IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2017

A Review on High Ranked Features based NIDS

Dipali G. Mogal[1], Sheshnaryan R. Ghungrad[2], Bapusaheb B. Bhusare[3]

PG Student, MSS's College of Engineering & Technology, Jalna (MH) India[1]

Assistant Professor, MSS's College of Engineering & Technology, Jalna (MH) India[2]

Abstract: With the rapid growth in the network traffic day by day the new threats are evolved affecting network security. The benchmark KDD dataset which was generated a decade ago has become outdated as it does not inclusively reflect modern normal behaviors and contemporary synthesized attack activities. In this paper we have used a new UNSW-NB15 data set and compared with the KDD data set and its version. As the network packets consist of a wide variety of features containing some irrelevant and redundant features which reduces the efficiency of detecting attacks, and increase False Alarm Rate (FAR). So to choose the relevant features and remove the redundancy we used central points of attribute values and association rule mining algorithms which help in reducing the processing time by selecting the most frequent values. These algorithms are applied on KDD99 and UNSW-NB15 data sets to get the high rank features.

Keywords: UNSW-NB15 and KDD99 data set, Central point, association rule mining, features Selection.

I. INTRODUCTION

With the tremendous growth of the internet network, a huge increase in the number of attacks has been witnessed. Intrusion detection system is vast area of research in information security. Intrusion detections technique is categories into Signature detection and Anomaly detection. Signature or misuse detection searches for well-known patterns of attacks, and it can only detect an attack if there an accurate matching behavior against an already stored patterns (known as signatures). Anomaly detection establishes a normal activity profile for a system which evolves itself by collecting and understanding the information about the system and determines the behavior of the system based on it. [3] IDS are classified into two types: host-based (HIDS) and network-based (NIDS), HIDS resides on a particular host and looks for attacks on that host while NIDS resides on a separate system monitoring network traffic and searching for attacks. The main issue lies in selecting a good classification technique for making a decision by minimizing the error. Therefore, the key point is to select an effective classification approach to develop an intrusion detection system which has high detection efficiency and low false alarm rate [1].

Description of the KDD99 Data Set

The IST group of Lincoln laboratories at MIT University generated the first version of the KDD99, namely DARAP98 by performing a simulation with normal and attack traffic in a military network (U.S. Air Force LAN) environment. The simulation contained a raw tcpdump files which continued for 9 weeks. By dividing the data set into training set and testing set. The training set consisted of compressed binary tcpdump files around 4 GBs from 7 weeks of the simulated network traffic processed into 5 million connection records. The testing set contained 2 million connection records from two weeks.

From the DARPA98 data set 41 features for each vector with the class label were extracted using Bro-IDS tool, and called KDD99 data set. Which are divided into 3 groups: intrinsic features, content features and traffic features and the attack records were classified into 4 vectors: DoS, U2R, R2L and Probe. The training set of KDD99 included 22 attack types and the testing set contained 15 attack types[2][18]. NSLKDD [19] is the upgraded version of the KDD99 data sets. Table I shows the distribution of attack and normal records in the NSLKDD data set for training and testing sets.

TABLE I: NSLKDD dataset description

Category	**Training set**	**Testing set**
DoS	45,927	7,458
Probe	11,656	2,422
U2R	52	67
R2L	995	2,887
Normal	67,343	9,710
Total Records	125,973	22,544

Due to public availability of all these versions of the data set, which are still applied to evaluate NIDSs, However, many researchers have stated three major disadvantages [4] which can affect the trust of NIDSs evaluation.

1. The attack data packets have a time to live value (TTL) of 126 or 253, while the packets of the network traffic mostly have a TTL of 127 or 254. However, TTL values of 126 and 253 of the attack types do not happen in the training vectors.
2. The probability distribution of both the testing set and training set are different from each other, because of inserting new attack records while testing. Which leads to skew or bias (one side) classification methods towards

DOI10.17148/IJARCCE.2017.6380

some records rather than balance between the attack and normal vectors.
3. The data set is outdated; hence, it does not give a full representation of modern normal and attack activities.

Description of the UNSW-NB15 Data set

The UNSW-NB 15 data set [2][20] was created by utilizing an IXIA Perfect Storm tool to extract a hybrid of modern normal and contemporary attack activities of network traffic. A tcpdump tool was used to capture pcap files of raw network traffic around 100 GB. In order to make analysis of packets easier each pcap file contains 1000 MB. Argus and Bro-IDS techniques were executed in a parallel to generate 49 features with the class label. This data set contains 2, 540,044 records which were divided into a training set and a testing set. The training set involved 175,341 records, while the testing set contained 82,332 records containing attack types and normal records which are reflected in Table II.

TABLE II: UNSW-NB15 dataset description

Category	**Training set**	**Testing set**
Normal	56000	37000
Analysis	2000	677
Backdoor	1746	583
DoS	12264	4089
Exploits	33393	11132
Fuzzers	181846	6062
Reconnaissance	10491	3496
Shell code	1133	378
Generic	40000	18871
Worms	130	44
Total Records	175341	82332

The involved features of the UNSW-NB 15 data set are classified into 6 groups as follows flow, basic, content, time, general purpose and connection, labeled features which are detailed described in Table II.

1. Flow features this group includes the identifier attributes between hosts, such as client-to-serve or server-to-client.
2. Basic features this category involves the attributes that represent protocols connections.
3. Content features this group encapsulates the attributes of TCP/IP; also they contain some attributes of http services.
4. Time features this category contains the attributes of time, for example, arrival time between packets, start/end packet time and round trip time of TCP protocol.
5. Additional generated features this category can be further divided into two groups: (1) General purpose features (from number 36 - 40) which each feature has its own purpose, in order to protect the service of protocols. (2) Connection features (from number 41-47) are built from the flow of 100 record connections based on the sequential order of the last time feature.
6. Labeled features this group represents the label of each record.

The involved attacks of the UNSW-NB15 data set were categorized into 9 types as fuzzers, analysis, backdoor, denial of service, exploit, generic, reconnaissance, shell code, worm.

The UNSW-NB15 data set has several advantages when compared to the KDD data set.

1. It contains real modern normal behaviors and contemporary synthesized attack activities.
2. The probability distribution of the training and testing sets are similar.
3. It involves a set of features from the payload and header of packets to reflect the network packets efficiently.
4. The complexity of evaluating the UNSWNB15 on existing classification systems showed that this data set has complex patterns. This means that the data set can be used to evaluate the existing and novel classification methods in an effective and reliable manner.

Comparison of the KDD99 and UNSW-NB15 data set

Table III shows a comparative analysis among the KDDCUP99 and UNSW-NB15 data sets. The UNSW-NB15 data set has different attack families which ultimately reflect modern low foot print attacks.

TABLE III: Comparison of KDD99 and UNSW-NB15 dataset

#	**Parameters**	**KDD99**	**UNSW- NB15**
1	No. of network	2	3
2	No. of distinct IP address	11	45
3	Simulation	Yes	Yes
4	Feature extraction tools	Bro-IDS tool	Argus, Bro-IDS and new tools
5	Duration of data collected	5weeks	15-16 hours
6	Format of data	3 types (tcdump, BSM, dump files)	pcap files
7	No. of features extracted	42	49
8	Attack families	4	9

II. LITERATURE SURVEY

IDDM (Intrusion Detection using Data Mining Technique) [5] is a real-time NIDS for misuse and anomaly detection. It applies Association rules, Meta rules, and Characteristic rules. It uses data mining technique to produce description

DOI10.17148/IJARCCE.2017.6380

of network data and perform analysis using this information.

MADAM ID (Mining Audit Data for Automated Models for Intrusion Detection) [6] is an offline IDS to produce anomaly and misuse intrusion detection models. This employs Association rules and frequent episodes to replace hand-coded intrusion patterns and profiles with the learned rules.

In order to achieve greater accuracy and decrease false acceptance rate, we need to build NIDS which extract and choose the relevant features from raw network traffic. Feature extraction captures attributes from network packets in which some of the attributes are redundant or irrelevant; which reduces the accuracy of detection and increases the false acceptance rate. Feature selection, removes redundant and noisy attributes from high dimensional data sets and selects a subset of relevant attributes to establish a reliable NIDS model.

As real time intrusion detection is not possible as huge number of data flows upon the Internet. Feature selection techniques can help to reduce the computation and complexities. The procedure for feature selection requires four basic steps: Generating subset, Evaluation of the subset, stopping criterion, Validation [7].

Blum and Langley [8] divide the feature selection techniques into 3 types:

a) Filter technique uses learning algorithm to measure the overall performance of selected features [9].

b) Wrapper it wraps around the learning algorithm by using one pre-specified classifier to observe the features utilizing a search algorithm. The performance evaluation of various feature set is done and the best performance are selected for further. Wrapper method is costly than the filter method [10].

c) The combination of both approaches is the hybrid technique [9] [10]. It can be used to get most effective performance having a specific learning algorithm.

Relevant Feature Selection Model Using Data Mining for Intrusion Detection System [11]: To build a lightweight intrusion detection system, a relevant feature selection model was developed to select the best features set which uses seven different feature evaluation methods to select and rank relevant features. This model has four different stages, Data Pre-Processing, Best Classifier Selection, Feature Reduction, and Best Features Selection. Redundant vectors in the training dataset were eliminated which leads to skew or bias classification of the learning algorithm. From the reduced training dataset only four class-based datasets have been constructed: DOS, PROBE, R2L, U2R each of these four datasets contains the attack type records and the normal class records. The results show that some features have no contribution to detect any intrusion attack type and some features detect all attack types. A set of 11 best features were chosen and tested against the complete features set. With this model a high detection rates was achieved along with speed up in the detection process.

Analysis of Feature Selection Techniques for Network Traffic Dataset [12]:In this paper they analyzes the performance of various classifier and feature selection techniques considering various parameter such as accuracy, number of features, tpr, fprand time taken. This technique reduces features by 82.93 % and gives better accuracy. The accuracy decrease as a result of features reduction i.e.0.91% in Naive bayes, 0.54 % in J48, and 0.56 in PART classifier. For network traffic dataset, CFS subset evaluation technique reduces the features by 75.61 %.

Intrusion Detection System Using Feature Selection and Classification Technique [13]: Optimal Feature Selection (OFS) algorithm and two classification techniques were used for securing the system. Instead of using all the 41 features of the KDD'99 cup data set which takes much time for detecting and classifying the record this developed system selects only the important features that help in reducing the time taken for detecting and classifying the records. The rule based classifier and SVM was implemented which achieves a greater accuracy. The results show that it reduces the FPR and the computation time.

Feature selection and design of intrusion detection system based on k-means and triangle area support vector machine [14]: A hybrid IDS using machine was developed, which was based on triangle area support vector machine (TASVM). In which information gain was calculated for each attack class, the ten most relevant features were selected and the remaining features were discarded. The linearly scaling method was implemented to reprocess data for unifying their ranges after that k-means clustering algorithm was used on the selected subset to produce five clustering centroids. Then two centroids were chosen randomly and one data point to form triangle and calculate these triangle areas which were used in generating a new feature vector for this data. Accordingly, they trained and tested a hybrid IDS with these feature subset in Lib SVM.

Z. Yanyan and Y. Yuan [15] developed a partition-based ARM algorithm. The algorithm was configured to scan the training set twice. In first scan, the data set was partitioned to execute into memory easily, in the second scan, the item sets for the training set were generated. This algorithm has very high complexity.

B. Nath, D. Bhattacharyya, and A. Ghosh [16] stated a review of some existing dimensionality reduction techniques based on ARM methods. Some of these algorithms support single objective and others multi-objective. The results showed that the multi-objective ARM can be used to solve several real datasets. This study is related to our work in customizing ARM as feature selection.

III. PROPOSED FEATURE SELECTION METHOD

Associations rule mining (ARM) is a data mining method to compute the correlation of two or more than two

DOI10.17148/IJARCCE.2017.6380

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2017

attributes in a data set, because it can find the strongest item sets between observations [15]. In this paper, we build a Central Point Algorithm based on ARM as a feature selection method to adopt the relevant features from the UNSW-NB15 and the KDD-99 data sets. The goal of ARM is to generate the strongest item sets among features by computing support and confidence of each rule in a data set [17]. Many researchers have utilized ARM approaches in NIDSs. Following are the steps to obtain high ranked features from the data sets.

- Choose an input data set, for example UNSW-NB15 or KDD99 data set.
- Execute Central Points (CP) Algorithm to compute the central points of attribute values.
- The output of CP is the input to ARM Algorithm to calculate the high ranked attributes.
- Divide the data set into training set and testing set to learn classifiers.

Central Points of Attribute Values

The data set records are divided into equal parts using equation 1, to reduce the processing time. The aim of partitioning is to make easier during the processing and identify statistical characteristics (e.g. mean, median, mode), from different parts of records of the data set to retrieve the relevant attributes.

$$p = \text{No. ofpartions} = \frac{\text{No .ofr ecords}}{\text{No .ofattributes}} \qquad (1)$$

In each part of the data set, we calculate the mode which is the most frequent value of a feature. The attribute values of a network data set can be numeric or categorical. In CP Algorithm, the central points of attribute values (mode) are described. In line 1 and 2, for loops assign all data values. From line 3 to 12, check attribute values either categorical or numerical, and then compute the mode for each data part (p). Lines 13 to 17 repeat the steps until finishing all parts. Line 18 retrieves the mode of all data parts to be input for computing the ARM.

CP Algorithm: Central points of attribute values

```
Input : d data set, p
1. for (i= 1 to length(row)) do
2.   for(j= 1 to length(col)) do
3.     if(d[r][c]!=categorical) then
4.         pre[r][c] =mode($d_{1:p}$ )
5.         if(pre[r][c] !=0 ) then
6.           centers[r][c] = +pre[r][c]
7.         end if
8.     else
9.       pre[r][c]=count($d_{1:p}$ )
10.      if(pre[r][c] > pre[r][c]+1) then
11.        centers[r][c] = +pre[r][c]
12.      end if
13.      p= p -1
14.      row =row-(row/p)
15.     end if
16.   end for
17.  end for
18. return centers
```

Feature Selection through Association Rule Mining (ARM)

To explain the ARM, let r = $\{f_1, f_2, f_3\ldots f_N\}$ be a set of features and D be a data set consisting of T transactions $t_1,t_2,t_3\ldots.t_N$. Each transaction t_j, $1 \leq j \leq N$ is a set of features such that $t_j \subseteq$ r. Association rule (f_1 (*i.e.*, *antecdent*) $\Longrightarrow f_2$ (*i. e.*, *precedent*)) subjects to the constraints of (1) $\exists t_j$, f_1, $f_2 \in t_j$, (2) $f_1 \subseteq$ r, $f_2 \subseteq$ r, and (3) $f_1 \cap f_2 \in \emptyset$.

The ARM subjects to two methods: support and confidence to create rules. Support determines the frequency of row values that denotes the association percentage, as reflected in equation (2). Confidence is the frequency of a precedent if the antecedent has already occurred as in equation (3).

$$\text{sup}(\text{f1} \Rightarrow \text{f2}) = \frac{|\#\text{tj}|\text{f1,f2}\epsilon\text{tj}|}{N} \qquad (2)$$

$$\text{conf}(\text{f1} \Rightarrow \text{f2}) = \frac{|\#\text{tj}|\text{f1f2}\epsilon\text{tj}|}{|\#\text{tj}|\text{f1}\epsilon\text{tj}|} \qquad (3)$$

The ARM finds out all repeated item sets and identifies the strongest rules in the frequent item sets. The strongest ARM in D is realized, if the support of a rule is greater than a user-specified minimum support ($sup \geq \text{min}sup$), and confidence of a rule is greater than minimum confidence thresholds ($conf \geq \text{min}conf$).

It is clear that the Central Points of attribute values of CP Algorithm is considered as an input to ARM Algorithm to reduce the processing time. ARM Algorithm generates the highest ranked attributes based on the ARM. Line 1 is a loop to all CP. From line 2 to 14, check if the rules do not accomplish the ARM constraints, remove it. Otherwise, compute support and confidence. In Line 15, all rules order descending based on the values of support and confidence. From Line 17 to 21, the strongest features are selected based on the number of required features.

ARM Algorithm: Feature selection

```
Input: centers (C), minimum support(min_sup), label (L),minimum confidence (min_conf), No. of required feature (X).
Output: F (feature subset)
1. for (i= 1 to length(C))do
2.        if(C[i] ==C[i+1]) then
3.                count = count+1
4.        else
5.                count=1
6.        end if
7.        filter_C[i] = C-C[i]
8. end for
9. for(j=1 to length(filter_C))do
10.       if(count<=1) then
11.               sup[j] = count[j]/ length(filter_C)
12.               conf[j] = count[j]/length(D[j])
13.       end if
14. end for
```

DOI10.17148/IJARCCE.2017.6380

```
Sort(filter_C, sup, conf)
for(m=1 to X )do
 if(sup>=min_sup&&conf>= min_conf)then
        F+ = extracted_features(r, L)
end if
end for
return F
```

IV. CONCLUSION

In this review paper, we propose a hybrid feature selection technique based on the central points (CP) of attribute values and Association Rule Mining (ARM). The CP technique helps to reduce the processing time by selecting the most frequent values. The ARM is customized to choose the highest ranked features by removing irrelevant or noisy features. This algorithm is executed on the UNSW-NB15 and the KDD99 data set. Ultimately, the proposed feature selection technique has two advantages: reduce processing time and improve the evaluation of decision engines. To discriminate between attack and normal records, clustering and classification techniques of data mining will be used further.

REFERENCES

[1] Moustafa N & Slay J. “The significant features of the UNSW-NB15 and the KDD99 data sets for Network Intrusion Detection Systems”. 4th International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security 2015.

[2] Moustafa N. & Slay J. “UNSW-NB15: A Comprehensive Data set for Network Intrusion Detection”. Paper presented at the Military Communications and Information Systems Conference, Canberra, Australia, 'in press' 2015.

[3] Dartigue, H.Jang and W.Zeng, “A new data-mining based approach for network intrusion detection”, Communication Networks and Services Research Conference. CNSR'09. Seventh Annual. IEEE, 2009, p 372-377.

[4] M. Tavallaee, E. Bagheri, W. Lu, and A.-A. Ghorbani, “A detailed analysis of the kdd cup 99 data set,” in Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defence Applications, 2009.

[5] Tomas Abraham, "IDDM: INTRUSION Detection using Data Mining Techniques”, Technical report DSTO electronics and surveillance research laboratory, Salisbury, Australia, May2001.

[6] Wenke Lee and Salvatore J.Stolfo, "A Framework for constructing features and models for intrusion detection systems”, ACM transactions on Information and system security (TISSEC), vol.3, Issue 4, Nov 2000.

[7] Das M. and Liu H. “Feature Selection for Classification”, Intelligent Data Analysis, 1(3), pp 131–56, 1997

[8] George H John, Ron Kohavi and Karl PEger, “Irrelevant Features and the Subset Selection Problem”, Proc. of the 11th International Conference. On Machine Learning, Morgan Kaufmann Publishers, pp 121-129, 1994.

[9] Liu, H. and Yu, L., “Towards integrating feature selection algorithms for classification and clustering”, IEEE Transactions on Knowledge and Data Engineering, 17(4), pp 491-502.

[10] R. Kohavi and G.H. John. “Wrappers for Feature Subset Selection”, Artificial Intelligence.97 (1-2), pp 273-324.

[11] A I. Madbouly, Amr M. Gody,Tamer and M. Barakat, “Relevant Feature Selection Model Using Data Mining for Intrusion Detection System”, IJETT – Volume 9 Number 10 - Mar 2014.

[12] Raman Singh, Harish Kumar and R K Singla “Analysis of Feature Selection Techniques for Network Traffic Dataset”, International Conference on Machine Intelligence Research and Advancement, IEEE, pp. 21-23,Dec. 2013.

[13] S Balakrishnan, Venkatalakshmi K, and Kannan A “Intrusion Detection System Using Feature Selection and Classification Technique”. IJCSA Volume 3 Issue 4, November 2014.

[14] Pingj Tang, R Jiang and Mingwei Zhao “Feature selection and design of intrusion detection system based on k-means and triangle area support vector machine”. Second International Conference future network on, pp 144 – 148, IEEE 2010.

[15] Z. Yanyan and Y. Yuan, “Study of database intrusion detection based on improved association rule algorithm,” in 3rd IEEE International Conference, CS and IT, vol. 4 IEEE, 2010, pp. 673–676.

[16] B. Nath, D. Bhattacharyya, and A. Ghosh, “Dimensionality reduction for association rule mining,” International Journal of Intelligent Information Processing, vol. 2, no. 1, 2011.

[17] Agrawal, R., Imieliński, T., & Swami, A. “Mining association rules between sets of items in large databases”. Paper presented at the ACM SIGMOD Record 1993.

[18] Kddcup1999, April 2015. Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[19] NSLKDD. Available on: http://nsl.cs.unb.ca/NSLKDD/2009.

[20] UNSW-NB15,May2015.Available: http://www.cybersecurity.unsw.adfa.edu.au/ADFA%20NB15%20Datasets/