



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Review on Data Mining Techniques to reduce false alarms in Intrusion Detection Systems

¹ Shruti Dubb

M Tech. Scholar (CSE Department)
Sri Sai University
Palampur, Himachal Pradesh (INDIA)

Abstract: In today's network world, there is a rapid increase in intrusions which are handled by intrusion detection systems. These intrusion detection systems produce alerts when any malicious activity occurs or any attack is detected, but these systems have a major issue, which is, generating a high number of false alarms. This issue is being solved with the help of different data mining techniques. This paper consists of different data mining techniques used to reduce false alarms. It also differs the classification and clustering algorithms to be used.

Keywords: Intrusion, Intrusion detection systems, False alarms, Data mining, Classification, Clustering.

I. Introduction

Intrusion detection systems are developed to oversee the systems and provoke alarms when they find any suspect in the network. There are mainly two techniques to detect intrusions. First is, Misuse based technique and second is, Anomaly based technique. Misuse based technique is used to detect patterns of known attacks only. Whereas anomaly based intrusion detection technique is used to detect unknown attacks or new attacks. Anomaly based intrusion detection are very useful but they have a major problem, that is these type of IDS creates high false alarms. A false alarm means high rate of false positives and high rates of false negatives. This issue is an important issue in anomaly based intrusion detection systems which is to be solved. There are many researchers doing research for decreasing false alarms through data mining approaches.

Various data mining approaches are being used to solve the issue of false alarms in intrusion detection systems. Following is the collection of most recently used and most effective approaches to make intrusion detection systems, false alarm free.

II. Methodologies to improve Anomaly based IDS

There are three kind of approaches being used to deduct false alarms in intrusion detection systems

- Data mining of features
- Classification of alerts
- Event correlation

Data mining of features works on the features of dataset of networks. According to the different criterion, features are classified into normal and abnormal classes. [1]

Classification of alerts works on the alert checking and its categorization of alerts into false alerts and true alerts.

Event correlation works by getting together similar alerts and allowing analyst to look only few alerts instead of hundreds of alerts. [2]

We will discuss about feature data mining approaches in intrusion detection systems to cut false alarms

III. Categories of Intrusion Detection Systems

A. Signature based detection

Signature detection involves looking out network traffic for a series of malicious bytes or packet sequences. The most advantage of this method is that signatures are terribly simple to develop and perceive if we all know what network behaviour we tend to try to spot. As an example, we tend to would possibly use a signature that appearance for explicit strings inside exploit explicit buffer-overflow vulnerability. The events generated by signature- primarily based IDS will communicate the explanation for the alert. As pattern matching may be done additionally expeditiously on trendy systems that the quantity of power required to perform this matching is least for a rule set. Limitations of these signature engines are that they solely observe attacks whose signatures are antecedent keep in database; a signature should be created for each attack; and novel attacks cannot be detected. Signature primarily based detection doesn't work well once the user uses advanced technologies such as generators, payload encoders and encrypted knowledge channels. The potency of the signature primarily based systems is greatly diminished, because it has got to produce a replacement signature for each variation. As the signatures keep on increasing, the system engine performance decreases. Due to this, several intrusion detection engines are unit deployed on systems with multi processors and multi Gigabit network cards. IDS creators create not known signatures before the assailant will, thus on forestall the novel attacks on the

system. The distinction of speed of creation of the new signatures between the developers and attackers confirm the potency of the system. [1]

B. Anomaly primarily based detection

The anomaly primarily based detection relies on shaping the network behaviour. The network behaviour is in accordance with the predefined behaviour, then it is accepted as an alternative it triggers the event within the anomaly detection. The accepted network behaviour is ready or learned by the specifications of the network directors. The necessary introduce shaping the network behaviour is that the IDS engine capability to cut through the numerous protocols at all levels. The Engine should be ready to method the protocols and perceive its goal. Although this protocol analysis is computationally costly, the advantages it generates like growing rules will lead to less false positive alarms. [1]. the foremost downside of anomaly detection is shaping its rule set. The potency of the system depends on however well it's enforced and tested on all protocols. Rule shaping method is additionally tormented by numerous protocols used by numerous vendors. Apart from these, custom protocols conjointly create rule shaping a tough job. For detection to occur properly, the elaborated data concerning the accepted behaviour of network is created by developer. However once the principles area unit outlined and protocol is designed then anomaly detection systems performs better.

If bad behaviour of the user falls beneath the accepted behaviour, then it goes overlooked. associate degree activity such as directory traversal on a targeted vulnerable server, that complies with network protocol, easily goes overlooked because it doesn't trigger any out-of-protocol, payload or information measure limitation flags. the foremost advantage of anomaly primarily based detection over signature-based engines is that a unique attack that a signature doesn't exist may be detected if it falls out of the traditional traffic patterns. This is often determined once the systems notice new automatic worms. If the new system is infected with a worm, it always starts scanning for other vulnerable systems at associate degree accelerated rate filling the network with malicious traffic, so inflicting the event of a transmission control protocol affiliation or information measure abnormality rule.

IV. Data Mining approaches in IDS

A. Statistical techniques

These approaches use algorithms which will learn from knowledge looking on relationship between knowledge. This method has been there for hundreds of years. it's primarily a sub field of arithmetic. Applied mathematics ways cannot handle giant knowledge set thus these are applied for smaller data with less attributes. Some rule includes:

♦ Markov process model

Andrei Markov model could be a model employed in haphazardly dynamic systems wherever it's assumed that future states rely solely on this state and not on the sequence of events that preceded it. [3]. The Intrusion detection in this model is done by work the system at fastened intervals and keeping track of its state; a chance for every state at a given quantity I_s . The modification of the state of the system occurs once Associate in event happens and the behavior is detected as anomaly if the chance of incidence of that state is low. The transitions between sure commands verify the anomaly detection wherever command sequences were necessary.

♦ Time series model

Time series analysis includes ways for analyzing statistic knowledge so as to extract purposeful statistics and alternative characteristics of the info. Statistic foretelling is that the use of a model to predict future values supported antecedently ascertained values. [4]. whereas multivariate analysis is usually utilized in such some way on check theories that this prices of 1 or a lot of freelance statistic have an effect on this value of yet again series, that focuses on scrutiny values of one statistic or multiple dependent statistic at totally different points in time. Interval timers along with Associate in event counter or resource live are major elements during this model. Order and inter- arrival times of the observations likewise as their values are keep. If the chance of incidence of a replacement observation is simply too low then it's thought of as anomaly. The disadvantage of this model is that it's a lot of computationally costly.

♦ Mean and variance model

The standard deviation use to quantify the number of variation or dispersion of a group of knowledge values. a customary deviation near zero indicates that the info points tend to be terribly near the mean of the set, whereas a high variance indicates that the info points are detached over a wider vary of values. In applied mathematics mean, customary deviation, or any alternative correlations are referred to as a flash. If the event that falls outside the set interval on top of or below the instant is aforesaid to be abnormal. The system is subjected to vary by considering the aging knowledge and creating changes to the applied mathematics rule knowledge base. There are 2 major benefits over Associate in operational model. First, previous information isn't needed determinative the conventional activity in order to set limits; Second, determinative the confidence intervals depends on ascertained user knowledge, as it varies from one to other user. Threshold model has not this flexibility. the foremost variation on the mean and variance model is to grant higher weights for the recent activities.

B. Machine learning approaches

This technique uses rule which will learn from knowledge while not counting on any kind of rules. Machine learning is recent development and it's one among subfield of computing. Main advantage is that it will handle big selection of attributes [5]. Some rule include -

♦ **Naïve Bayesian rule**

In naive Bayes classifiers area unit a family of easy probabilistic classifiers supported applying theorem with robust (naive) independence assumptions between the options. Naive Bayes classifiers area unit extremely scalable of parameters linear within the number of variables (features/predictors) during a learning downside. Maximum-likelihood coaching will be done by evaluating a closed-form expression that takes linear time instead of by dearly-won unvarying approximation as used for several alternative forms of classifiers. [5]

♦ **Genetic rule**

Genetic rule could be a programming technique that uses a problem-solving strategy. Given a particular downside to unravel, the input to the GA could be a set of potential solutions thereto downside, encoded in some fashion, and a metric known as a fitness operate that enables every candidate to be quantitatively evaluated. These candidates is also solutions already famed to figure, with the aim of the GA being to enhance them, however additional usually they're generated arbitrarily. The GA then evaluates every candidate in step with the fitness operate. during a pool of generated candidates, of course, most won't work all, and these are going to be deleted. However, strictly out of the blue, a couple of could hold promise they could show activity, notwithstanding solely weak and imperfect activity, toward resolution the matter.

♦ **SVM**

Support Vector Machines (SVMs) are a decent willdidate for intrusion detection systems that can give time period detection capability, agitate giant spatiality of knowledge. SVMs plot the coaching vectors in high dimensional feature area through nonlinear mapping and labeling every vector by its category. the information is then classified by deciding a group of support vectors, that area unit members of the set of coaching inputs that define a hyper plane within the feature area. SVM could be a technique for resolution a range of learning, Classification and prediction issues. [7] the fundamental SVM deals with two-class issues within which the information area unit separated by a hyper plane outlined by variety of support vectors. Support vectors area unit a set of coaching knowledge accustomed outline the boundary between the 2 categories. In things wherever SVM cannot separate 2 categories, it solves this downside by mapping input file into high-dimensional feature areas employing a kernel operate. In high-dimensional area it's potential to make a Hyper plane that enables linear separation.[6] Compared with the ANN, the SVM have 2 benefits. Firstly, the worldwide optimum will be derived. Secondly, the over fitting downside will be simply controlled by the selection of an acceptable margin that separates the information. Classes of samples and measures happiness completely different categories decide the consequences on the target operate. That assures eliminating or reducing the impact of the noise and outlier samples for the target operate.

♦ **Neural Networks**

Artificial neural network consists of a group of process components that area unit extremely interconnected and rework a group of inputs to a group of desired outputs. The results of the transformation is set by the characteristics of the weather and therefore the weights related to the interconnections among them. With enhancing the connection among different nodes the network is ready to adapt to the required outputs. not like professional systems, which might give the user with a definitive answer if the characteristics that area unit reviewed specifically match those that are coded within the rule base, a neural network conducts Associate in Nursing analysis of the data and provides a likelihood estimate that the information matches the characteristics that it's been trained to acknowledge [6]. A classification rate may also be computed if the system is intended to perform attacks multi classification.

♦ **K-NN rule**

k-NN could be a kind of instance-based learning or lazy learning wherever the operate is merely approximated domestically and every one computation is delayed till classification. The k-NN rule is among the best of all machine learning algorithms. each for classification and regression, it will be helpful to assign weight to all the neighbor's contribution, in order that the nearer neighbors contribute additional to the typical than the additional distant ones. as an example, a standard weight theme consists in giving every neighbor a weight of $1/d$, wherever d is that the distance to the neighbor. The neighbors area unit taken from a group of objects that the category (for k-NN classification) or the article property worth (for k-NN regression) is understood. this may be thought because the training set needed for the rule, there is no specific coaching step is needed.. A disadvantage of the k-NN rule is that it's sensitive to the native structure of the information. The rule has nothing to try and do with and isn't to be confused with k-means another standard machine learning technique.

♦ **Decision trees**

A decision tree will be accustomed visually and expressly represent selections and higher cognitive process. In data processing a choice tree describes knowledge however not decisions; rather the ensuing classification tree will be Associate in input for higher cognitive process. This page deals with call trees in data processing. There

area unit several specific decision-tree algorithms. Notable ones include: ID3 (Iterative Dichotomiser 3) [8], C4.5 (successor of ID3) [8], CART (Classification and Regression Tree). This type of approach leads to unbiased predictor choice and doesn't need pruning.

V. Current algorithms in use

Neethu B [5] uses machine learning approaches, which includes N-B classification and principle component analysis. KDD 99 dataset is used, 42 features have been used in KDD99 dataset and each connection can be categorized into five main classes (normal class and anomaly classes (DOS, Probe, U2r,r2l)). The naïvebayes Updatable classifier in the WEKA is used for doing the experiment. The experiment is conducted with 40 instances. The result of the classification is accuracy of the system is calculated to be 89%.

Mr. Sunil Choudhary, Pankaj Dalal [6] used support vector machine for classification procedure of kdd 99 dataset and achieved false alarm rate of 0.5 and rate of classification up to 98%

Krishan kumar, Gulshan kumar [8] used full dimension features of the NSL-KDD dataset, proposed feature selection method for network intrusion detection system with reduced 12 dimension features with the use of the J48 classifier. The False positive rate for each class decreases from 0.121 to 0.104. Classification accuracy achieved by reduced feature set is improved than using 41-variable dataset with J48 standalone classifier.

VI. Conclusion

There are lots of data mining approaches being used to improve false alarm rate in intrusion detection systems. Various algorithms and techniques have been discussed. Among them SVM can be used as a better option. Decision trees are used to classify large dataset properly. Different range of datasets can be used. Multiclass SVA classifier gives better results than normal binary class. Features are extracted in different ways and then reduced to classify them properly. Different approaches can be used with varying data. Also hybrid techniques can give better results.

References

- [1] Prof. Shivendu Dubey, Neha Tripathi., A Survey on Intrusion Detection Systems , 2015 IJSRSET, Volume 1 , Issue 6, Themed Section: Engineering and Technology.
- [2] El Mostapha Chakir , Chancerel Codjovi, Youness Idrissi Khamlichi, Mohammed Moughit , False Positives Reduction in Intrusion Detection Systems Using Alert Correlation and Data mining Techniques, Volume 5, Issue 4, 2015 International Journal of Advanced Research in Computer Science and Software Engineering.
- [3] Asieh Mokarian, Ahmad Faraahi, Arash Ghorbannia Delavar, False Positives Reduction Techniques in Intrusion Detection Systems-A Review , IJCSNS International Journal of Computer Science and Network Security, VOL.13 No.10, October 2013
- [4] V. Jyothisna V. V. Rama Prasad, A Review of Anomaly based IntrusionDetection Systems, International Journal of Computer Applications (0975 8887) ,Volume 28 No.7, September 2011
- [5] Neethu B, Classification of Intrusion Detection Dataset using machine learning Approaches , International Journal of Electronics and Computer Science Engineering.
- [6] Mr. Sunil Choudhary, Pankaj Dalal, An Architecture for Network Intrusion Detection System Based on Dag Classification using Hybrid Enesemble and Ensemble Method, 2015
- [7] Majid Ghonji Feshki , Managing Intrusion Detection Alerts Using Support Vector Machines, IJCSS), Volume (9), Issue (5), 2015.
- [8] Krishan Kumar , Gulshan Kumar , Yogesh Kumar, Feature Selection Approach for Intrusion Detection System, International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), Vol.2 , No.5, Pages :47-53 (2013), Special Issue of ICCECT 2013