be controlled by figuring the likelihood of the information Y concerning event X increased by the event X likelihood standardized as to the likelihood of the information Y. This implies the likelihood of an assault on irregular information can be determined by first deciding the likelihood that arbitrary information associated with the assault and afterward increasing it by the particular sort of assault likelihood [9].

The process of Navie Bayes classification work, as per the following:
Each information in the preparation set that was ordered into a class is watched [10]. In the event that the likelihood of the example is known and a presumption of each characteristic being autonomous with indistinguishable dissemination, at that point the arrangement is likewise thought to be conveyed ordinarily and subsequently the assessed estimation of the example mean likelihood is acquired. Consequently it is normal that, for each element in the dataset, the mean worth likelihood is resolved and typical conveyance is used for assessing the persistent qualities.

$$N(x \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$$

The likelihood of characterization is determined by deciding the example mean once more. This is finished by including all the passages in the preparation information that compare to the classification and separating by the aggregate of tests. The likelihood of an occasion having a place with a particular class from the example mean qualities is gotten from the past advances. Consequently this count can be performed for all class marks and the example is distributed to the class with the greatest likelihood.

Support Vector Machine (SVM)
Binary classification can be settled utilizing SVM (Kim and Cha, 2005). A SVM maps straight frameworks into the non-direct space. SVM uses the bit work for planning the information. A hyper-plane is worked by isolating the space utilizing different part works specifically polynomial and spiral premise capacities [11].

SVM yields a general ideal arrangement as it includes a quadratic programming issue. Accept there are 'N' preparing tests {(x1,y1), ((x2,y2)… (xN,yN)}, were xi ϵ Rd and yi ϵ { +1, - 1}. A hyper-plane of the structure (w,b) is built, where 'w' speaks to the weight vector and 'b' indicates the inclination. The spot result of the weight vector and inclination is spoken to by w.s. Another item's' is classified by

$$f(s) = \mathrm{sign}(w.s + b) = \mathrm{sign}(\sum_{i}^{N} \alpha_i y_i (s_i.s) + b)$$

Appearance in the dab item structure for the preparation vectors si is training. A lagrangian multiplier αi is related with each preparation point which indicates the centrality of each information point. The estimation of αi > 0 shows the fall of focuses nearer to the hyper-plane when the maximal edge hyper-plane is resolved. Such focuses are named as the help vectors. The rest of the information focuses have αi=0. The theory portrayal indicates the focuses that untruth near the hyper-plane. The focuses are used for building a free edge regarding the consistency of the classifier.

Conclusion

This paper has discussed about intrusion detection system and various data mining methods that have been applied to intrusion detection system by different researcher. It was indicated that data mining techniques helps in intrusion discovery from numerous points of view. Consequently data mining techniques can add to make better and more powerful intrusion detection system.

Reference

1) Fadi solo,Mohammadnoor and Ali Bou Nassif, Data Mining Techniques in Intrusion Detection System: A Systematic Literature Review, IEEE September 2018.
2) Jagjeet , Vinay Kumar and Vinod Kumar, Intrusion Detection using Data Mining Techniques: A Study through Different Approach, International Journal Science & Technology Research Excellence, Vol.3 Issue 4, July-August 2013.
3) Lidio Mauro and Lima De Campos, Network Intrusion Detection System Using Data Mining, Communication in Computer and Information Science, Vol.311, PP104-113 September 2012.
4) Zibusiso Dewa and Leandros A.Maglares, Data Mining and Intrusion Detection Systems, International Journal of Advanced Computer Science and Application Vol.7, November 2016.
5) Parmod Kumar and Sunil Kumar Intrusion Detection System in Clustering: A Review, International Journal of Advance Research in Computer Science and Management Vol.viii, August 2017.
6) Amudha Arul and Karthick Subburathinam, Classification Techniques for Intrusion Detection An Overview, International Journal of Computer Applications 76(16):33-40, August 2013.
7) Kajal Rai and Mandalika Syamala Devi, Decision Tree Based Algorithm for Intrusion Detection, International Journal Advanced Networking and Applications, Vol:07 Issue:04 January 2016.
8) Vivek Kshirsagar and Madhuri S.Joshi, Rule Based Classifier Models for Intrusion Detection System, International Journal of Computer Science and Information Technologies, Vol.7(1), 2016.
9) Dr.Saurabh Mukherjee and Neelam Sharma, Intrusion Detection Using Naive Bayes Classifier with Feature Reduction, Sciverse Sciencedirtect Proccedia Technology4 (2012) 119-128.
10) Mrutyunjaya Panda and Manas Ranjan Patra, Network Intrusion Detection Using Naive Bayes, International Journal of

Computer Science and Network Security, Vol.7, December 2007.

11) Snehal A.Mulay,P.R.Devale, Intrusion Detection System Using Support Vector Machine, International Journal of Computer Applications, Vol.3,June 2010.

12) Christos Douligeris, Aikaterini Mitrokotsa, ”DDoS attacks and defense mechanisms: classification and state-of-the-art” ,Computer Networks: The International Journal of Computer and Telecommunications Networking, Vol. 44, Issue 5 , pp: 643 - 666, 2004.

13) Z. Chen, L. Gao, K. Kwiat, Modeling the spread of active worms, Twenty-Second Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM), Vol. 3, pp. 1890 1900, 2003.

14) Mukherjee et al., "Network Intrusion Detection", Network, IEEE, Volume 8, Issue 3, May/June 1994, pp. 26-41

15) Thuraisingham, Bhavani, “Data Mining for Malicious Code Detection and Security Applications", IEEEJACM International Joint Conference on Web Intelligence and Intelligent Agent Technologies, 15-18 sept 2009

16) Lu et al., "Exploiting Efficient Data Mining Techniques to Enhance Intrusion Detection Systems", International Conference on Information Reuse and integration, Aug 2005 pp.512-517

17) Lee W and D. Xiang, "Information-theoretic measures for Anomaly detection", In Proc. of the 2001 IEEE Symp. on security and privacy, Oakland, CA, pp130143, IEEE society press, May 2001.

18) Ming-Yang Su, Kai-Chi Chang, Hua-Fu Wei, Chun-Yuen Lin, "A Real-time Network Intrusion Detection System Based on Incremental Mining Approach", ISI, June 17-20, 2008, Taipei, Taiwan.

19) Cohen, W. W., ”Fast effective rule induction”, In A. Prieditis and S. Russell (Eds.), Proc. of the 12th International Conference on Machine Learning, Tahoe City, CA, pp. 115123. Morgan Kaufmann, 9-12 July, 1995.

20) Neri, F., ”Comparing local search with respect to genetic evolution to detect intrusion in computer networks”, In Proc. of the 2000 Congress on Evolutionary Computation CEC00, La Jolla, CA, pp. 238243. IEEE Press, 16-19 July, 2000.

21) MohammedA.Ambusaidi, MemberIEEE, XiangianHe, SeniorMember,“Building An Intrusion Detection System Using a Filter-Based Feature Selection Algorithm, IEEE Transaction On Computers, Vol.65, October 2016.

22) C.F.Tasi,Y.F.Hsu,C.Y.LinandW .Y.Lin,”Intrusion detection by machine learning:Areview”,ExpertSystwithAppl.,vol36,no.10.pp.11994-12000,2009