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 4, April 2017

NIDS using Machine Learning Classifiers on UNSW-NB15 and KDDCUP99 Datasets

Dipali Gangadhar Mogal[1], Sheshnarayan R. Ghungrad[2], Bapusaheb B. Bhusare[3]

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

Assistant Professor, MSS's College of Engineering & Technology, Jalna (MH) India[2,3]

Abstract: The benchmark KDD dataset for intrusion detection system generated a decade ago has become outdated as it does not reflect modern normal behaviors and contemporary synthesized attack activities. In this paper we have used a new UNSW-NB15 data set for NIDS. Pre-processing on this datasets is done using Central Points of attribute values with apriori algorithm to select high ranked feature and remove irrelevant features which causes high false alarm rate. The evaluation of the dataset is performed using machine learning classifiers algorithm: Naïve Bayes and Logistic Regression. The results show that the decrease in false alarm rate and detection accuracy is improved even after reducing the dataset by eliminating the features and further more reduce in the processing time.

Keywords: Central Point (CP) of attribute values, Apriori, Naïve Bayes (NB), and Logistic Regression (LR).

I. INTRODUCTION

As networks are considered as the engine of communications, attackers endeavor to penetrate them to steal valuable information or disrupt computer resources. A Network Intrusion Detection System (NIDS) is technique to protect computer resources against malicious activities [1]. 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 construction of NIDS needs to extract and choose the relevance features of raw network traffic to reduce the processing time .Feature extraction captures attributes from network packets. Some of these attributes are redundant or irrelevant; thus reducing the accuracy of detection. Feature selection, on the other hand, removes redundant and noisy attributes from high dimensional data sets and selects a subset of relevant attributes to establish a reliable NIDS model [8].
An association rule mining (ARM) technique generates feature correlations from a data set, as it can find out related isomorphism between data set observations [9]. The association rule mechanism is applied to extract suitable behaviors from user activities. Associations rule mining (ARM) is a data mining method to compute the correlation of two or more than two attributes in a data set, because it can find the strongest item sets between observations [13]. 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 [14][16].

II. PROPOSED SYSTEM ARCHITECTURE

In this section, we describe the proposed system which is shown in Figure 1. The architecture consists of three stages: Pre-Processing, Decision Engine Techniques, Results Analysis.

The proposed system includes the following procedure:

- Choose an input data set, for example UNSW-NB15 [17] or KDDCUP99 [18] data set.
- Execute Central Points (CP) Algorithm [2] to compute the central points of attribute values and apriori algorithm for feature selection using association rule mining (ARM) [4][15].
- Divide the dataset into two parts training and testing sets.
- The original datasets and reduced datasets are evaluated using machine learning classifiers algorithms: Naïve Bayes and Logistic regression.
- Finally result analysis is performed in terms of detection accuracy with respect to processing time.

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 4, April 2017

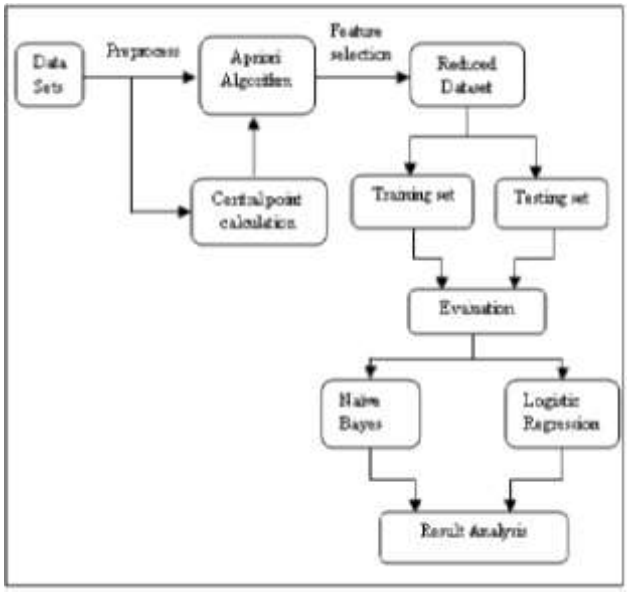

Figure1: Proposed System Architecture

Stage I: Pre Processing

In this stage we reduce the dataset size by eliminating the attributes of a given data set by using central points of attribute values and Apriori algorithm to decrease the FAR. This algorithm is designed to be implemented in a short processing time, due to its dependency on the central points of feature values with partitioning data records into equal parts. This algorithm is applied on the UNSW-NB15 [3] and the KDDCUP99 data sets to adopt the highest ranked features.Pre-processing on this datasets is done using Central Points of attribute values with Apriori algorithm to select high ranked feature and remove irrelevant features which causes high false alarm rate.

Stage II: Decision Engine Techniques

In the Decision engine, we used Naïve Bayes (NB) [10] techniques & Logistic Regression (LR) [11]. Naïve Bayes is based on the Bayesian theorem it is particularly fit when the dimensionality of the inputs is more. Estimation of Parameter for naive Bayes models uses the method of maximum likelihood. It can also performs better in many complex real world situations .NB requires small amount of training data. It is a conditional probability model which creates the classification of the two classes: normal (0) or attack (1). It is computed using the maximum a posterior, as denoted as:

$$P(C|\ I) = \text{argmax}_{\omega\epsilon\{1,2,..N\}} P(C_\omega) \prod_j^N = 1P(I_j|C_\omega)$$

Such that denotes the label of class, I is the observation of each class, ω is the class number, P (C|I) refers to the probability of the class given a specified observation$P(C_\omega)\prod_j^N = 1P(I_j|C_\omega)$ and is multiplication of all the probabilities of the instances conditionally to their classes to achieve the maximum outcome.

Logistic Regression is a classification method. It returns the probability that the binary dependent variable may be predicted from the independent variables. Maximum Likelihood Estimation is a statistical method which estimates the coefficients of the model. It is used in various fields, including machine learning, most medical fieldsetc.

$$P(D=1|X_1; X_2; \ldots; X_k) = \frac{1}{(1+e^{-(\alpha+\sum\beta iXi)})}$$

The model is defined as logistic if the expression for the probability, given the Xs, is 1 over 1 plus e to minus the quantity α plus the sum from i equals 1 to k of β_i times X_i. The terms α and β_i in this model represent unknown parameters that we need to estimate based on data obtained on the Xs and on D for a group of subjects.The LR algorithm constructs the correlation between a dependent variable (L) and independent variables (F). It utilizes the maximum likelihood function to estimate the regression parameters [11].

Stage III: Result Analysis

The result analysis is done for the Pre-processing and the Decision Engine Techniques.

In the first stage we pre-process the KDDCUP99 and UNSW-NB15 datasets using apriori and central point with apriori algorithms. We considered the processing time required for both the methods apriori and central point with apriori and

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 4, April 2017

evaluate that CP with apriori requires less processing time as compared to apriori algorithm which is shown in the Table I.

Table I: Pre-processing time taken

Dataset	Apriori Time	CP with Apriori Time
KDDcup99	175.59 sec	23.4 sec
UNSW-NB15	150.11 sec	4.13 sec

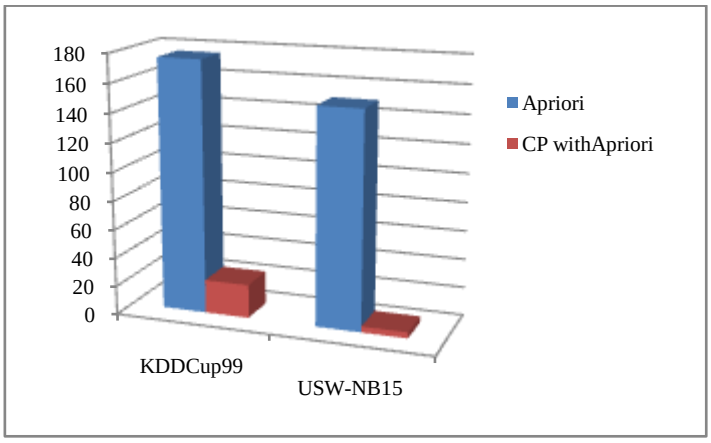

Figure 2: Pre-processing time taken

In the decision engine techniques we evaluate the machine learning algorithms: NB and LR on KDDCUP99 and UNSW-NB15 datasets using a performance metric as detection accuracy and execution time.The performance analysis is shown in Table II& III and the performance comparison plot are shown in Figure 3 & 4.

Table II: Performance analysis of NB & LR for KDDCUP99 dataset

Data set	Classifiers	NB		LR	
		Acc	Time	Acc	Time
KDD cup99	KDDcup99	99.82	5.72	99.93	109.66
	Reduced with Apriori	99.82	5.75	99.93	113.42
	Reduced with CP + Apriori	99.82	4.72	99.94	65.54

Table III: Performance analysis of NB & LR for UNSW-NB15 dataset

Data set	Classifiers	NB		LR	
		Acc	Time	Acc	Time
UNSW-NB15	UNSW-NB15	99.96	0.96	99.89	62.22
	Reduced with Apriori	96.03	1.07	99.93	52.72
	Reduced with CP + Apriori	98.96	0.76	99.96	44.47

The detection accuracy for KDDCUP99 dataset using NB and LR is maintained even after the data loss in the dataset using the reduction techniques apriori and CP with apriori. The execution time taken by NB for KDDCUP99 dataset and reduced CP with apriori dataset varies slightly. The execution time taken by LR for reduced KDDCUP99 CP with apriori dataset is 65.54 seconds which is less as compared to 113.42 seconds for reduced KDDCUP99 with apriori dataset and 109.66 seconds for original KDDCUP99 dataset.

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 4, April 2017

The detection accuracy of NB for UNSW-NB15 dataset differs slightly for reduced UNSW-NB15 with apriori dataset but maintained for reduced UNSW-NB15 CP with apriori dataset. The execution time taken by NB for UNSW-NB15 dataset is also reduced.

The detection accuracy of LR for UNSW-NB15 dataset is maintained and improved for reduced UNSW-NB15 dataset. The execution time taken by LR for UNSW-NB15 dataset is also reduced.

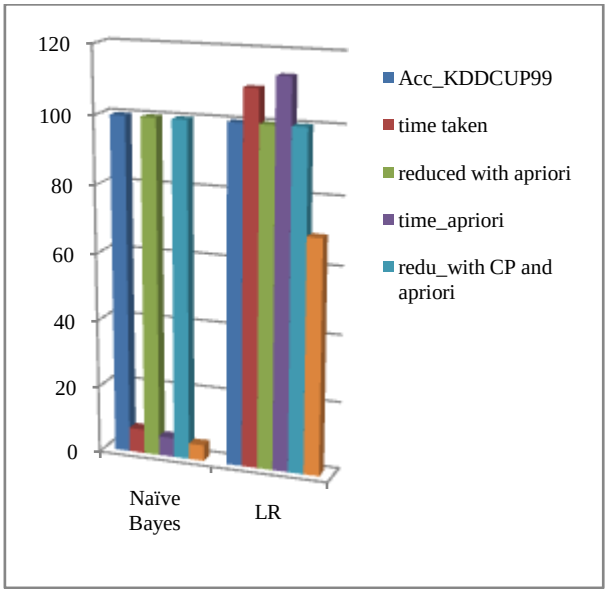

Figure 3: Performance comparison plot of NB & LR for KDDCUP99 dataset

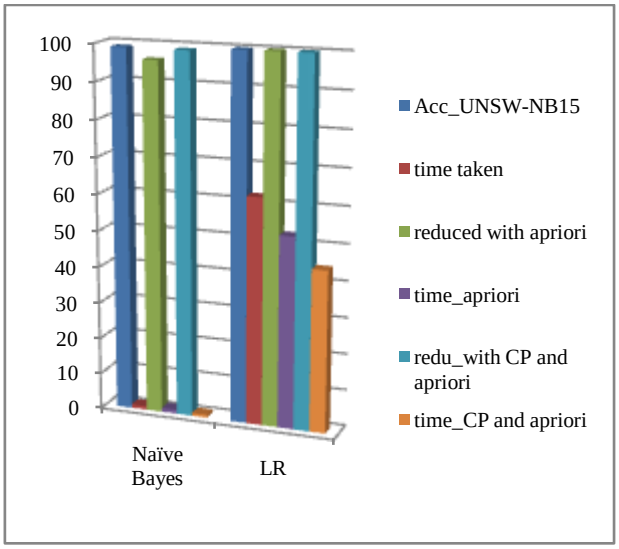

Figure 4: Performance comparison plot of NB & LR forUNSW-NB15 dataset

III. CONCLUSION AND FUTURE SCOPE

Proposed a Network Intrusion Detection System (NIDS) based on central points (CP) of attribute values and apriori algorithm of Association Rule Mining (ARM) for pre-processing. The CP helps to improve the apriori algorithm by reducing the processing time to choose the high ranked features by eliminating the irrelevant features. These algorithms are executed on the KDDCUP99 and UNSW-NB15 datasets.To discriminate betweenattack and normal records, Naïve Bayes and Logistic Regression are used. The experimental results showthat, the pre-processing has reduced the processing time and improved the evaluation of the decision engine. The proposed system is able improve the detection accuracy and decrease the false alarm rate by reducing the processing time. In the future, using enhance algorithm for reducing redundancy in the dataset will help in reducing the processing time.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 4, April 2017

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. "A hybrid feature selection for network intrusiondetection systems: Central point's" 2015
[3] 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.
[4] Dipali G. Mogal, S. R. Ghungrad."A Review on High Ranked Features based NIDS" IJARCCE Vol. 6 Issue 3 March 2017.pp 349-353.
[5] 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.
[6] 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.
[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] Panda, M., & Patra, M. R..(2007)Network intrusion detection using naive bayes. International journal ofcomputer science and network security, 7(12), 258-263.
[11] Kleinbaum, D. G., & Klein, M. (2010). Analysis of Matched Data Using Logistic Regression: Springer.
[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] 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.
[14] B. Nath, D. Bhattacharyya, and A. Ghosh, "Dimensionality reduction for association rule mining," International Journal of Intelligent Information Processing, vol. 2, no. 1, 2011.
[15] Agrawal, R., Imieliński, T., & Swami, A. "Mining association rules between sets of items in large databases". Paper presented at the ACM SIGMOD Record 1993.
[16] Zhang, C., & Zhang, S. (2002). Association rule mining: models and algorithms: Springer-Verlag.
[17] Kddcup1999, April 2015. Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
[18] UNSW-NB15, May2015.Available: http://www.cybersecurity.unsw.adfa.edu.au/ADFA%20NB15%20Datasets/