International Research Journal of Mathematics, Engineering and IT
Vol. 3, Issue 9, September 2016 IF- 3.563 ISSN: (2349-0322)
© Associated Asia Research Foundation (AARF)
Website: www.aarf.asia Email : editor@aarf.asia , editoraarf@gmail.com

DATA MINING TECHNIQUES FOR INTRUSION DETECTION: A REVIEW

[1]Shikha Attri and [2]R C Gangwar and [3] Rajeev Bedi

[1]Post-Graduate Student, Computer Sc. & Engg, IKG Punjab Technical University, Kapurthala(Pb) India.

[2]Associate Professor, Department of Computer Sc, Beant College of Engg. & Tech, Gurdaspur(Pb) India.

[3]Assistant Professor, Department of Computer Sc, Beant College of Engg. & Tech, Gurdaspur(Pb) India.

ABSTRACT

With significant advancement of web, security of system activity is turning into a major issue PC system framework. Cyber attacks on system are expanding day-by-day. Intrusion is considered as most pitched attack on system traffic. Intrusion recognition framework has been utilized for finding out intrusion and to protect the security objectives of data from attacks. Data mining systems are utilized to screen and investigate extensive measure of system information and group this system information into anomalous and typical information. Since information originates from different sources, system traffic is substantial. Data mining methods such as classification and clustering are connected to design of intrusion detection framework. A viable Intrusion detection framework requires high recognition rate, low false caution rate and additionally high precision. This paper exhibits the audit on IDS and diverse Data mining methods connected on IDS for the powerful detection of pattern for both malicious and typical activities in the system, which creates secure data framework. This paper also presents two distinct clustering algorithms known as K-Means Clustering and Hierarchical Clustering Algorithm. K-Means clustering results indegeneracy and is not suitable for large databases.

INTRODUCTION

Data mining technology has been emerged as a means for identifying patterns and trends from large quantities of data. It is a withdrawal of hidden predictive information or knowledge from large databases. In Intrusion Detection System, information comes from various sources like online data, network log data, alarm messages etc. Since the variety of different data sources is too complex, the complexity of the operating system also increases. Also, network traffic is huge, so the data analysis is very hard. The data mining technology have the capability of extracting large databases; it is of great importance to use data mining techniques in intrusion detection. By applying data mining technology, intrusion detection system can widely verify the data to obtain a model, thus helps to obtain a comparison between the abnormal pattern and the normal behaviour pattern. An important problem in intrusion detection is how effectively it can separate the attack patterns and normal data patterns from a large number of network data and how effectively it generates automatic intrusion rules after collected raw network data. To accomplish this various data mining techniques are used [16].

Instruction Detection System

Data security is essential for all organizations furthermore for home PC people. This security is required as every single noteworthy information is transferred and oversaw on the web. Subsequently it is essential to spare data from interlopers. Client requires checking best procedure to spare the framework from unmistakable sorts of attacks. Intrusionis a sort of attack. In various territories it is said by particular name. For instance in deficiency like any unapproved client login into other client profile, information driven assaults on applications, illicit access to mystery data, assaults against helpless administrations, host-based assaults like benefit heightening and numerous different infections like, Trojan stallions or worms and so on. The yield of any framework is transfer upon its execution; integrity, accessibility and privacy, so these sorts of Intrusions specifically attack on these variables to minimize framework execution.

Instruction Detection System it is unrealistic to outline a totally secure framework. So a framework known as Intrusion Detection Systems (IDSs) is intended to improve security of PC framework [1]. To keep any further harm, an IDS is used to check, evaluate and report unlawful or unapproved system work with goal that important moves might be made to spare information [2]. Transfer on wellspring of information, IDS is isolated into 2 sorts;

Network-based - Network intrusion detection systems (NIDSs) finds network packets collected from a network segment.

Host-based - Host- based intrusion detection systems (HIDSs) for instance IDES (Intrusion Detection Expert System) [3] finds audit trails or framework calls generated by individual hosts.

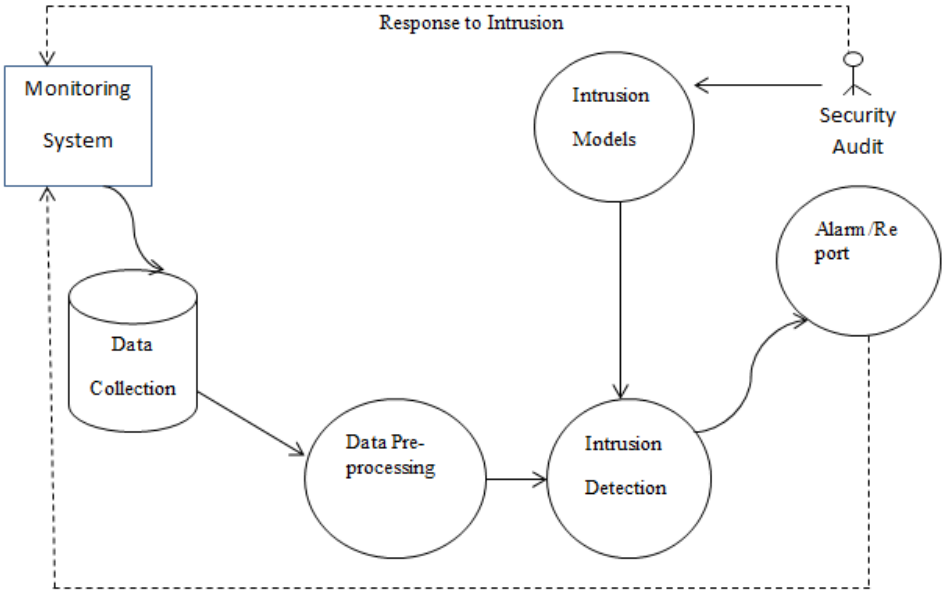

Figure1: Overall structure of Intrusion Detection System

1.1 Types of IDS

IDSs can likewise be ordered by recognition approaches they utilize. Essentially, there are two recognition techniques: misuse detection and anomaly detection. The significant reverence between the two techniques is that misuse detectiondistinguishes intrusions in view of components of known attacks while abnormality recognition dissects the properties of ordinary conduct. IDSs that utilize both recognition strategies are called hybrid detection-based IDSs. Case of hybrid detection-based IDSs are Hybrid NIDS utilizing Random Forests [4] and NIDES [5]. The accompanying subsections clarify the two detection approaches.

1.1.1 Misuse Detection

Misuse detection gets intrusion as far as the attributes of known attacks. Any activity that complies with the example of a referred to attack or helplessness is considered as intrusive. The fundamental issues in abuse recognition framework are the manner by which to compose a mark that incorporates every single conceivable variety of the related attacks. Furthermore, how to compose marks that doesn't likewise coordinate non-nosy movement. Block diagram of abuse based recognition framework is as taking after. Abuse recognition distinguishes interruptions by coordinating checked occasions to examples or signature of assaults. The attack signatures are the qualities connected with effective known attacks. The real favorable position of abuse identification is that the strategy has high exactness in identifying known attacks. Be that as it may, its recognition capacity is restricted by the signature database. Unless new assaults are changed into signatureand added to the database, abuse based IDS can't recognize any assault of this write. Deferent procedures, for example, master frameworks, signature investigation, and state move examination are used in misuse detection.

1.1.2 Anomaly Detection System

It depends on the typical conduct of a subject (e.g. a client or a framework). Any activity that altogether goes amiss from the ordinary conduct is considered as intrusive. That implies on the off chance that we could set up an ordinary movement profile for a framework, then we can signal all framework states shifting from built up profile. There is a vital distinction between anomaly based and misuse based strategy that the anomaly based attempt to distinguish the compliment of terrible conduct and misuse based recognition framework attempt to perceive the known awful conduct. For this situation we have two potential outcomes: (1)False positive: Anomalous exercises that are not intrusive but rather are flagged as intrusive. (2) False Negative: Anomalous exercises that are intrusive yet are flagged as non-intrusive. The square graph of inconsistency location framework is as taking after:

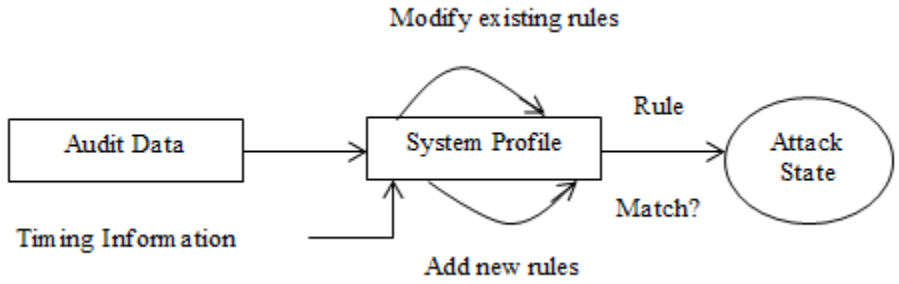

Fig 1.1 Misuse Detection Systems

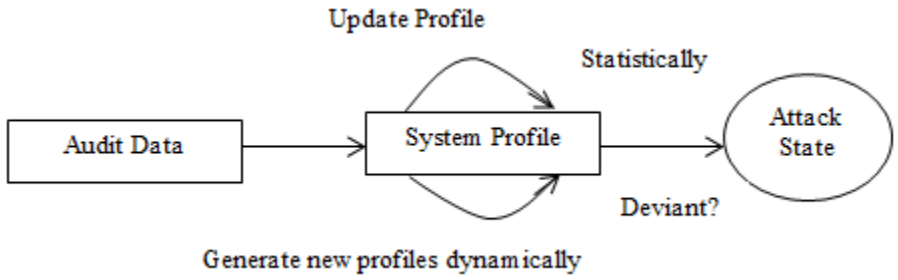

Fig 1.2 Anomaly Detection Systems

TABLE 1
Comparison between Misuse Detection and Anomaly Detection

Misuse Detection Systems	Anomaly Detection Systems
Advantages	**Advantages**
High detection rate, Accuracy for known behaviors	Can examine unknown and more complicated intrusions
Simplest and effective method	Rate of missing report is low
Low false alarm rate	Detect new and unforeseen voluntaries
Disadvantages	**Disadvantages**
It can detect only known attacks	Needs to be trained, and trained model carefully otherwise it tends to false positive
Need a regular update of the rule which are used	Low detection rate and high false alarm rate
Often never differentiate between an attack attempt and a successful attack	It can't identify new attack
Rate of missing report is high	intrusion detection depend upon new model

Data Mining Techniques for Network Intrusion Detection

Numerous analysts have researched the organization of data mining algorithms and systems for intrusion detection [13,15-23, 32,33]. Cases of these methods incorporate in [16-18]:

Feature selection data analysis:The fundamental thought in features determination is to evacuate highlights with practically no prescient data from the first arrangement of components of the review information to frame a subset of suitable elements [24]. Feature selection altogether decreases computational multifaceted nature coming about because of utilizing the full unique list of capabilities.

Classification analysis*:* The objective of arrangement is to appoint objects (intrusions) to classes in view of the estimations of the item's elements. Classification algorithms can be utilized for both misuse and anomaly detections [16]. In misuse detection, system movement information are gathered and named as "ordinary" or "intrusion". In anomaly detection, the typical conduct model is found out from the preparation dataset that are known not "ordinary" utilizing learning algorithms.

Clustering analysis:Clustering appoint objects (interruptions) to gatherings (groups) on the premise of separation estimations made on the items. Rather than order, clustering is an unsupervised learning process subsequent to no data is accessible on the names of the preparation information. In inconsistency recognition, bunching and exception examination can be utilized to drive the ID model [16].

Association and correlation analysis:The principle target of association rule investigation is to find affiliation connections between particular estimations of elements in substantial datasets. This finds concealed examples and has a wide assortment of uses in business and exploration. Association rules can choose separating qualities that are valuable for intrusion detection. It can be connected to discover connections between framework traits depicting system information. New qualities got from amassed information may likewise be useful, for example, summary counts of trafficmatching a particular pattern.

Stream data analysis*:* Intrusions and malicious attacks are of element nature. In addition, information streams may identify intrusions as in an occasion might be typical all alone, yet thought to be malignant if saw as a feature of a succession of occasions [16]. In this way, it is important to perform intrusion detection in information stream, constant environment. This recognizes arrangements of occasions that are habitually experienced together, find consecutive examples, and distinguish exceptions.

Distributed data mining:Intruders can work from a few unique areas and assault a wide range of destinations. Distributed data mining techniques might be used to break down system information from a few system areas, this distinguishes disseminated assaults and keep aggressors in better places from hurting our information and assets.

Visualization and querying tools*:* Visualization data mining devices that incorporate components to view classes, affiliations, bunches, and exceptions can be utilized for review any strange examples identified. Graphical UI connected with these apparatuses permits security investigators to comprehend intrusion detection results, assess IDS execution and settle on future upgrades for the framework.

LITERATURE SURVEY

Memon V I et al. [6] introduced work is a gathering of three data mining techniques to diminish false alarm rate in IDS that is known as a hybrid IDS which has k-Means, K-closest neighbor and Decision Table Majority strategy for anomaly detection. Displayed hybrid IDS assessed over the KDD-99 Data set; such kind of information set is utilized worldwide for computing the execution of different IDS. At first bunching executed by means of k-Means over KDD99 information sets then executed two-arrangement technique; KNN took after by DTM. The introduced framework can distinguish the intrusions and classify them into four sorts: Remote to Local (R2L), Denial of Service (DoS), User to Root (U2R) and Probe.

Wankhade K et al. [7] displays a hybrid data mining approach incorporating feature selection, clustering, bunching, partition and consolidation and grouping outfit. A methodology for assessing the quantity of the group centroid and selecting the appropriate early bunch centroid is exhibited.

Dhakar M et al. [8], in context to improve execution, the work displays a model for IDS. This enhanced model, known as REP (Reduced Error Pruning) based IDS Model provides yield with more noteworthy exactness alongside the expanded number of legitimately grouped occasions. It utilizes the two classification of grouping methodologies to be specific, K2 (BayesNet) and REP (Decision Tree). Here REP gives a powerful grouping alongside the pruning of tree with speedy choice learning ability.

Zubair Md. Fadlullah et al. [9], in this article, they highlighted the significance on outlining suitable intrusion detection frameworks to battle assaults against cognitive radio systems. Additionally, we proposed a basic yet viable ID, which can be effectively actualized in the auxiliary clients' cognitive radio programming. Authorsdesigned IDS utilizes non-parametric cusumalgorithm, which offers anomaly detection. By taking in the typical method of operations and sys-tem parameters of a CRN, the proposed IDS can distinguish suspicious (i.e., strange or anomalous) conduct emerging from an attacks. Specifically, we displayed a case of a jamming attack against a CRN auxiliary client, and exhibited how our proposed IDS can identify the assault with low recognition latency. In future, their work will perform further examinations on the most proficient method to upgrade the detection sensitivity of IDS.

MueenUddin,et al [10], This paper has concentrated on the proficiency and execution of new IDS: known as signature-based multi-layer IDS utilizing mobile agents. It then talks about the advancement of another signature based ID utilizing mobile agents. The proposed framework utilizes mobile agents to exchange rule-based signatures from substantial reciprocal database to little signature database and after that consistently overhaul those databases with new signatures recognized.

R. China Appala Naidu et al. [12] utilized three Data mining systems SVM, Ripper rule and C5.0 tree for Intrusion detection furthermore looked at the proficiency. By test result, C5.0 decision tree is proficient than other. All the three Data mining system gives higher than 96% detection rate.

RoshanChitrakar et al. [13] Proposed a hybrid approach to intrusion detection by utilizing k-ModelIDS grouping with Naïve Bayes classification and watched that it gives preferred execution over K-Means clustering procedure took after by Naïve Bayes classification additionally time unpredictability increments when expand the quantity of information focuses.

RoshanChitrakar et al. [14] proposed a hybrid approach of consolidating k-methodclustering with Support Vector Machine Machineprocedure and delivered better execution contrasted with

k-model IDS with Naïve Bayes classification. The methodology demonstrates change in both Accuracy and Detection Rate while diminishing False Alarm Rate when contrasted with the k-model ids grouping approach took after by Naïve bayesclassification procedure.

SumaiyaThaseen et al. [15] Analyzed distinctive tree based characterization systems for IDS. Exploratory results demonstrate that Random tree model lessens false alert rate and has most elevated level of accuracy.

ALGORITHM

K-Means: K-Means algorithm is veryfamoustechniqueof clustering checking that motive to divide 'n' information objects into 'k' clusters in which each information object refers to cluster with nearest mean. It utilizes Euclidean metric as anequal measure. Distance equation to find distance among2 objects is:

$$L(x,y) = L(y,x) = |x-y| = \sqrt{\sum_{j=1}^{m}(xj-yj)^2}$$

Procedure K-Means

Step 1: Choose k objects from L as initial cluster centers
Step 2: Assign every object to cluster according to mean value of objects in cluster.
Step 3: Update cluster means, i.e., calculate mean value of objects for everycluster.
Step 4: Until no change

Important properties of K-Means algorithm:

1. Efficient in processing large data sets.

2. Works only on numerical values.

3. Clusters have convex shapes.

Hierarchical Clustering

In suggested agglomerative clustering scheme, start by Mclusters at level q=1 as given by optimized GGM model of $l(s)$which in case of supervised learning is$l\ (s) = \sum_{d=1}^{A}\sum_{M=1}^{M_d} l\big(s|m,d\ L\ (m)L(d)\big)$where M_d is optimal number of components forClass$d$. At everylargerstage in hierarchy two clusters iscombineddepends on

aequalitymeasure among pairs of clusters. Processis repeated until we reach one cluster attop level. That is, at level $q = 1$ there are M clusters and 1 cluster at last level,$l = 2M - 1$. let$l_q\ (s|m)$be density for $k'th$ cluster at level$q$ j and $l_q(m)$as its mixingproportion, i.e., density model at level j is $l\ (s) = \sum_{m=1}^{M-q+1} L_q\ (m) L_q(s|q)$If clusters m and n at level q are merged into$g$at level $q + 1$ then.

$$l_{q+1}(s|g) = \frac{l_q(s|m)\ .\ l_q(m) + l_q(s|n)\ .L_q(n)}{L_q(m) + L_q(n)}\ , L_{q+1}(g) = L_q(m) + L_q(n)$$

Natural distance measure among cluster densities is Kullback-Leibler (KL) divergence[11], since it reflects dissimilarity among densities in probabilistic space.Problem is that KL only obtains an analytical expression for first level in hierarchy while distances for subsequently levels have to be approximated.

Conclusion

Since investigation of intrusion detection started to pick up energy in the security group approximately ten years ago, various differing thoughts have developed for standing up to this issue. Intrusion detection frameworks differ in the sources they use to acquire information and in the particular methods they utilize to examine this information. Most frameworks today group information either by misuse detection or anomaly detection: every methodology has its relative merits and is joined by an arrangement of restrictions. It is likely not reasonable to expect that an intrusion detection framework be prepared to do effectively grouping each occasion that happens on a given framework. Immaculate discovery, similar to impeccable security, is basically not a feasible objective given the many-sided quality and quick advancement of current frameworks. An IDS can, in any case try to lift the bar for intrusion or attacks by plunging the viability of enormous classes of interruption or assaults and rising the work issue required to get a framework compromise. A decent intrusion detection framework guarantees to permit more prominent trust in the aftereffects of and to enhance the scope of intrusion detection, making this a basic segment of any exhaustive security architecture.

References

[1] LI Yongzhong,YANGGe,XU Jing Zhao Bo “A new intrusion detection method based on Fuzzy HMM “IEEE Volume 2, Issue 8, November 2008.

[2] Tarem Ahmed, Boris Oreshkin and Mark Coates, “Machine Learning Approaches to Network Anomaly Detection” in Workshop on Tackling Computer Systems Problems with Machine Learning Techniques, 2007.

[3] Li Tian, “Research on Network Intrusion Detection System Based on Improved K-means Clustering Algorithm”, Computer Science-Technology and Applications, IFCSTA 2009.

[4] ZhenglieLi ”Anomaly Intrusion Detection Method Based on K-Means Clustering Algorithm with Particle Swarm Optimization“ Springer Volume 4, Issue 2, April 2011.

[5] SK Sharma, P Pandey, SK Tiwari “An improved network intrusion detection technique based on k-means clustering via Naïve bayes classification” IEEE Volume 2, Issue 2, February 2012.

[6] V. I. Memon and G. S. Chandel, “A Design and Implementation of New Hybrid System for Anomaly Intrusion Detection System to Improve Efficiency”, International Journal of Engineering Research and Applications (IJERA), Volume 4, Issue 5, pp. 01-07, May 2014.

[7] K. Wankhade, S. Patka and R. Thool, “An efficient approach for Intrusion Detection using data mining methods”, International Conference on Advances in Computing, Communications and Informatics (ICACCI), INSPEC, pp. 1615-1618, August 2013.

[8] M. Dhakar and A. Tiwari, “A New Model for Intrusion Detection based on Reduced Error Pruning Technique” International Journal of Computer Network and Information Security, pp. 51-57, 2013.

[9] Zubair Md. Fadlullah, Hiroki Nishiyama, “An Intrusion Detection System (IDS) for Combating Attacks Against Cognitive Radio Networks” IEEE 2013.

[10] MueenUddin ,Azizah Abdul Rehmanletl “Signature-based Multi-Layer Distributed Intrusion Detection System using Mobile Agents” International Journal of Network Security, Volume 15, Number 1, pp.79-87, Jan. 2013.

[11] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, 1995.

[12] R.ChinaAppala Naidu and P.S.Avadhani, “A Comparison of Data Mining Techniques for Intrusion Detection”, International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), pp-41-44, IEEE, 2012.

[13] RoshanChitrakar and Huang Chuanhe, “Anomaly based Intrusion Detection using Hybrid Learning Approach of combining k- Medoids Clustering and Naïve Bayes Classification”, IEEE,2012.

[14] RoshanChitrakar and Huang Chuanhe, “Anomaly Detection using Support Vector Machine Classification with k-Medoids Clustering”, IEEE, 2012.

[15] SumaiyaThaseen and Ch. Aswani Kumar, “An Analysis of Supervised Tree Based Classifiers for Intrusion Detection System”, International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME), IEEE, February 2013.

[16] Vaishali B Kosamkar et al. “Data Mining Algorithms for Intrusion Detection System: An Overview”, ICRTITCS, 2012.