International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 07 | July-2018 www.irjet.net p-ISSN: 2395-0072

A Comprehensive Study of Association Rule Mining using NSL-KDD Dataset for Intrusion Detection

Vasanthi K[1]., Vinutha H.P.[2]

[1]Student, Dept. of Computer Science and Engineering, BIET College, Karnataka, India.
[2]Assistant Professor, Dept. of Computer Science and Engineering, BIET College, Karnataka, India.

Abstract- Network intrusion detection includes a set of malicious actions that compromise the integrity, confidentiality and availability of information resources. Several techniques for mining rules from KDD intrusion detection dataset enables to identify attacks in the network. But little research has been done to determine the association patterns that exist by the attributes in the dataset. In this proposed method focus on the association rule mining using NSL KDD'99 intrusion dataset. Since the dataset constitutes different kinds of data like binary, discrete & continuous data, same technique cannot be applied to determine the association patterns. The proposed method is used to generate attack rules that will detect the attacks in network audit data using anomaly detection. Rules are formed depending upon various attack types. Apriori approach is used to eliminate the non-frequent item set from the binary, discrete and continuous values of the dataset that are used.

Keywords: Data Mining[1], NSL-KDD dataset[2], Feature selection[3], Apriori algorithm[4], Association Rules[5].

1. INTRODUCTION

Intrusion Detection System (IDS) is used to monitor network traffic and suspicious activity and alerts the system or network administrator. In some cases, the IDS may also respond to anomalous or malicious traffic by taking action such as blocking the user or source IP address from accessing the network [1]. There are network based and host based intrusion detection systems. Some IDS detect by looking for specific signatures of the known threats similar to the way antivirus software typically detects and protects against malware. There are also a different type of IDS that detect attacks based on comparing traffic patterns against a baseline and looking for anomalies. There are IDS that simply monitor and alert whereas some other IDS perform an action or actions in response to a detected threat [2].

NSL-KDD dataset is a simulated dataset that is used for research work. KDD dataset covers four important classes of attacks: Denial of Service (DOS), user-to-root (U2R), remote-to-local (R2L) and probing attack. The selected network simulator dataset that NSL KDD dataset is separated into labelled and unlabelled records. Each marked record comprised of 41 properties and one target esteem. Association rule mining is for the most part used to locate the interesting related logical rules from a huge database relying on the client characterized support and certainty.

1.1 Categories of Intrusion Detection System

IDS can be classified into the following two categories:

- **Signature Based IDS**

It means that, it regulates the small amount of data packets which has been used during the period of usage of interconnections and which perform analysis of the recognized packets and packets which has been stored in the collection of data referred as the dataset. The work of signature positioned intrusion detection system is that it simulates the work behavior of antivirus software. But the only difference is that we always compare the data packets which has been identified during the period of interconnections usage and with the packets which has been already analyzed and stored in the collection of data known as dataset [4].

- **Anomaly Based IDS**

Anomaly based intrusion detection system it is inconsistency stationed, display arranges action and corresponding sequence activities and thinks across a built up standard. The system identifies the custom standard of the system, also tells about the type of data to be transferred, rules which has been utilized, the standard port ad user interface which has been used and signals the user when action is recognized as strange, or conditionally different from normal activities, than the standard procedure. Intrusion detection system approach is mainly designed to identify the activities which is anomaly than the standard activity [5].

2. LITERATURE SURVEY

HuyAnh Nguyen et.al [1] in 2008 proposed on Classifier Alternative Procedure. As framework horrendous risks has been extended in entirety and validity over the cross as of late years, assaults unmistakable confirmation structure is intensely changing into a crucial part to secure the framework. In context of liberal volumes of security survey data and furthermore stunning and dynamic properties of intrusion manages, overhauling execution of IDS changes into an essential open issue that is getting powerfully thought from the examination gathering. The weakness to research if certain checks to excellent part strike classes constitutes the motivation for the nitty along these lines. They survey execution of a vigilant course of action of classifier checks the selected dataset. In perspective of assessment happens, most important figuring's for each assault game plan is picked, selected and categorical estimation decision built up framework has been identified. It's reenactment output examination shows that recognizable execution change and relentless interference distinguishing proof can be ace as they apply the proposed models to see different kinds of framework strikes.

Hui Wang et.al [2] in 2009 proposed on Mining Association regulations related to assault detection process or system. Assault or assault malicious activities identification is a most powerful weapon to shield interconnections from assaults and has expanded progressively thought. Data mining has been exhibited as a crucial system to perceive intrusions. It has been the ebb and flow inquire about focus and example to apply data mining procedures in assault or assault identification activities for finding new sorts of assaults, in any case it is still in its soonest arranges. They overviews the new change of association regulations and syntax rules mining progresses for intrusion detection in wired and furthermore remote interconnections. The troubles and impelled changes of advances to use association lead mining for intrusion detection are discussed.

Lei Li et.al [3] in 2010 proposed on framework security is changing into an unquestionably basic main problem, during from that period the quick advance of the collections of the autonomous systems interconnections. Framework assault or assault activities identification system, it is the key redemption ensuring procedure, it is mainly for the most part utilized opposite to poisonous strikes. Learning the provided dataset and understanding the pattern trained dataset progress made generally related in arrange intrusion acknowledgment and unpleasantness structures by discovering customer individual direct benchmarks from the framework change data. Connection rules, regulations, related threshold and development regulations, and provisional rules are the most typical and custom game plan of data burrowing for interference area. Acknowledge the created Apriori estimation related to deadlock of dynamic thing sets of pre-processing the dataset, it suggested the duration reducing to see assaults in perspective of data searching or processing, it is the most improved Apriori figuring. Examination works out obviously demonstrate that the proposed technique is productive.

Hee-suChae et.al [4] in 2013 proposed on attribute selection analyzer for the chosen standard NSL KDD dataset. These days, interconnections development is extending a result of the growing usage of sagacious devices and the Internet. Measure of the assault identification contemplates focused mainly above the selection of properties the dataset and decreasing the huge dataset to small amount of dataset during the period of part related to the properties are insignificant more over abundance it happens broad assault identification activities and spoils program execution of an assault identification activities. The main aim behind here examination conducted mainly to perceive crucial picked properties to mainly built up the assault identification system activities based upon the statistics capable, flexible with general dataset and cogent. Based upon evaluation of the execution related to selection of attributes or properties from the selected dataset procedures; those are gain ratio, asymmetric, information gain, chi square and correlation based properties chosen. They propose another properties chosen strategy by utilizing the most custom and typical based on the available properties. They apply one of the compelling classifier decision tree count for surveying feature reduction strategy. They take a gander at between proposed procedure and distinctive systems.

Kamini Nalavade et.al [5] in 2014 Proposed work on Finding Frequent Item sets using Apriori Algorithm to Detect Intrusions in Large Dataset. With the improvement of hacking and manhandling mechanical assemblies and making of better methodologies for intrusion, Intrusion detection and shirking is transforming into the critical test in the domain of network security. The growing network action and data on Internet is making this errand moreover asking. There are distinctive procedures being utilized as a part of intrusion detections, however grievously any of the systems so far isn't absolutely faultless. The false positive rates makes it to an extraordinary degree hard to separate and react to attacks. Intrusion detection systems using data mining approaches make it possible to look for illustrations and rules in

tremendous measure of audit data. They address a model to arrange association rules to intrusion detection to plot and execute a network intrusion detection system. Their technique is used to create ambush rules that will recognize the attacks in network survey data using anomaly detection. This exhibits the association rules mining estimation is prepared for perceiving network intrusions. The KDD dataset which is energetically open online is used for our experimentation and results are inspected. Their point is to investigate diverse roads with respect to unmistakable parameters of apriori computation to produce a string intrusion detection system using association manage mining.

L.Dhanabal1 et.al [6] in 2015 Proposed tackle a brief analysis on NSL KDD Dataset for Assault or attack identification activities which is stand upon the variety of Classification Algorithms. The most excellent attack or assault identification activities should be developed if at all existence of openness of a productive selected standard dataset. The selected dataset related to network simulator with a countable measure of significant worth information which mirrors or reflects the progressing that can be simply train and assert an assault identification activities. The chosen network simulator NSL KDD data set is a refined type of its precursor KDD"99 data set. The NSL-KDD data set is penniless down and it is mainly used to inspect the feasibility of the distinctive categorical classification computations in perceiving the irregularities in the collection of system movement outlines. It is furthermore examined the connections of the traditions present in typically utilized collection of system tradition piled up horizontally or vertically with the assaults used by third person or attackers to make odd collection of system action. The examination mainly performed based upon the categorical division of the type of assault counts open in the wide famous tool for data analysis that is WEKA. The examination which has revealed various substances related to the holding among the traditions and interconnection assaults.

Noureldien An et.al [7] in 2016 proposed Classifier Alternative Procedure. As framework horrendous risks has been extended in entirety and validity over the cross as of late years, assaults unmistakable confirmation structure is intensely changing into a crucial part to secure the framework. It has been the ebb and flow inquire about focus and example to apply data mining procedures in assault or assault identification activities for finding new sorts of assaults, in any case it is still in its soonest arranges. Connection rules, regulations, related threshold and development regulations, and provisional rules are the most typical and custom game plan of data burrowing for interference area. They showed the relevance of every component in NSL KDD assault or attack activities identification of selected data set until the identification of the proposed categorical class of assault. Brutal group level related to dependence whereas dependence extent of single class of subset are utilized mainly to select the utmost isolating properties of categorical class. The main aim behind here examination conducted mainly to perceive crucial picked properties to mainly built up the assault identification system activities based upon the statistics capable, flexible with general dataset and cogent. Based upon evaluation of the execution related to selection of attributes or properties from the selected dataset procedures; those are gain ratio, asymmetric, information gain, chi square and correlation based properties chosen

RashmiRavindra et.al [8] in 2017 drove manage Assault or assault identification activities: Classification, Techniques and Datasets to Implement. With the increasing speed of the web, Security of interconnections action is transforming into an essential issue of PC interconnections system. As time is passing the amount of assaults on the interconnections are extending. Such assaults on interconnections are just the Intrusions. Assault or assault identification activities has been used for distinguishing intrusion and to shield the data and interconnections from assaults. Data mining strategies are used to screen and separate considerable measure of interconnections data and gathering these interconnections data into peculiar and normal data. Since data begins from various sources, interconnections action is tremendous. Data mining techniques, for instance, classification and gathering are associated with produce or assault identification activities. They demonstrates the classification of IDS, unmistakable Data mining strategies and datasets for the fruitful detection of case for both malevolent and conventional activities in interconnections, which makes secure data system.

3. PROBLEM STATEMENT

Vindictive or malicious activities on the system review data creates intrusion on the system. Contingent on the different sort of intrusion rules are to be created for binary data. Preprocess the dataset still consist of some redundant values. To avoid duplicate attributes the proper way of analyzing the dataset has been applied. To generate association rules based on the input file provided in the format of attribute record file format.

3.1 ELABORATION OF PROBLEM STATEMENT

- Furthermore the present IDS system will identify the type of vicious intrusion but the lack of features of prediction model.
- Current IDS system organizes the assault of vicious activities but does not give the bounded model of the enormous amount of data present in the KDD dataset. Bounded model consist of train standard KDD dataset about the range of minimum and maximum values of the each and every attribute.
- Computational data set which is huge and enormous for the most part contain uproarious, unlimited, or uninformative features which leads to basic roadblocks to data expose and data demonstrating.
- Current data pre-processing may avoid the duplicate elements, but to remove the least compelling attributes from the selected dataset there should be an analyzer which should identify the quality and rank of the properties.

4. PROPOSED METHODOLOGY

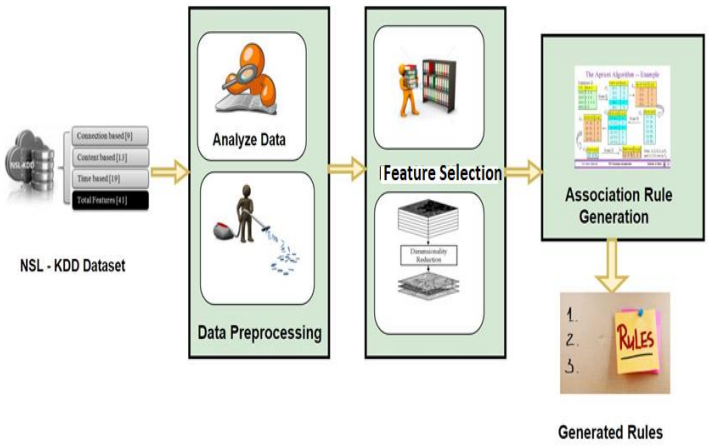


Fig.4.1: diagram of the proposed system

Figure 4.1 gives the design outline in this strategy the first and essential step is to pre-process the information. Since got dataset may contains numerous fragmented, good for nothing, inconsistence and boisterous information. At the phase of pre-processing NSL-KDD dataset the nonnumeric features are to be changed over into numeric element aside from convention write, administration and banner. Attack names must be related with its class like typical, DoS, Probe, R2L and U2R. In NSL-KDD dataset features may contain discrete or nonstop qualities these features are standardized with min-max ranges. In the wake of preprocessing we can separate the vital features by applying distinctive component extraction strategies. In each element extraction techniques every one of the 41 qualities are positioned with the assistance of rankers calculation alongside include choice strategy. Dimensionality can be diminished by expelling the superfluous features. At that point ensemble approaches are connected for various calculation as a base calculation.

Input NSL-KDD:

NSL-KDD will be provided as input. Which contains 41 attributes related values as input. Some examples of the attributes which is associated are as follows. Below diagram illustrates the labels of the properties which is present in the selected the dataset. The main idea behind choosing the named dataset is standard data means it always consist of the named data. Whereas the test data means absence of the named properties. These datasets play a major role in the prediction model.

Table 1: NSL-KDD Dataset Attributes

Toatal Attribute		
Duration	su_attempted	same_srv_rate
protocol_type	num_root	diff_srv_rate
service	num_file_creation	srv_diff_host_rate
flag	num_shells	dst_host_count
src_byte	num_access_file	dst_host_srv_count
dst_byte	num_outbound cmds	dst_host_same_srv_rate
land	is_host_login	dst_host_diff_srv_rate
wrong_fragment	is_gust_login	dst_host_same_src_port_rate
urgent	count	dst_host_srv_diff_host_rate
hot	srv_count	dst_host_serror_rate
num_failed_login	serror_rate	dst_host_srv_serro_rate
logged in	srv_serror_rate	dst_host_rerror_rate
num_compromised	rerror_rate	dst_host_srv_rerror_rate
root_shell	srv_rerror_rate	class

Fig 4.2: Properties names which is present in the chosen dataset

Analyze Data:

Here procedure follows the pre-processing of the collected dataset. It mainly identifies the duplicate properties which is present in the dataset and removes the unwanted data from the current dataset. It follows different steps that has

been illustrated in the pre-processing step. Properties selection analyzer procedure will applied to get the modified dataset of the selected dataset. Properties selection analyzer will remove the least significant properties from the dataset. So the modified dataset always consist of most significant properties. If input of dataset consist of the most significant properties means it always generates the meaningful and most appropriate regulations.

Feature Selection:

To build an Intrusion detection model feature selection is one of the most important step. It is the one of the important data pre-processing technique in Data Mining. Feature selection is also known as an attributes selection method. It is mainly applied to generate the dataset which consist of most significant properties. It utilizes the ranker method to generate the ranks corresponding to each properties. There are plenty of properties selection methods are available online. Out of which it should select the most optimum and best properties selection procedure. Procedure is also utilizing the ranker method it will generate the corresponding ranks for each properties. Here we have to use the chi-square algorithm and also applying the rankers method. Most important properties will get less rank and least significant properties will get high rank. Here less rank means most significant attribute and high rank means least significant attribute. Output of properties selection procedure will be the comma separated file which consist of properties names, rank and metric.

Feature selection attributes analyzer

Output of best properties identification and selection method is text files depicting ranks associated with 41 attributes and it average merit. To remove the low ranked features from the attribute list, we need to manually select the least valued or least ranked attribute and remove from the KDD data set. To avoid this manual work, we have implemented Feature Selection Attribute Analyzer.

Attribute analyzer accepts NSL KDD dataset as an input. Chi Square best properties identification and selection algorithm will be applied for the dataset. Ranker method will be applied for the NSL KDD dataset. It will generate the corresponding ranks associated for the 41 attributes that are defined in the NSL KDD dataset. Feature selection attribute analyzer will remove the redundant properties from the chosen dataset. It will remove the least ranked properties from the chosen dataset. The modified dataset consist of most significant ranked attributes.

Feature selection attribute analyzer accepts best properties identification and selection output as an input. Best properties identification and selection utilizes the ranker method and generates the ranks for all the 41 attributes from the chosen dataset. Best properties identification and selection algorithm generates most significant ranks for the important attributes. Whereas least ranks applied for the least significant attributes. Best properties identification and selection algorithm chosen is Chi Squared algorithm. Best properties identification and selection output is considered as an input for the analyzer. Procedure of attribute selection analyzer will remove the least significant attributes from the NSL KDD dataset. It will modify the selected interconnections simulator dataset and removes the redundant data from the dataset.

Output of feature selection method is text files depicting ranks associated with 41 attributes and it average merit. To remove the low ranked features from the attribute list, we need to manually select the least valued or least ranked attribute and remove from the KDD data set. To avoid this manual work, we have implemented Feature Selection Attribute Analyzer.

Steps for attribute selection analyzer are as follows:

Step 1: Property selection method output will be considered as an input for the attribute selection analyzer. Output of property selection will be in the format of comma separated file which consist of properties names, associated ranks and metric. The selected interconnections simulator dataset also be considered as an input for this step.

Step 2: Attribute selection analyzer will read the input provided. Here we already mentioned that input file consist of properties names and corresponding rank and metric. Based on the rank provided, most significant properties which is having low rank and least significant properties which is having high rank, few properties will be eliminated. Most of the time properties having least significant will be eliminated from the file. It removes the label as well as the corresponding data from the chosen information.

Step 3: Modified file will be created at the particular path. Modified file means least valued properties have been deleted from the chosen information. New modified file can be created or the present file can be updated by eliminating the properties. Options provided to the user.

The modified NSL KDD dataset is the most optimized dataset for further calculations such as classification and clustering. The modified NSL KDD dataset will be provided as an input for the association rule generator. One of the most important and major advantage from the feature selection attribute analyzer is that it will avoid the manual work and avoid errors while removing the particular columns or attributes from the dataset. Feature selection attribute analyzer provides the most optimized dataset that can be applied for the further classification and clustering.

Association Rules Generator:

Before applying the association rules generator, system should make sure that the data modified by the properties selection analyzer is of the correct format. To get successfully the association rules the modified dataset generated by the properties selection analyzer must be in nonnumeric format. Because apriori algorithm doesn't work well with the data consist of numeric values. It is considered as one of the limitation. To escape from this problem system will run the discretization to convert from data of format numeric to nonnumeric type. Once data set has been set properly, it can be provided as an input for the rules generator. Generates association rules by observing the statistical analysis done by using R studio 3.4.2. The output is based on the input provided, corresponding standard associations regulations and the rules are the output of the proposed system.

To generate association rules NSL KDD dataset, system should analyse the standard dataset file first. As NSL KDD dataset is huge, first system should analyse all the 42 attributes minimum value and maximum value range. To study NSL KDD dataset attributes minimum and maximum value, R-Studio has been taken into account. NSL KDD dataset in format of CSV (Comma separated value) provided as input and corresponding summary will be generated as output. Summary consist of relative all the attributes minimum, mean and maximum value.

To read table and to get the summary of the NSL KDD dataset following command line will be used.

NHIS<read.table("C:\\KDDTrain+_20Percentuse_normal.csv",header=T,sep=",")

Head (NHIS)

Summary (NHIS)

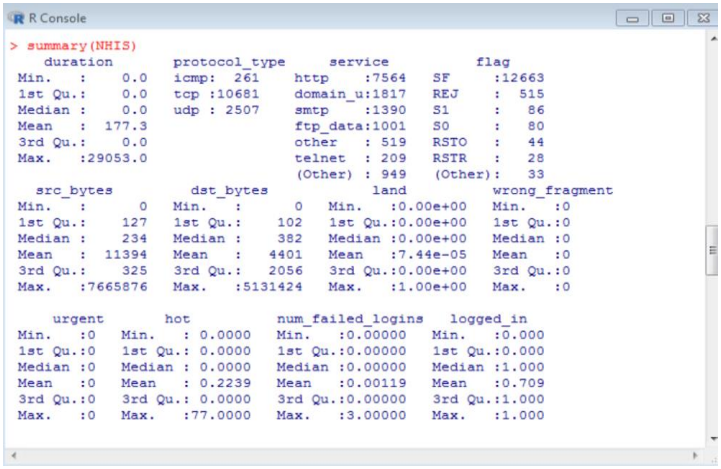


Fig.4.3: Summary of the selected dataset in R-Console

Above diagram 4.3 illustrates summary of the selected dataset. It consists of min, median, mean and maximum of each attribute from the CSV file. For all 42 columns or attributes it provides min, mean and maximum values. By reading the summary we can set the threshold for all 42 attributes range.

Attribute analyser will be run for NSL KDD dataset. It will remove the redundant data from input dataset. Here input dataset will be in the form of ARFF (Attribute Record File Format). Modified dataset will be generated by removing the redundant attributes from the input dataset. To remove the columns we follow ranker method. Columns which possess least rank will be removed by attribute analyser.

Association rules follow the below steps to generate the corresponding rules of the NSL KDD dataset.

Step 1: Output of attribute analyser will be provided as input to generate association rules.

Step 2: If output of attribute analyser consist of numerical values then we have to apply discretize method to convert numerical values to nominal or to corresponding string.

Step 3: We should apply Apriori algorithm for mining to generate frequent used large item sets.

Step 4: Set class index for the Apriori object and build associations for the Apriori object.

Step 5: Print out the details of Apriori algorithm output.

Step 6: Print out the generated sets of large item sets.

Step 7: Print out the extracted rules.

Association Rules Generator Procedure:

Input: Attribute analyser output will be treated as input in association rules generation. Modified NSL KDD dataset will be read as input.

Procedure: Modified NSL KDD dataset will be set as new data source set. Apriori algorithm will be taken into account to generate most frequent used large item set generation. To build the model we have to set the data class index and build the model.

Output: Computed rules, Corresponding confidence, Association rules corresponding output will be explained as below with an example.

Attribute analyser modified output arff file will be treated as input for association rules. Modified file means it consist of less number of columns or attributes. It will reduce the redundant data to relevant data. Unnecessary columns will be deleted from the arff file. Modified file will be given as input to the association rules generator. Apriori algorithm is considered as the most frequent item set Boolean association rules generator.

4.1 PREDICTION MODEL IMPLEMENTATION

Collecting standard dataset which consist of four important classes of assaults or attacks those are DOS, U2R, R2L, Probe and Normal [4]. Prediction model mainly we use two terms that is standard dataset and test dataset. Standard dataset means it consist of 42 attributes, 41 attributes says about properties of assaults whereas 42 attribute say about type of the attack. It is also referred as labeled dataset. Which means it consist of corresponding attribute name of the columns. Test dataset means it consist of 41 attributes. 41 attributes says about properties of attacks. And it doesn't consist last column that is type of attacks. Hence it is called as unlabeled dataset.

Prediction model works as follows:

Step 1: Prediction model reads the standard KDD dataset. Here standard KDD dataset means labeled dataset.

Step 2: Prediction model reads the test KDD dataset. Here test dataset means unlabeled records. Which means it doesn't consist of attribute names.

Step 3: Prediction model will analyze the both files (Standard KDD dataset and test dataset).

Step 4: Prediction model will scan all the attributes from unlabeled dataset.

Step 5: Prediction model will compare the attributes values from test dataset to standard dataset.

Step 6: Comparison in prediction model follows the line by line approach. Means each line in test dataset will be compared with all n number lines in standard dataset.

Step 7: If the test dataset attribute values matching with standard dataset values then corresponding type of attack will be returned by the prediction model.

Step 8: If the test dataset attribute values are not matching with standard dataset then type of attack will be returned as malicious.

Prediction Model Procedure:

Input: Standard KDD dataset (labeled dataset), Test KDD dataset (unlabeled dataset)

Analyze: Prediction model will analyze the standard KDD dataset and test dataset. Prediction model follows the line by line approach. It scans standard dataset and KDD dataset line by line.

Procedure:

```
Procedure_Read_Standard_Dataset(File Standard_Dataset)
        {
int n=Standard_Dataset_Total_Nr_Lines;
File Test_Dataset= File.Read(File Test_Dataset);
for(inti=0;i<n;i++)
{
if (StandardLines.contains(teststr))
{
intind=StandardLines.indexOf(teststr);
System.out.println("The test file index= " +
                                    TestFileIndex + "
belongs to the class
                                of attack "
+StandardClass.get(ind));
System.out.println(teststr + "========>" +
        StandardClass.get(ind));
}
```

}

}

Output: Returns test dataset type of attack.

5. EXPERIMENTAL RESULTS:

KDD dataset covers four important classes of attacks: Denial of Service (DOS), User-to-root (U2R), remote-to-local (R2L) and probing attack. The network simulator NSL KDD dataset is segregated into labelled and unlabelled records. Each marked record comprised of 41 properties and one target esteem. Association rule mining is for the most part used to locate the fascinating rules from a vast database relying on the client characterized support and certainty. For implementation process we have to use the some requirements are Intel Pentium IV Processor or above, 2 GB RAM, 20 GB HDD these are hardware requirements. Operating System is Windows 7 and higher version are used it is software requirements, WEKA API is responsible to fetch the data mining statistics from the corresponding dataset which is NSL KDD dataset. NetBeans IDE 8.0 development environment will be used NetBeans. Language environment is Java Version 8.0.WEKA Jar 3.7.0 Weka jar file contains classes related Weka API files.

5.1 PRE-PROCESSING THE DATA

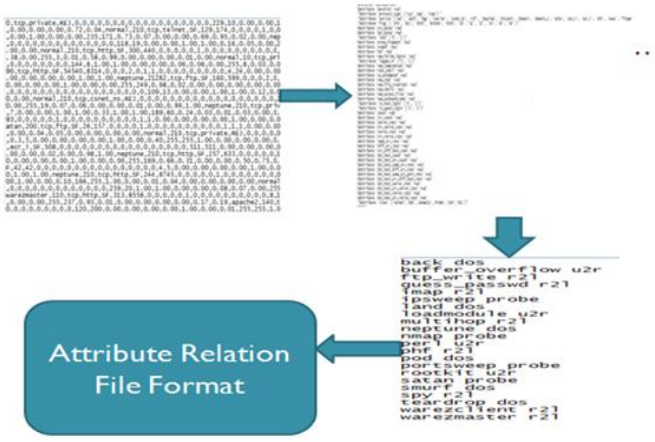


Fig.5.1: Pre-Processing the data

Diagram 5.1 illustrates pre-processing the data will read the complicated and clumsy data which is related to NSL KDD intrusion detection and it will pre-process the data to perform normalization. Here normalization means remove redundancy.

5.2 ARFF TO CSV

Here ARFF means attribute record file format and CSV means comma separated value. In this step program will accept attribute record file format as an input and generate comma separed value file as an output. It converts attribute record file formsat to comma separated value format. Program reads the @relation file name, attributes names,its values and generates corresponding comma separated file as an output.

The below diagram depicts the example file for ARFF format. ARFF to CSV program accepts above file format as an input. ARFF to CSV program reads @relation file name, attribute names and its values.

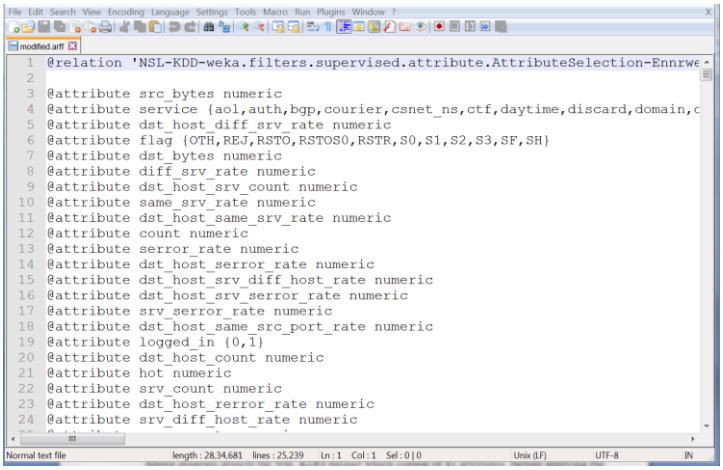


Fig.5.2: Image depicts attribute record file format

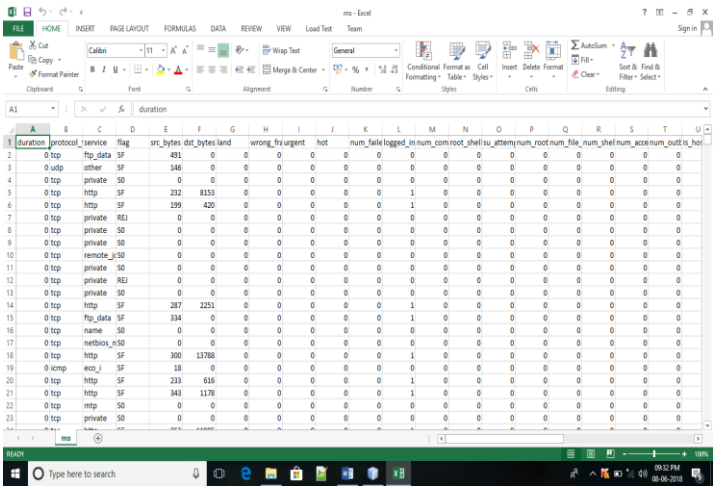


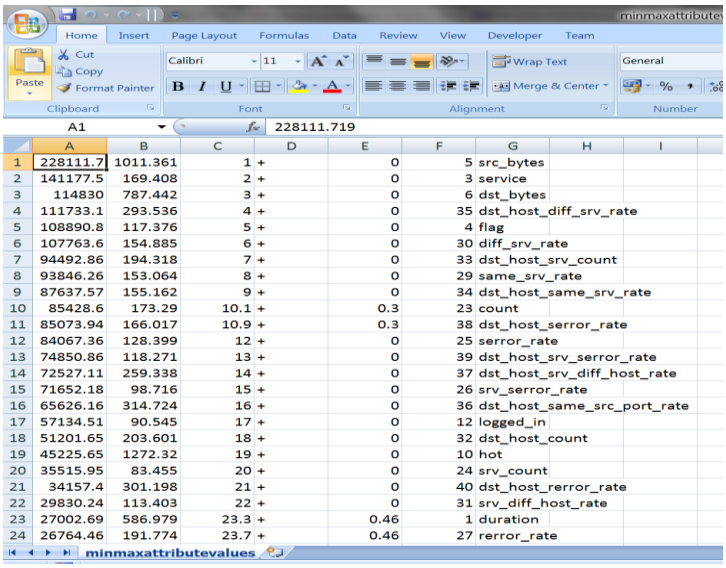
Fig.5.3: Image depicts converted arff file to corresponding csv format

Above diagram depicts that an output of ARFF to CSV program. Process will convert the given input file format from attribute record file format to comma separated value. CSV file has many advantages such its readability, presentation of data in column wise manner etc.

5.3 CSV TO ARFF

Here ARFF means attribute record file format and CSV means comma separated value. In this step program will accept comma separed value file as an input and attribute record file format file as an output. It converts comma separated value format to attribute record file format. Program reads the first row of the csv file and generates @relation file name and its corresponding attributes names. Other lines apart from table header lines indicates data rows, it will be presented in arff file region as @data. It converts the csv file format to corresponding arff format. It is vice versa of the arff to csv.

5.4 FEATURE SELECTION RESULTS

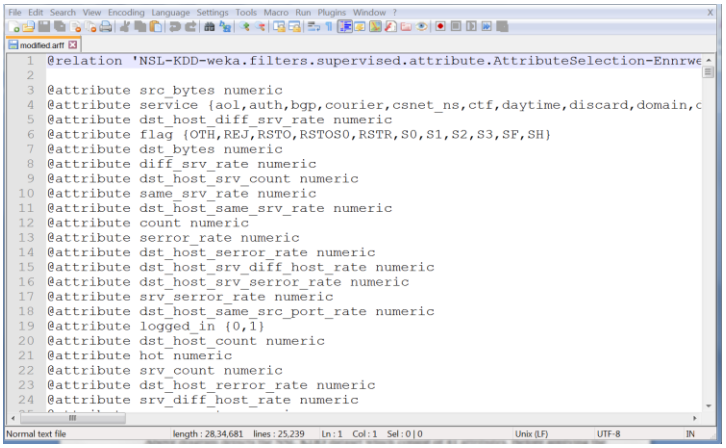


	A	B	C	D	E	F	G	H	I
1	228111.7	1011.361	1	+	0	5	src_bytes		
2	141177.5	169.408	2	+	0	3	service		
3	114830	787.442	3	+	0	6	dst_bytes		
4	111733.1	293.536	4	+	0	35	dst_host_diff_srv_rate		
5	108890.8	117.376	5	+	0	4	flag		
6	107763.6	154.885	6	+	0	30	diff_srv_rate		
7	94492.86	194.318	7	+	0	33	dst_host_srv_count		
8	93846.26	153.064	8	+	0	29	same_srv_rate		
9	87637.57	155.162	9	+	0	34	dst_host_same_srv_rate		
10	85428.6	173.29	10.1	+	0.3	23	count		
11	85073.94	166.017	10.9	+	0.3	38	dst_host_serror_rate		
12	84067.36	128.399	12	+	0	25	serror_rate		
13	74850.86	118.271	13	+	0	39	dst_host_srv_serror_rate		
14	72527.11	259.338	14	+	0	37	dst_host_srv_diff_host_rate		
15	71652.18	98.716	15	+	0	26	srv_serror_rate		
16	65626.16	314.724	16	+	0	36	dst_host_same_src_port_rate		
17	57134.51	90.545	17	+	0	12	logged_in		
18	51201.65	203.601	18	+	0	32	dst_host_count		
19	45225.65	1272.32	19	+	0	10	hot		
20	35515.95	83.455	20	+	0	24	srv_count		
21	34157.4	301.198	21	+	0	40	dst_host_rerror_rate		
22	29830.24	113.403	22	+	0	31	srv_diff_host_rate		
23	27002.69	586.979	23.3	+	0.46	1	duration		
24	26764.46	191.774	23.7	+	0.46	27	rerror_rate		

Fig.5.4: Chi - Square Feature Selection Output Method

There are various feature selection algorithms available. Out of all the algorithms, Chi Square feature selection algorithms has been chosen to generate the rank associated with the attributes. Ranker method generate the corresponding ranks for the attributes. It will identify the most significant attributes and least significant attributes.

5.5 ATTRIBUTE ANALYZER



```
@relation 'NSL-KDD-weka.filters.supervised.attribute.AttributeSelection-Ennrwe

@attribute src_bytes numeric
@attribute service {aol,auth,bgp,courier,csnet_ns,ctf,daytime,discard,domain,c
@attribute dst_host_diff_srv_rate numeric
@attribute flag {OTH,REJ,RSTO,RSTOS0,RSTR,S0,S1,S2,S3,SF,SH}
@attribute dst_bytes numeric
@attribute diff_srv_rate numeric
@attribute dst_host_srv_count numeric
@attribute same_srv_rate numeric
@attribute dst_host_same_srv_rate numeric
@attribute count numeric
@attribute serror_rate numeric
@attribute dst_host_serror_rate numeric
@attribute dst_host_srv_diff_host_rate numeric
@attribute dst_host_srv_serror_rate numeric
@attribute srv_serror_rate numeric
@attribute dst_host_same_src_port_rate numeric
@attribute logged_in {0,1}
@attribute dst_host_count numeric
@attribute hot numeric
@attribute srv_count numeric
@attribute dst_host_rerror_rate numeric
@attribute srv_diff_host_rate numeric
```

Fig.5.5: NSL - KDD dataset with redundant attributes of size 41 attributes

It accepts the NSL KDD dataset as input and apply Chi-Square feature selection algorithm and ranker algorithm to generate corresponding ranks for all the 42 attributes. Feature selection attribute analyzer output or result consist of most significant attributes or features which is possessing high rank. Least ranked attributes have been removed from the NSL KDD dataset. It is most difficult to achieve the same process by manually. If manually done in the sense it may lead to error prone. To avoid these manual work feature selection attribute analyzer has been implemented.

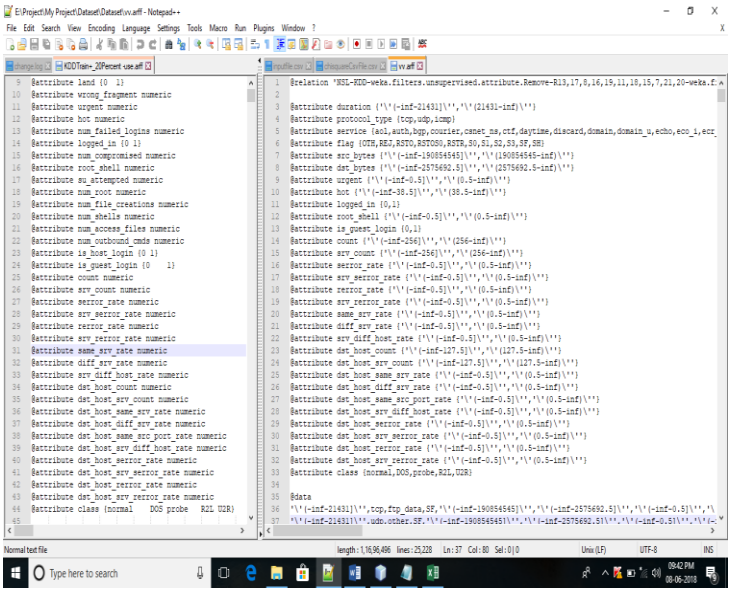


Fig.5.6: Illustrating Removal Features after Applying Attribute Analyzer

5.6 PREDICTION MODEL

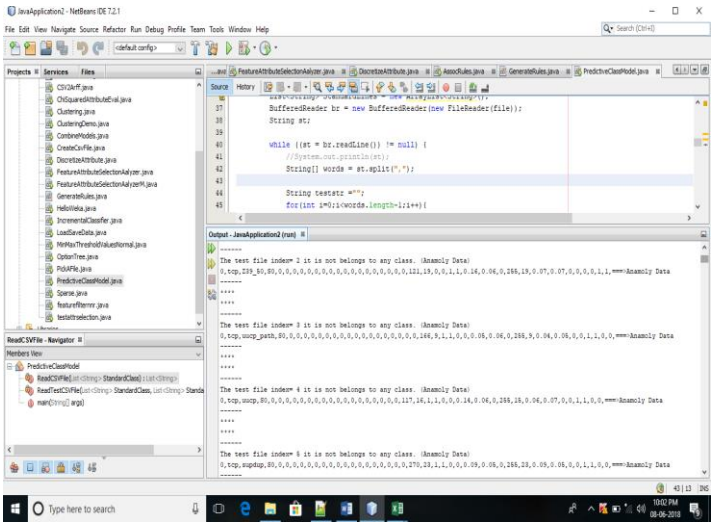


Fig.5.7: Output of prediction model illustrates test file index, class of assault and anomaly data

Prediction model is mainly implemented to perform test analysis. It mainly reads the standard dataset and compares the test dataset. Comparison process follows the line by line approach to find the class of attack or assault it belong to. Prediction model predicts the test model belongs to which class of attack. Prediction model predicts that test dataset belongs to either following of attacks those are normal, u2r, r2l, probe. If class of attack not belongs to any of the attack means it is considered as the anomaly of data.

5.7 ASSOCIATION RULES GENERATOR

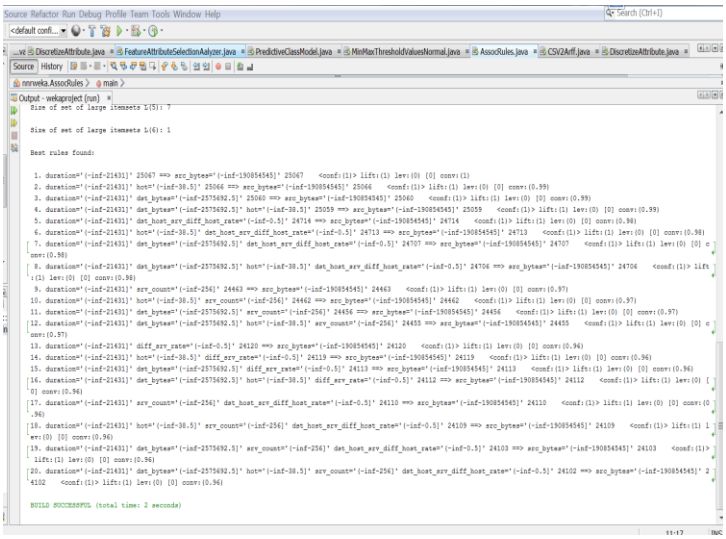


Fig.5.8: Associations rules generator output depicts the minimum support, confidence and generated rules

Above diagram depicts that a rules generator accepts the modified NSL KDD dataset from the feature selection attribute analyzer. Output from the feature attribute selection analyzer is modified NSL KDD dataset it consist of most significant attributes having highest rank. Least attributes have been removed from the dataset. Association rules generator apply the apriori algorithm. It generates the association rules along with the usage of metrics such as minimum support and confidence. The default number of rules generated by the association rules generator program is 10. It is possible to set the number of association rules to be more than 10. The generated association rules follows the descending order of the confidence and minimum supports. Which means rules having highest confidence is located at the top.

6. CONCLUSION

KDD dataset covers four important classes of attacks: Denial of Service (DOS), user-to-root (U2R), remote-to-local (R2L) and probing assault. The selected network simulator dataset that NSL KDD dataset is segregated into labelled and unlabelled records. Each marked record comprised of 41 properties and one target esteem. Association rule mining is for the most part used to locate the fascinating related logical rules from a vast database relying on the client characterized support and certainty. Arff to CSV will convert attribute record file format to comma separated format. Whereas CSV to Arff converts comma separated file format to attribute record file format. To analyse NSL KDD dataset these two file formats are most important file formats. Attribute analyser will analyse the 42 attributes resides in the chosen standard dataset and removes redundant attributes from the chosen standard network dataset. The most important step in attribute analyser is, it uses ranker method to remove the redundant data from the file. The most generalized approach for generating the association rules have been implemented. Number of rules which will be generated based on the Apriori algorithm most frequent item set generation. This generalized approach can be applied to any attribute record file format set. Confidence and support factors are the key factors for association rules generation. Association Rules possessing least support and confidence has been eliminated. Association rules generated from the rules generator is having highest confidence and highest support.

REFERENCES

[1] M.Sulaiman Khan, MaybinMuyeba, FransCoenen, "Weighted Association Rule Mining from Binary and Fuzzy Data".

[2] Tao, F., Murtagh, F., Farid, M, "Weighted Association Rule Mining Using Weighted Support and Significance Framework". In: Proceedings of 9th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 661- 666, Washington DC (2003).

[3] Lu, S., Hu, H., Li, F, "Mining Weighted Association Rules", Intelligent data Analysis Journal, 5(3), 211-255(2001).

[4]. M. Sulaiman Khan, MaybinMuyeba, FransCoenen, David Reid, "Mining Fuzzy Association Rules from CompositeItems".

[5] Michael Steinbach, PangNing Tan, HuiXiong, VipinKumar," Generalizing the Notion of Support".

[6] R. Aggarwal, T. Imielinski, A. Swami, "Mining association rules between sets of items in very large database," Proceedings of ACM SIGMOD conference, 1993.

[7] Y. Wang, Inyoung Kim, G. Mbateng, S.Y Ho, "A latestclass modeling approach to detect network intrusion", Computer Communications, 30, 93-100,2006.

[8] Agarwal, R. Srikant, "Fast Algorithms for MiningAssociation Rules". In: "20 th VLDB Conference,pp.487-499(1994).

[9] Flora S. Tsai, "Network Intrusion Detection UsingAssociation Rules". In International Journal of RecentTrends in Engeerin, vol 2, No 2, November 2009.

[10] D. Newman,"KDD cup 1999 Data", The UCI KDDArchive, Information and Computer Science, University of California, Iravin.

[11] SubhadipSamanta, "Genetic Algorithm: An Approach for optimization (Using MATLAB)", 2014, IJLTET, Vol. 3 Issue 3