International Journal of Security and Its Applications
Vol. 9, No. 5 (2015), pp. 69-76
http://dx.doi.org/10.14257/ijsia.2015.9.5.07

Intrusion Detection System: A Review

Sanjay Sharma and R. K. Gupta

Department of CSE & IT
Madhav Institute of Technology and Science, Gwalior (M.P.), India

Abstract

With the incredible expansion of network-based services and responsive information on networks, network protection and security is getting more and more significance than ever. Intrusion poses a serious security risk in network surroundings. The ever rising new intrusion or attacks type poses severe difficulties for their detection. The human labeling of the accessible network audit information instances is generally tedious, expensive as well as time consuming. This paper focuses on study of existing intrusion detection task by using data mining techniques and discussing on various issues in existing intrusion detection system (IDS) based on data mining techniques.

Keywords: *Data Mining, Intrusion Detection System, Attack, Clustering*

1. Introduction

Intrusion detection System (IDS) is a type of security management system for computers and networks [1]. An intrusion detection system (IDS) inspects all outbound and inbound network action and find out the doubtful patterns that may point to a network or system intrusion or attack from someone trying to crack into or conciliation a system. IDS gathers and observed information from different areas inside a network of systems to find out probable safety breaches, which contain together called intrusions (attacks exterior from the association) and misuse (attacks from inside the association). IDS use susceptibility assessment, it is an expertise which is design and developed to appraise the security of a network [2]. Data mining techniques can be used to detect intrusions. Applications of data mining have presented a collection of research efforts on the use of data mining in computer security. In the context of security of the data we are looking for the information whether an information security breach has been experienced [3]. This data could be collected in the perspective of discovering attacks or intrusions that aim to break the privacy and security of services, information in a system or alternatively, in the context of discovering evidence left in a computer system as part of criminal activity. There are four major categories of networking attacks: Denial of Service, Probing, User to Root and Remote to Local.

Intrusion detection system is the area where data mining concentrate heavily. There are two fold reasons for this first an IDS is very common and very popular and extremely critical activity. Second, large volume of the data on the network is dealing so this is an ideal condition for the data mining to use it. The data mining technology has the enormous benefits in the data extracting attributes and the rule, so it is significant to use data mining methods in the intrusion detection [4]. A significant problem of IDS is how to efficiently divide the normal behavior and the abnormal behavior from a huge number of raw information's attributes, and how to effectively generate automatic intrusion rules following composed raw data of the network. To accomplish this, different data mining methods must be studied, like classification, correlation analysis of data mining methods and so on [4]. The ever rising new intrusion or attacks type poses severe difficulties for their detection. The human labeling of the accessible network audit information instances

ISSN: 1738-9984 IJSIA

is generally tedious, expensive as well as time consuming. This paper focuses on study of existing intrusion detection task by using data mining techniques and discussing on various issues in existing IDS based on data mining techniques.

1.1 *Types of Networking Attacks*

Following are the four major categories of networking attacks:

Denial of Service (DoS): In DoS attack, legitimate networking requests are not served because attacker makes the resources either too busy or full to serve the request. Hence the legitimate user cannot access the services of a machine or network resources. Example: apache, mail bomb, back etc.

Probing (Probe): In probing, attacker scans a machine or a network device for gathering the information about weaknesses or vulnerabilities that can be exploited later to compromise the target system. Example: saint, mscan, nmap *etc*.

User to Root (U2R): In U2R attacks, an authorized user attempt to abuse the vulnerabilities of the system in order to gain privilege of root user for which they are not authorized. Example: perl, xterm, Fd-format etc.

Remote to Local (R2L): In this type of attacks, a remote user tries to gain access as a local user to a local machine by sending packets to a machine over the internet. An external intruder exploits vulnerabilities of the system to access the privileges of a local user. Example: xlock, phf, guest *etc*.

2. Survey Work

Memon V I, Chandel G S [5] presented work is a grouping of three data mining methods to decrease false alarm rate in IDS that is called a hybrid IDS which has k-Means, K-nearest neighbor and Decision Table Majority method for anomaly detection. Presented hybrid IDS evaluated over the KDD-99 Data set; such type of data set is used worldwide for calculating the performance of various IDS. Initially clustering executed via k-Means over KDD99 data sets then executed two-classification method; KNN followed by DTM. The presented system can detect the intrusions and categorize them into four types: Remote to Local (R2L), Denial of Service (DoS), User to Root (U2R) and Probe.

Wankhade K, Patka S, Thool R [6] presents a hybrid data mining approach encompassing feature selection, filtering, clustering, divide and merge and clustering ensemble. An approach for evaluating the number of the cluster centroid and selecting the suitable early cluster centroid is presented.

Dhakar M, Tiwari A [7], in perspective to enhance performance, the work presents a model for IDS. This improved model, named as REP (Reduced Error Pruning) based IDS Model gives output with greater accuracy along with the augmented number of properly classified instances. It uses the two algorithms of classification approaches namely, K2 (BayesNet) and REP (Decision Tree). Here REP provides an effective classification along with the pruning of tree with quick decision learning capability.

Subramanian P.R and Robinson J.W [8] have discussed on network security through Intrusion Detection Systems (IDSs) with data mining approaches. This model uses binary classifier (C4.5) and multi boosting technique. Here binary classifier is used to classify bit by bit transmission of the packet and used for each type of attack to improve the accuracy and to reduce the variance and bias multi boosting technique is used.

Chandolikar N.S and Nandavadekar V.D [9] presented an approach for intrusion detection using J48 decision tree classifier and also compared with some other tree based algorithms in which J48 tree shows the best performance. To evaluate the performance of the algorithm correctly classified instances, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Root Relative Squared Error and Kappa statistics measures are used.

Barot V and Toshniwal D [10] presented a hybrid model that ensembles Naive Bayes (statistical) and Decision Table Majority (rule based) approaches. Naive Bayes predicts quickly because of less complex functioning of it and processes training data set only once to store statistics. Decision Table Majority (DTM) is a classifier that matches each of the attribute values all together. This model uses sequential reclassification approach for combining rule base classifier. Here correlation based feature selection (CFS) algorithm is used for attribute selection using BestFirst search. Author used KDDCUP'99 data set for their experiment.

Om H and Kundu A [11] presented a hybrid model that combines K-Means and two classifier methods: K-nearest neighbor and Naive Bayes. This model uses entropy based feature selection method for attribute selection. It applies K-Means clustering algorithm for clustering purpose (used number of clusters five) which is followed by K-nearest neighbor (KNN) and Naïve Bayes classification algorithms for detecting intrusions. The model shows better approach than only K-Means and K-Means, KNN. Author also used the KDD99 cup data set for performing their experiment.

Thakur M R & Sanyal S [12] suggested a multi-dimensional method towards intrusions or attacks detection. Network system usage various parameters like destination and source IP addresses; destination and source ports; outgoing and incoming network traffic information rate and amount of CPU cycles per request are split into numerous dimensions. Observing raw bytes of information corresponding to the values of the network factors, an established function is inferred throughout the training phase for every measurement. This grown-up function takes a measurement value as an input and returns a value that represents the level of anomaly in the system usage relating to that dimension. This mature function is referred as *entity Anomaly pointer*. *Entity Anomaly pointer* recorded for every of the measurement are then used to produced a Universal *Anomaly Pointer*, a function with n variables (n is the number of dimensions) that provides the U*niversal Anomaly Factor,* a pointer of anomaly over the system usage based on all the measurements measured together. The U*niversal Anomaly pointer* inferred throughout the training phase is then used to find out anomaly over the network traffic throughout the detection phase. Network traffic data encountered through the detection phase is fed back to the system to develop the maturity of the *Entity Anomaly Pointers* and hence the U*niversal Anomaly Pointer.*

Pathak V and Ananthanarayana V. S [13] have suggested a multi-threaded K-Means clustering approach. In this approach they have used six threads which run in parallel. Out of which five threads are used to cluster the data and the last sixth thread is used to take decision classify the data. Out of five threads, each is used to identify particular type of attack and normal or abnormal data. Author used KDD99 training data set for their experiment. Proposed approach i.e. multi-threaded K-Means gives better result in comparison to K-Means.

Wang P and Wang J Q [14] discussed about data mining which is popularly known as an important way to mine useful information from large volumes of data which is noisy, fuzzy, and random. In this, present the whole techniques of the IDS along with data mining method in details. Author mainly discussed about three data mining based approaches: Classification, Association and Sequence rules. Also discussed the system architecture of the IDS.

Muda Z, Yassin W, Sulaiman M.N and Udzir N.I [15] describe a hybrid learning approach that combines K-Means clustering method and Naive Bayes classification method. In the proposed approach, firstly K-Means (as a pre-classification) cluster all the data in to corresponding groups and then Naive Bayes classifier is used to classify the resultant clusters into attack classes as a final task. Because of this, the data that has been misclassified in the earlier stage (K-Means) may be classified correctly in the consequent

classification task (Naive Bayes). Here author took number of clusters (K) = 3 and KDD Cup 99 benchmark data set for evaluating the performance of their approach.

Dewan Md. Farid, Nouria Harbi [16] offered a learning algorithm for adaptive network intrusion detection using Naive Bayesian classifier and ID3 algorithm which performs good detections and keeps less false positives and also eliminates redundant attributes in addition to contradictory examples from training data set that make complex detection model. Author also addresses some difficulties of data mining such as handling continuous attribute, missing attribute values and reducing noise in training data. This model used Knowledge Discovery Data Mining (KDD) CUP 99 dataset for experiment.

Bharti K K, Shukla S and Jain S [3], a number of techniques are available for IDS. Data mining methods are the proficient methods available for IDS. Data mining techniques may be supervised or unsupervised. Various researchers have applied various clustering algorithm for IDS, but all of these are trouble from class ascendancy, force assignment and No Class problem. This work proposed a model that is based on feature selection (as a first phase), K-Means clustering model generation (as a second phase) and classification (as a third phase). This model used CfsSubSetEval method along with BestFirst search for feature selection. In the final phase of the proposed model i.e. classification phase, author used J48 and Random Forest. To evaluate the performances of proposed model KDD Cup 1999 dataset is used. This model used precision and recall as a performance metric.

Panda M, Patra M R [4] evaluated the performance of different rule based classifiers like: NNge (Non-Nested Generalized Exemplars), RIDOR (Ripple-Down Rules), JRip (Extended Repeated Incremental Pruning) and Decision Table using ensemble methods in order to make a proficient network IDS, by combining AdaBoost with different base learners. This model used KDD Cup 99 data set for performing experiment.

2.1 Some Popular Data Mining Methods used in Various Researches:

K-Means: The K-Means algorithm is one of the most popular methods of clustering analysis that aims to partition ‘n’ data objects into ‘k’ clusters in which each data object belongs to the cluster with the nearest mean. It uses Euclidean metric as a similarity measure.

The basic algorithm is:

1. Select k objects as initial centroids.
2. Assign each object to the closest centroid.
3. Recalculate the centroid of each cluster.
4. Repeat steps 2 and 3 until centroids do not 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.

ID3: ID3 (Iterative Dichotomiser 3) invented by Ross Quinlan used to generate a decision tree from a data set and also a precursor to the C4.5 algorithm. It uses the entropy of attributes.

The algorithm can be summarized as:

1. Take all attributes using the data set S and calculate their entropies.
2. By using the attribute which has minimum entropy split the data set S into subsets.
3. Make a decision tree node containing that attribute.
4. By using remaining attributes recurs on subsets.

Important properties of ID3 algorithm:

1. Usually produces small trees.

2. Only one attribute is tested at a time for making a decision.
3. Classifying continuous data may be computationally expensive.

Naive Bayes: Naive Bayes is one of the most efficient learning algorithms. It is based on a strong independence assumption with quite simple construction. It analyzes the correlation between independent variable and dependent variable to obtain a conditional probability for every correlation. By using Bayes theorem we write:

$$P(H|D) = P(D|H)\,P(H) / P(D)$$

Here D may be a data record and H is a hypothesis represents data record D. P (H|D) is the posterior probability of H conditioned on D and P (H) is the prior probability. Similarly, P (D|H) is the posterior probability of D conditioned on H.

Important properties of Naive Bayes algorithm:

1. It is very easy to construct and training is also easy and fast.
2. Highly scalable.

K-NN: K-NN (K-Nearest Neighbor) is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure. It is a type of instance based learning or lazy learning. It uses Euclidean distance as a distance metric.

The K-NN algorithm for determining the class of a new object C:

1. Calculate the distance between object C and all objects in the training data set.
2. Select K-nearest objects to C in the training data set.
3. Assign C to the most common class among its K-nearest neighbors.

Important properties of K-NN algorithm:

1. It is simple to implement and use.
2. Needs lot of space to store all objects.

3. Issues in Traditional IDS

The proposed work also aims to investigate different issues over IDS. When an intruder attacks a system, the ideal response would be to stop his activity before any damage or access to sensitive information occurs. This would require recognition of the attack as it takes place in real time. A single instance of suspect behavior on a host in a network may not warrant any serious action [10]. However, repeated suspect behavior across several hosts in a network may indeed suggest an attack, with a response definitely warranted. This would be very difficult for a human to recognize because IDS typically produce a lot of false positive alerts [7].

In intrusion detection, removal of unrelated data requires the knowledge of external environment variables like human interference, knowledge of the network, vocabulary etc. Application of traditional data mining algorithms on high dimensional data does not provide optimal results as they look at a single dimension. With the number of intrusion and hacking or attacking incidents around the world on ascend; the significance of having responsible IDS in place is better than ever. Having an Intrusion Detection System (IDS) does not solve the problem as organizations frequently have problems addressing the challenges of detecting, alerting and responding to unauthorized access into their computing environment [9]. One of the major challenges faced in today's IDS is its ability to effectively measure their performance. Measure of the effectiveness of intrusion detection refers to its ability to efficiently and correctly classify the events as being malicious or not.

Typically malicious activities create abnormal behavioral patterns, for example, a lot of Internet Control Message Protocol (ICMP) host unreachable messages in a network are abnormal. These activities can be considered to be an alert [6]. A false positive alert in IDS is an attack alarm that is raised incorrectly. This IDS alert is a false positive if ICMP HOST unreachable error is raised because of faulty router. As the overall number of

alerts generated by an IDS is overwhelming for a human operator to handle, there is a need to reduce the alerts that falsely indicate security issues. These alerts require investigation of audit events for diagnosis by human operators or automated agents. Most of the time the diagnosis information associated with the alerts is so poor that it requires the operator to go back to the original data source to understand the diagnosis and assess the real severity of the alert. Care must also be taken to remove or minimize the number of undetected attacks (false negatives) in order to increase the effectiveness of Intrusion Detection [13].

4. Discussion

This section discusses the limitations and advantages of various existing methods:

An intrusion detection system based on genetic algorithm approach has proposed by author [17]. This approach showed a reasonable detection rate but can be improved by using better equations or heuristic in the detection process.

A hybrid learning approach [16] that combines Naive Bayesian classifier and ID3 algorithm overcomes the problem of moderate detection rate and false positives but needs improvement for false positives in remote to user (R2L) attack.

A Y-means clustering algorithm [18] has improved detection rate and low false alarm rate. But it cannot solve the real time anomaly detection, because it cannot revise the data set dynamically during the process.

A clustering based algorithm uses SOM and K-Means [19] overcomes the drawback of traditional SOM which cannot give the precise clustering results, and it also overcomes the drawback of traditional K-Means that depends on the initial value and it is also hard to locate a suitable center of the cluster.

A parallel clustering ensemble algorithm [20] forms the clusters more rapidly to mass data. It also achieves high detection rate but false alarm rate. Because it is an ensemble approach it requires extra memory.

A modified dynamic K-Means algorithm called MDKM [21] has fine detection rate but its false alarm rate is moderate.

A hybrid learning approach [15] that uses K-Means clustering and Naive Bayes classification overcomes the problem of moderate detection rate and high false alarm rate of existing methods.

A hybrid intrusion detection system [11] that combines K-Means and two classifiers: k-nearest neighbor and naive bayes overcome the shortcoming of high false alarm rate in existing method.

Intrusion detection system that combines K-Means, fuzzy neural network and SVM classifiers [22] and by utilizing data mining methods such as neuro-fuzzy and radial basis support vector machine (SVM) for helping IDS to achieve a better detection rate.

Back-propagation neural network based IDS [23] requires a very large amount of data and takes time to ensure the results accuracy.

Boosted decision tree approach [24] for intrusion detection system is an ensemble approach and its detection rate is fine but has moderate false alarm rate. Because it combines a number of decision trees, it becomes complex and needs more time and space.

A triangle area based nearest neighbors approach [25] for intrusion detection system has a reasonable detection rate but unfortunately, a potential drawback of this technique is the false alarm rate.

5. Conclusion

Since the study of intrusion detection began to gain momentum in the security community roughly ten years ago, a number of diverse ideas have emerged for confronting this problem. Intrusion detection systems vary in the sources they use to obtain data and in the specific techniques they employ to analyze this data. Most systems

today classify data either by misuse detection or anomaly detection: each approach has its relative merits and is accompanied by a set of limitations. It is likely not realistic to expect that an intrusion detection system be capable of correctly classifying every event that occurs on a given system. Perfect detection, like perfect security, is simply not an attainable goal given the complexity and rapid evolution of modern systems. An IDS can, still endeavor to hoist the bar for intruder or attackers by dipping the efficacy of big classes of intrusion or attacks and rising the work issue required to get a system compromise. A good intrusion detection system promises to allow greater confidence in the results of and to improve the coverage of intrusion detection, making this a critical component of any comprehensive security architecture.

References

[1] K. Jungwon, J. B. Peter, A. Uwe, G. Julie, T. Gianni and T. Jamie, "Immune System Approaches to Intrusion Detection – A Review", Natural Computing: an international journal, vol. 6, Issue 4, **(2007)** December.
[2] E. J. Derrick, R. W. Tibbs and L. L. Reynolds, "Investigating New Approaches to Data Collection, Management and Analysis for Network Intrusion Detection", ACMSE, Winston-Salem, N. Carolina, USA, **(2007)** March 23-24, pp. 283-287.
[3] K. K. Bharti, S. Shukla and S. Jain, "Intrusion detection using clustering", Special Issue of IJCCT, International Conference [ACCTA-2010], vol. 1, Issue 2, **(2010)** August 3-5, pp. 3-4.
[4] M. Panda and M. R. Patra, "Ensembling Rule Based Classifiers for Detecting Network Intrusions", IEEE International Conference on Advances in Recent Technologies in Communication and Computing, **(2009)**, pp. 19-22.
[5] 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) ISSN: 2248-9622, vol. 4, Issue 5, (Version 1), **(2014)** May, pp. 01-07.
[6] 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), Print ISBN:978-1-4799-2432-5 INSPEC Accession no. 13861274, **(2013)** August 22-25, pp. 1615-1618.
[7] 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, **(2013)**, pp. 51-57.
[8] P. R. Subramanian and J. W. Robinson, "Alert over the attacks of data packet and detect the intruders", Computing, Electronics and Electrical Technologies (ICCEET), IEEE International Conference on ISBN: 978-1-4673-0211-1, **(2012)** March 21-22, pp. 1028-1031.
[9] N. S. Chandolikar and V. D. Nandavadekar, "Efficient algorithm for intrusion attack classification by analyzing KDD Cup 99", Wireless and Optical Communications Networks (WOCN), 2012 Ninth International Conference on ISSN :2151-7681, **(2012)** September 20-22, pp. 1 - 5.
[10] V. Barot and D. Toshniwal, "A New Data Mining Based Hybrid Network Intrusion Detection Model" IEEE International Conference on Print ISBN: 978-1-4673-2148-8, **(2012)** July 18-20.
[11] H. Om and A. Kundu, "A hybrid system for reducing the false alarm rate of anomaly intrusion detection system", Recent Advances in Information Technology (RAIT), IEEE International Conference on Print ISBN:978-1-4577-0694-3, **(2012)** March 15-17, pp. 131-136.
[12] M. R. Thakur and S. Sanyal, "A Multi-Dimensional approach towards Intrusion Detection System" International Journal of Computer Applications, vol. 48, no. 5, **(2012)** June, pp. 34-41.
[13] V. Pathak and V. S. Ananthanarayana, "A novel Multi-Threaded K-Means clustering approach for intrusion detection" Software Engineering and Service Science (ICSESS), IEEE 3rd International Conference on Print ISBN: 978-1-4673-2007-8, **(2012)** June 22-24, pp. 757-760.
[14] P. Wang and J. Q. Wang, "Intrusion Detection System with the Data Mining Technologies" IEEE 3rd International Conference on Print ISBN**:** 978-1-61284-485-5, **(2011)** May.
[15] Z. Muda, W. Yassin, M. N. Sulaiman and N. I. Udzir, "Intrusion Detection based on K-Means Clustering and Naive Bayes Classification", 7th IEEE International Conference on IT in Asia (CITA), **(2011)**.
[16] D. Md. Farid, N. Harbi and M. Z. Rahman, "Combining naive bayes and decision tree for adaptive intrusion detection", International Journal of Network Security & Its Applications (IJNSA), vol. 2, no. 2, **(2010)** April.

[17] M. S. Hoque, Md. A. Mukit and Md. A. N. Bikas, "An implementation of intrusion detection system using genetic algorithm", International Journal of Network Security & Its Applications (IJNSA), vol. 4, no. 2, **(2012)** March.
[18] Y. Guan and A. A. Ghorbani and N. Belacel, "Y-Means: A Clustering Method For Intrusion Detection", In Proceedings of Canadian Conference on Electrical and Computer Engineering, Montreal, Quebec, Canada, IEEE, **(2003)** May 4-7, pp. 1083-1086.
[19] W. Huai-Bin, Y. Hong-Liang, X. Zhi-Jian and Y. Zheng, "A clustering algorithm use SOM and K-Means in Intrusion Detection", International Conference on E-Business and E-Government, IEEE, **(2010)**, pp. 1281-1284.
[20] H. Gao, D. Zhu and X. Wang, "A Parallel Clustering Ensemble Algorithm for Intrusion Detection System", Ninth International Symposium on Distributed Computing and Applications to Business, Engineering and Science, IEEE, **(2010)**, pp. 450-453.
[21] L. Han, "Using a Dynamic K-means Algorithm to Detect Anomaly Activities", Seventh International Conference on Computational Intelligence and Security, IEEE, **(2011)**, pp. 1049-1052.
[22] A. M Chandrashekhar and K. Raghuveer, "Intrusion Detection Technique by using K-means, Fuzzy Neural Network and SVM classifiers", International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, INDIA, **(2013)** January 4-6.
[23] S. T. F. Al-Janabi and H. A. Saeed, "A Neural Network Based Anomaly Intrusion Detection System", Developments in E-Systems Engineering, IEEE, **(2011)**, pp. 221-226.
[24] M. Gudadhe, P. Prasad and K. Wankhade, "A New Data Mining Based Network Intrusion Detection Model" International Conference on Computer and Communication Technology (ICCCT), IEEE, **(2010)**, pp. 731-735.
[25] C. F. Tsai and C. Y Lin, "A triangle area-based nearest neighbors approach to intrusion detection" Pattern Recognition, vol. 43, no. 1, **(2010)**, pp. 222-229.

