International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

Enhancing Security and Event Management Using Association Rule Mining

M. Nithya[1], A. Komathi[2]

[1]M.Phil Scholar, Department of Computer Science & Information Technology, Nadar Saraswathi College of Arts and Science, Theni, Tamilnadu, (India)

Head and Assistant Professor, Department of Computer Science & Information Technology, Nadar Saraswathi College of Arts and Science, Theni District, Tamilnadu, (India)

Abstract: ***Security data and event management system is the industry-specific term to secure the data from the unauthorized one on the collection of knowledge usually log files or event logs from various sources into a central repository for analysis. The design of Security Information and Event Management system and so the rule of algorithm for the correlation analysis. The information flow in and out of the atmosphere, however this information is being accessed, modified, and monitored at totally different points, and the way all the security solutions relate to every alternative in several things. Varied association rules to find normal and abnormal patterns with attack types. Here the system is to calculate the difficulty level to generate the rules by classification and the association rule to mine the abnormal types. The testing dataset is NSL KDD dataset filtered into 4 anomaly class and one normal class. The dataset is processed using WEKA tool.***

Keywords: SIEM, NSL-KDD Dataset, Classification Rule, Association Rule, Weka Tool

1. Introduction

Introduction of Data Mining

Data mining refers to extracting or "mining" knowledge from large amounts of data. The term is actually a misnomer. Data mining (the analysis step of the "Knowledge Discovery in Database") process, or KDD, a field at the intersection of computer science and statistics, is the process that attempts to discover patterns in large data sets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database system.

Data Mining Elements

Data mining consists of five major elements

- Extract, transform, and load transaction data on to the data warehouse system
- Store and manage the data in a multidimensional database system
- Provide data access to business analysts and information technology professionals
- Analyze the data by application software
- Present the data in useful format such as a graph or table

Intrusion Detection System

Several types of IDS technologies exist due to the variance of network configurations. Each type has advantages and disadvantage in detection, configuration, and cost.

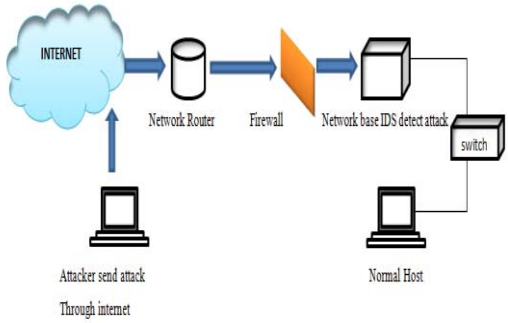

Figure1.1: Network Intrusion Detection System works

1) **NIDS (Network Intrusion Detection Systems)**
2) **HIDS (Host Intrusion Detection Systems)**
3) **Signature Based**
4) **Anomaly Based**

An Introduction to Security and Event Management

Security Information and Event Management (SIEM) is as an umbrella term for the set of activities, tactics and technologies from different disciplines of Information security to proactively protect, prevent and respond to information security incidents which intend to cause harm to the Confidentiality (C), Integrity (I) and Availability (A) of an Information system. Intelligent intrusion detection systems can only be built if there is availability of an effective data set.

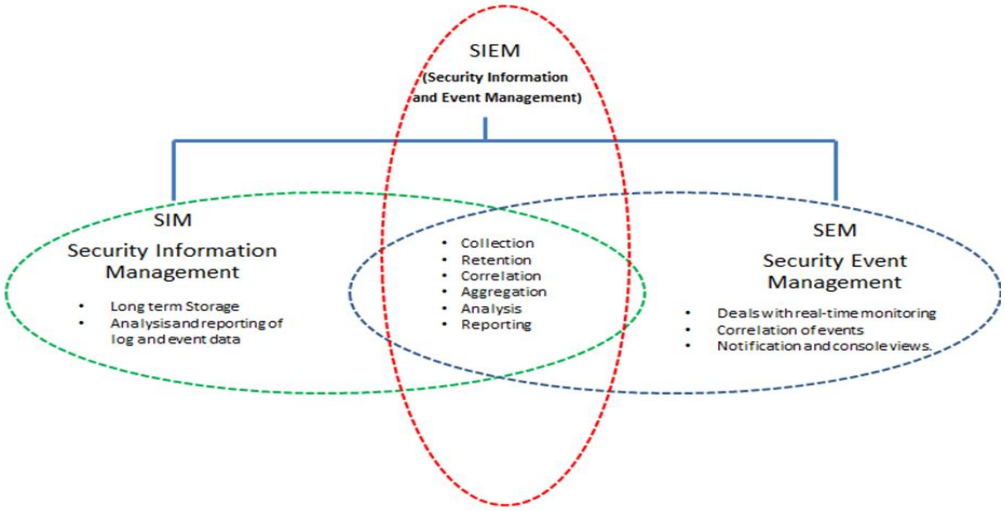

Figure1.2: Logical Anatomy of SIEM

The NSL-KDD data set is a refined version of its predecessor KDD‟99 data set. The monitoring process is automated by an intrusion detection system (IDS). The IDS can be made of combination of hardware and Software. At any point of time a web server can be visited by many clients and they naturally produce heavy traffic.

Each network connection can be visualized as a set of attributes. The traffic data can be logged and be used to study and classify in to normal and abnormal traffic. In order to process the voluminous database, machine learning Techniques can be used. Data mining is the process of extracting interested data from voluminous data sets using machine learning techniques.

To identify relationship existing between various features in a security event for normal and attack scenarios. Once the association has been identified, association rules are generated which are further used to generate classifiers which help of the system to decide that if an event can be an attack or not.

The analysis of the NSL-KDD data set is made by using various algorithms available in the WEKA data mining tool. The NSL-KDD data set is analyzed and categorized into different rule depicting the four common different types of attacks. An depth analytical study is made on the test and training data set. Here the test data set is used.

2. Literature Review

The work done so far in the domain for enhancing security and event Management systems is based one of various methods like self-adaption, artificial neural network and machine learning concepts.

“Self-Learning SIEM System Using Association Rule Mining”, Ravi Raman Tiwari et al. Volume: 04, Issue: 01, June 2015 International Journal of Data Mining Techniques and Applications, ISSN: 2278-2419, Pages: 506-517

Association rule mining method for incorporating self learning in the modern day SIEM systems. The system they proposed can generate classification-based directives and can help the system to take appropriate course of action when a particular set of conditions are met. The classifiers generated using the FP-Growth algorithm is quite effective and capable to identify most of the attack types but they were unable to detect all the attack types. This may be due to the fact that for few attack types sufficient data was not available and the data set may have been containing redundant connection records. For certain attack types for which the data was unambiguous, sufficient and non redundant the classifiers generated were very accurate. They concluded that in order to obtain better results the process should be conducted sufficient and non-redundant data.

“Selection of Relevant Feature for Intrusion Attack Classification by Analyzing KDD Cup 99”, N.S. Chandolikar.et.al, MIT International Journal of Computer Science & Information Technology, Vol. 2, No. 2, Aug. 2012, pp. (85-90) ISSN No. 2230-7621 © MIT Publications

Data mining can improve intrusion based security attacks detection system by adding a new level of observation to detection of network data indifferences. It is highly required to identify appropriate features to categorize into different types of attack. Feature selection abbreviates the size of network data which improve finally performance of intrusion detection system. One R algorithm which is used for experimentation is an efficient algorithm of feature selection in KDD cup dataset. This feature identification helps to improve efficiency of intrusion detection system.

“A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification algorithms”, L.Dhanabal et al Vol. 4, Issue 6, June 2015International Journal of Advanced Research in Computer and Communication Engineering, ISSN (Online) 2278-1021

The analysis results on the NSL-KDD dataset show that it is a best candidate data set to simulate and test the performance of IDS. The CFS method for dimensionality reduction reduces the detection time and increase the accuracy rate. This analysis conducted on the NSL-KDD dataset with the help of figures and tables helps the researcher to have clear

understanding of the dataset. It also brings to light that most of the attacks are launched using the inherent drawbacks of the TCP protocol. In future, it is proposed to conduct an exploration on the possibility of employing optimizing techniques to develop an intrusion detection model having a better accuracy rate.

"Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection ", S. Revathi et al., International Journal of Engineering Research & Technology (IJERT), Vol. 2 Issue 12, December – 2013, ISSN: 2278-0181.

In this paper, analyzed the NSLKDD dataset that solves some of the issues of KDD cup99 data. The analysis shows that NSL-KDD dataset is very ideal for comparing different intrusion detection models. Using all the 41 features in the dataset to evaluate the intrusive patterns may leads to time consuming and it also reduce performance degradation of the system. Some of the features in the dataset are redundant and irrelevant for the process.

CFS Subset is used to reduce the dimensionality of the dataset. The experiment has been carried out with different classification algorithms for the dataset with and without feature reduction and it's clear that Random Forest shows a high test accuracy compared to all other algorithms in both the cases. So in the case of reduced feature set this analysis shows that Random Forest is speeding up the training and the testing methods for intrusion detection that is very essential for the network application with a high speed and even providing utmost testing accuracy. In future we can try to improve the Random Forest algorithm to build an efficient intrusion detection system.

3. Problem Definition

The NSL KDD dataset are the dataset contains the information how the system or Network is intruded by the hackers or attackers. The dataset which contain only the details of anomaly and normal classes in both the testing and training type where the existing works are made on those types.

The proposed works are related with the anomaly class with the different types of class such as Dos, Probe, U2R, R2L attacks and these types of attacks are in the testing data of NSL KDD dataset with 21 difficulty level in TXT format.

The experiment made on the testing data to calculate the difficulty level of those attacks using the classification and the association rule to show the accuracy, confidence, confirmation value, frequency of counter instances, ROC analysis using Apriori, Predictive Apriori, Tertius.

4. Methodology

Data Preprocessing

One of the most critical steps in the data mining process is the preparation and transformation of the initial dataset. Raw data are seldom used for data mining; many transformations may be needed to produce features more useful for selected data mining methods such as prediction or classification. In the real world of data mining application, the situation is reversed. More effect is expended preparing data than applying data mining methods.

Importance of Preprocessing

Data in the real world is dirty

- Incomplete
- Noisy"
- Inconsistent

Measures of Data Quality

A well-accepted multidimensional view of data quality:

- Accuracy
- Completeness
- Consistency
- Timeliness
- Believability
- Interpretability
- Accessibility

Major Tasks in Data Preprocessing

a) **Data cleaning:** Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
b) **Data integration:** Integration of multiple databases, data cubes, or files
c) **Data transformation:** Normalization and aggregation
d) **Data reduction:** Obtains reduced representation in volume but produces the same or similar analytical results
e) **Data discretization:** Part of data reduction but reduces the number of values of the attributes; particular importance especially for numerical data

Data cleaning tasks

Fill in missing values
Identify outliers and smooth out noisy data
Correct inconsistent data
Resolve redundancy caused by data integration

Dataset Description

The inherent drawbacks in the KDD cup 99 dataset has been revealed by various statistical analyses has affected the detection accuracy of many IDS modeled by researchers. NSL-KDD data set is a refined version of its predecessor. It contains essential records of the complete KDD data set. There are a collection of downloadable files at the disposal for the researchers. They are listed in the Table 4.1

Table 4.1: List of NSL-KDD Data set files and their description

S. No	*Name of the file*	*Description*
1.	KDDTrain+ .ARFF	The full NSL-KDD train set with binary labels in ARFF format
2.	KDDTrain+.TXT	The full NSL-KDD train set including attack type labels and difficulty level in CSV format
3.	KDDTrain+_20Percent.ARFF	A 20% subset of the KDDTrain+ .ARFF file
4.	KDDTrain+_20Percent.TXT	A 20% subset of the KDDTrain+.TXT file
5.	KDDTest+ .ARFF	The full NSL-KDD test set with binary labels in ARFF format
6.	KDDTest+.TXT	The full NSL-KDD test set including attack type labels and difficulty level in CSV format
7.	KDDTest-21+.ARFF	A subset of the KDDTest+ .ARFF file which does not include records with difficulty level of 21 out of 21
8.	KDDTest-21+.TXT	A subset of the KDDTest+.TXT file which does not include records with difficulty level of 21 out of 21

In each record there are 41 attributes unfolding different features of the flow and a label assigned to each either as an attack type or as normal.

The 42nd attribute contains data about the various 5 classes of network connection vectors and they are categorized as one normal class and four attack class. The 4 attack classes are further grouped as DoS, Probe, R2L and U2R.The description of the attack classes

The attack classes present in the NSL-KDD data set are grouped into four categories:
1. DOS
2. Probing
3. U2R
4. R2L

5. Classification and Prediction

Classification
Data Classification is a two step process consisting of a learning step and a Classification step. A classifier is built describing a predetermined set of classes or concepts. This is the learning step, where a Classification builds the classifier by analyzing or learning form a training set made up database tuples and their associated class labels.
- M5 rules

Prediction
Predicting is to identify one thing based purely on the description of another or related thing. Not necessarily future events, just unknowns based on the relationship between thing that know and a thing need to predict.
- Apriori
- Predictive Apriori
- Teritius

Converting TXT into CSV

The testing data of NSL-KDD dataset are changed into CSV format since the work related with the different types of attacks and the difficulty level of attacks deals in the KDD test dataset consists of only the normal and anomaly.

KDD test TXT format data contains the details of the attack types, and the difficulty level of those attacks. So the TXT format data are covert into CSV format with the help of MS-Excel.

Import the TXT format data into the Excel and the data are separated by comma are act as delimiters to form cells and data are fed and finally save as in CSV Format.

Filtering
The CSV format data set are filtered as normal and anomaly data set, the normal set are kept as it is. But the anomaly dataset are filtered as four different data attack types. NSL-KDD dataset is a reduced version of the original KDD 99 dataset. NSL-KDD consists of the same features as KDD 99. The KDD99 dataset consists of 41 features and one class attribute. The class attribute has 21 classes that fall under four types of attacks:
Probe attacks
User to Root (U2R) attacks
Remote to Local (R2L) attacks
Denial of Service (DoS) attacks
Anomaly of the data set are filtered to the DoS attack class into a single DoS attack types similarly other attack classes are grouped into their types

6. Sample Rules

M5 Rule
Rule: 1
IF same_srv_rate <= 0.215
count <= 200.5
count <= 149.5
count > 107.5
THEN
dl = -0.0046 * logged_in - 0.006 * count - 0.0215 * same_srv_rate - 0.004 * srv_diff_host_rate + 0.0459 * attack_type = udpstorm, back, teardrop, pod, apache2, processtable, land, smurf, neptune + 0.0026 * attack_type = teardrop, pod, apache2, processtable, land, smurf, neptune + 0.004 * attack_type = processtable, land, smurf, neptune + 21.1728 [1332/35.982%]

Apriori
1. dst_host_srv_diff_host_rate=0 226 ==> num_failed_logins=0 226 conf:(1)
2. num_failed_logins=0 226 ==> dst_host_srv_diff_host_rate=0 226 conf:(1)
3. srv_diff_host_rate=0 221 ==> num_failed_logins=0 221 conf:(1)

4. srv_diff_host_rate=0 221 ==> dst_host_srv_diff_host_rate=0 221 conf:(1)
5.srv_diff_host_rate=0 dst_host_srv_diff_host_rate=0 221 ==> num_failed_logins=0 221 conf:(1).

Predictive Apriori
1. num_failed_logins=0 226 ==> dst_host_srv_diff_host_rate=0 226 acc:(0.99499)
2. dst_host_srv_diff_host_rate=0 226 ==> num_failed_logins=0 226 acc:(0.99499)
3. srv_diff_host_rate=0 221 ==> num_failed_logins=0 dst_host_srv_diff_host_rate=0 221 acc:(0.99499)
4. logged_in=0 202 ==> num_failed_logins=0 dst_host_srv_diff_host_rate=0 202 acc:(0.99499)
5. attack_type=neptune 141 ==> num_failed_logins=0 logged_in=0 141 acc:(0.99497)

Teritus
1. /* 0.168527 0.183908 0.000000 */ dl = 21 ==> attack_type = neptune
2. /* 0.140889 0.071942 0.000000 */ logged_in = 1 ==> count = 4 or attack_type = apache2 or dl = 16
3. /* 0.138771 0.067164 0.000000 */ attack_type = processtable ==> count = 1
4. /* 0.138771 0.067164 0.000000 */ count = 1 ==> attack_type = processtable
5. /* 0.124454 0.257895 0.000000 */ attack_type = neptune ==> logged_in = 0

The overall classification of M5Rules and number of rules generated are shown in these table

Table 7.1: M5 Rules generated

Attacks(instances)	*No. of. Rules generated*	*Time Taken to build*
Dos(20162)	11	12.7 seconds
Probe(2421)	14	2.86 seconds
U2R(67)	1	0.28 seconds
R2L(2887)	9	7.68 seconds

Table 7.2: Association Rules Generated

Attacks	*Apriori (Confidence 1.000) No. of. cycles performed*	*Predictive Apriori (Accuracy)*	*Tertius (Instances/Rules)*
Dos	16	0.99499	200/16
Probe	15	0.99493	100/107
U2R	15	0.99442	67/35
R2L	2	0.99427	50/45

The above table 7.1 shows how the dataset has been classified to generate the number of rules by the M5 rules and time taken to build the model and table 7.2 shows the association rule mining of datasets with the Apriori (Confidence rate:1.000andnumber of cycles performed), Predictive Apriori(Accuracy rate: 0.99499), Tertius(Number of instances/Number of rules with hypotheses explored, considered)

7. Conclusion

The NSL KDD datasets are refined version of KDD cup 99 dataset with the 41 attributes and 42nd attributes contains data about the network connection vectors and they are categorized as one normal class and four attack class.

The dataset attributes of each network connection vector are categorized as

- Basic features
- Content Related features
- Time Related features
- Host Based Traffic features

The NSL KDD datasets contain the test and train datasets in ARFF Format. But the both datasets contain the details of only normal and anomaly type of data. The TXT format dataset contain the details of different types of attacks and also with the difficulty level of those attacks.

The work is done on the TXT format of test datasets with 21 difficulty level by converting TXT format to CSV format and then filtering is made on the datasets to separate the 4 types of attacks (DoS, Probe,U2R, R2L) into 4 types of datasets.

The datasets are loaded into the weka environment and attribute selection method gives the relevant attributes for each type of attacks. With those attributes the M5 rules are used to generate the rules to calculate the difficulty levels of those attacks.
And the attribute types are converted and sample the instances by reservoir sample for generating the Association Rule in

Apriori to show the confidence
Predictive Apriori to show accuracy
Tertius to show the confirmation values, frequency of counter instances and ROC Analysis

8. Future Work

The NSL KDD dataset contain testing and training dataset to show the anomaly and normal classes. To differentiate the attacks in the testing and training dataset the proposed work is made on the testing dataset of TXT format file. In the future the same can be done for training dataset to generate the rules by association rule mining.

More work should be done in the preprocessing stage so that the time taken to build the model can be reduced and the accuracy rate can be increased.

Modification made on the attack types since the data are unambiguous, insufficient and non redundant so that the rules can be generated with more accuracy rate.

FP- Growth algorithm is not used in the proposed work, it will be implemented in the future to generate the rules.

References

[1] http://www.cs.unb.ca/downloasds/iscx for NSL-KDD dataset
[2] Ravi Raman Tiwari et al."Self-Learning SIEM System Using Association Rule Mining", Volume: 04, Issue: 01, June 2015 International Journal of Data Mining Techniques and Applications, ISSN: 2278-2419, Pages: 506-517.

[3] L.Dhanabal et al ."A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification algorithms", Vol. 4, Issue 6, June 2015International Journal of Advanced Research in Computer and Communication Engineering, ISSN (Online) 2278-1021.
[4] Guillermo Suarez–Tangil et al. "Providing SIEM Systems with Self–Adaptation", January, 2013, University of Madrid.
[5] Hongyan Liu et al. "Mining Frequent Patterns from Very High Dimensional Data: A Top-Down Row Enumeration Approach".
[6] Hemant Khandelwal ."An Approach for Information Security using Data Manning Technique", IPASJ International Journal of Information Technology (IIJIT) Volume 3, Issue 8, August 2015 ISSN 2321-5976
[7] Karan Bajaj et al. "Dimension Reduction in Intrusion Detection Features Using Discriminative Machine Learning Approach".
[8] S. Revathi et al. "Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection", International Journal of Engineering Research & Technology (IJERT), Vol. 2 Issue 12, December – 2013, ISSN: 2278-0181.
[9] Pooja Bhoria et al. "Determining feature set of DOS attacks", International Journal of Advanced Research in Computer Science and Software Engineering , Volume 3, Issue 5, May 2013, ISSN: 2277 128X, pp. 875-878
[10] Meenu Choudhary et al. "Performance Analysis Of Data Reduction Algorithms Using Attribute Selection In Nsl-Kdd Dataset", [IJESAT] [International Journal of Engineering Science & Advanced Technology] Volume-4, Issue-2, , ISSN: 2250-3676 Page no 214-219.
[11] Mohammad Khubeb Siddiqui et al. "Analysis of KDD CUP 99 Dataset using Clustering based Data Mining", International Journal of Database Theory and Application, Vol.6, No.5 (2013), pp.23-34.
[12] Vinod Rampure et al. "A Rough Set Based Feature Selection on KDD CUP 99 Data Set", International Journal of Database Theory and Application Vol.8, No.1 (2015), pp.149-156.
[13] Laheeb M.Ibrahim et al. "A Comparison Study For Intrusion Database (KDD99, NSL-KDD) Based On Self Organization Map (SOM) Artificial Neural Network", Journal of Engineering Science and Technology, vol. 8, no. 1 (2013) , page no 107 – 119.
[14] S. Saravanakumar et al. "Development and Implementation of Artificial Neural Networks for Intrusion Detection in Computer Network", IJCSNS International Journal of Computer Science and Network Security, VOL.10 No.7, July 2010, page no 271-275.
[15] Neesha Sharma et al. "Comparative analysis of association rule mining algorithms", IJSRD - International Journal for Scientific Research & Development Vol. 2, Issue 2, 2014 , ISSN (online): 2321-0613.
[16] Mukesh Sharma et al. "Evaluating the performance of apriori and predictive apriori algorithm to find new association rules based on the statistical measures of data sets", International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 6, August – 2012, ISSN: 2278-0181.
[17] N.S. Chandolikar.et.al. "Selection of Relevant Feature for Intrusion Attack Classification by Analyzing KDD Cup 99", MIT International Journal of Computer Science & Information Technology, Vol. 2, No. 2, Aug. 2012, pp. (85-90) ISSN No. 2230-7621 © MIT Publications
[18] Preeti Aggarwal et al. "Analysis of KDD Dataset Attributes-Class wise For Intrusion Detection" , 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015)
[19] Przemysław Kukiełka et al. "Adaptation of the neural network-based IDS to new attacks detection".
[20] Paresh Tanna et al. "Using Apriori with WEKA for Frequent Pattern Mining", International Journal of Engineering Trends and Technology (IJETT) – Volume 12 Number 3 - Jun 2014

Author Profile

M.Nithya is a M.Phil. scholar, completed her Master degree in Computer Science and Information Technology. She is the receiver of PG Merit Scholarship from UGC for the University Rank Holders for the year 2013 – 2015. She already published 4 paper in a journal and presented 4 paper in conferences and published. She attended many conferences at national and international levels, workshops and seminars. Her area of interest is Data Mining, Network security, Software engineering and testing.

A.Komathi completed her M.C.A., M.Phil. now doing Ph.D. in Bharathiar University. Now she is working as Vice Principal, Head and Assistant professor in Department of CS&IT, Nadar Saraswathi College of Arts and Science, Theni. Her area of interestsis Data Structure, Biometric Authentication, Data Compression and Wireless Sensor Network. She organized and also attended many conferences, seminars and workshops. So far presented 18 papers in international conference and published 12 papers in the international journals. She is member of staff selection committee in NSCAS, Board of studies member in MTWU. Her area of research is Wireless Sensor Network. She served as research supervisor for M.Phil. Scholars.