ISSN (e): 2250 – 3005 || Volume, 08 || Issue, 9|| Sepetember – 2018 ||
International Journal of Computational Engineering Research (IJCER)

Analysis Of Data Normalization With Multilevel Classifiers For Intrusion Detection

Dharmendra Kumar, Ravi Singh Pippal
[1]Research Scholar, [2]Professor,Department of Computer Science and Engineering,Vedica Institute of Technology, RKDF University, Bhopal, India.
Corresponding Author: Ravi Singh Pippal

ABSTRACT:*As network applications grow rapidly, network security mechanisms require more attention to improve speed and accuracy. The evolving nature of new types of intrusion poses a serious threat to network security: although many network security tools have been developed, the rapid growth of intrusive activities is still a serious problem. Intrusion detection systems (IDS) are used to detect intrusive network activity. Machine learning and data mining techniques have been widely used in recent years to improve intrusion detection in networks. In this research work the proposed model for intrusion detection is based on normalized features and multilevel classifier. The work is performed in divided into four stages. In the first stage data is normalized as well as in second stage multilevel classifiers are used. Mean Range, Statistical and frequency normalization techniques are analyzed with Multi-SVM, Multilevel SVM_ELM and MultiSVM_ELM classifiers. In result analysis the detection rate and false alarm rate is evaluated and it is concluded that with mean range normalization outperforms best as compared to other normalization techniques.*
KEYWORDS:*Intrusion Detection, Multilevel Classifier, Machine Learning, Classification, Detection Rate, FAR*

Date of Submission: 03-11-2018 Date of acceptance: 17-11-2018

I. INTRODUCTION

Intrusion Detection Systems (IDS) are security tools that detect attacks on a network or host computer. An IDS is based on the host or network. A host-based IDS detects attacks on a host computer, while a network-based IDS, also known as a network intrusion detection system (NIDS), detects intruders in a network by analyzing network traffic and typically installed in the gateway network or server, host-based intrusion detection systems can be divided into four types: (a) file system monitor, (b) log file scanners, (c) link analyzers, (d) IDs based on kernels [1, 2]. Based on the data analysis technique, there are two broad categories of IDS titles, which are mainly based on signatures and anomalies. A signature-based system detects attacks by analyzing network data for attack signatures stored in its database. This type of IDS detects previously known attacks whose signatures are stored in their database. On the other hand, an IDS anomaly appearance - deviations from the traditional behavior of the subjects. The anomaly-based systems are able to detect new attacks [3-7].

Here are some very common methods used by intruders to take control of computers: Trojan horses, backdoors, denial of service, viruses transmitted via email, package tracking, identity theft and so on. a network package has 42 features and four simulated attacks like [8-12]:

DoS (Denial of Service): excessive use of bandwidth or unavailability of system resources resulting from denial of service attacks. Examples: tear and smurf.

User root (U2R) Attack: Initially, access to malicious users on a normal user account, obtained after logging in to root exploiting system vulnerabilities. Examples: Perl, Load Module and Eject attacks.

Probe attack: access to all network information before launching an attack. Examples: ipsweep, nmap attacks.

Root to Local Attack (R2L): exploiting some of the vulnerabilities of the network, the attacker gets local access by sending packets to a remote machine.

Machine learning techniques can be effective in detecting intruders. Many intrusion detection systems are based on machine learning techniques [13,14,15]. Learning algorithms are created in the offline data set or in real data from academic or organizational networks. To make an IDS model faster with more accurate detection rates, selection of important features from the input dataset is highly essential. Feature selection in learning process while design the model leads to reduction in computational cost, over fitting, model size and improve accuracy. Some existed work in feature selection for intrusion detection. Intrusion detection datasets contain huge amount of observations or records with higher dimensional data. Most of the machine learning algorithm are not

perform well in case of unscale data. In KDD Cup 99, the attribute like duration, source byte, dst byte contains high variations as a result the performance of the algorithm degrades [16].
Attribute normalization is very important for many anomaly detection tasks but it is often ignored. To the best of our knowledge, this is the first study that evaluates the impact of attribute normalization on the classification performance. There are generally four steps for intrusion detection:

a) Attribute Normalization
b) Feature Selection
c) Model Building
d) Intrusion detection

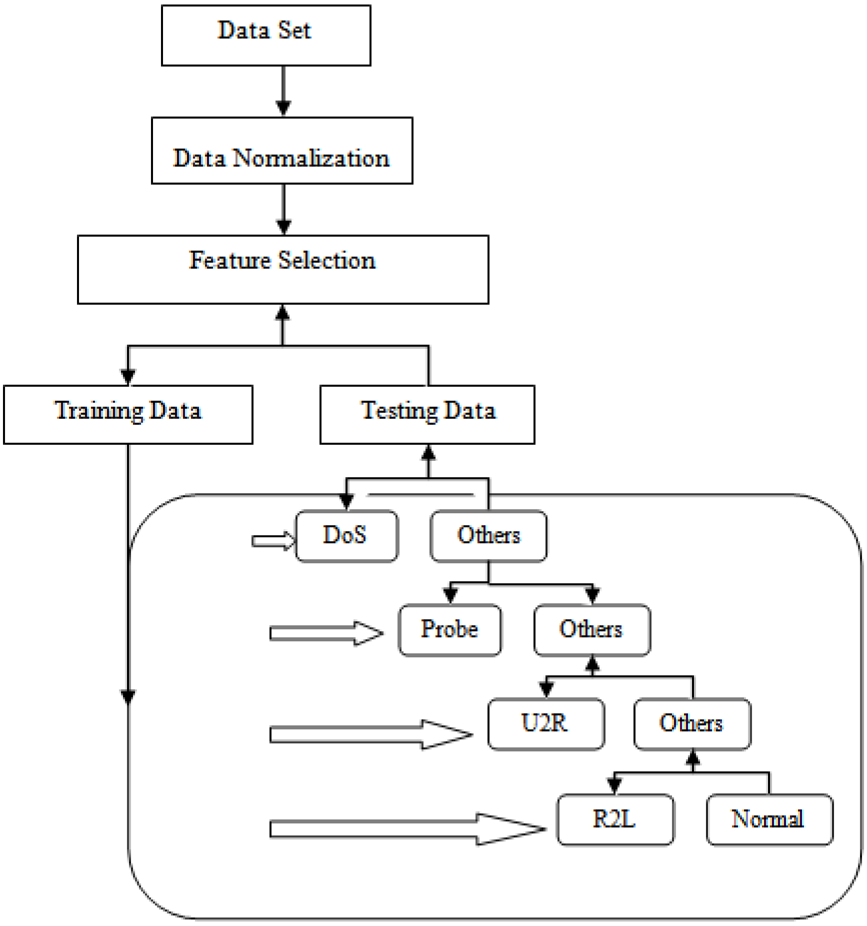

Figure 1: Flow Diagram of Intrusion Detection System

II. INTRUSION DETECTION

IDS learns models from training data so that only the known attack can be detected, new attacks cannot be identified. This section describes the proposed hybrid model for intrusion detection. TheKDD-99 dataset is used as a benchmark to evaluate the performance of the proposed model [17]. The algorithm flow of the proposed method is described as follows:
Following steps will be used to build the proposed model for intrusion detection:
Step 1: Convert the symbolic attributes protocol, service, and flag to numerical.
Step 2: Normalize data to [0,1].
Step 3: Separate the instances of dataset into two categories: Normal, DOS, R2L, U2R and Probe.
Step 4: The data set is divided as training data and testing data.
Step 5: Train classifier with these new training datasets.
Step 6: Test model with dataset.
Step 7: Finally computing and comparing Detection rate and False alarm rate for classifiers.
The algorithm flow diagram of intrusion detection model is illustrated in figure 1. The proposed framework consists of three phases i.e. Attribute normalization, feature reduction and Intrusion Detection Phase. Below each stage is described individually in details.

A. **Attribute Normalization**

This paper is focuses on attribute normalization which is further required for intrusion detection which is illustrated in figure 2. Besides the original attributes, in this paper, data attributes are normalized for further processing.

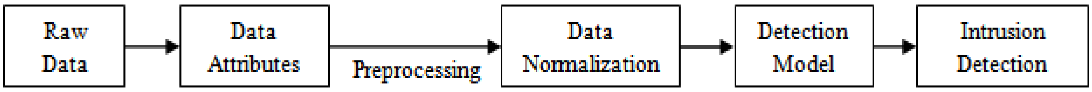

Figure 2: Steps of Intrusion Detection

In this paper, performance of three data normalization technique are analyzed in intrusion detection which are discussed below:

i. Mean Range Normalization

If we know the maximum and minimum value of a given attribute, it is easy to transform the attribute into a range of value [0,1] by:

$$\mathrm{Data}_i = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} \tag{i}$$

Where, x_i = original data of the feature or attribute
$\min(x_i)$= minimum value of data attribute
$\max(x_i)$= maximum value of data attribute
Normally xi is set to zero if the maximum is equal to the minimum.

ii. Statistical Normalization

The purpose of statistical normalization is to convert data derived from any Normal distribution into standard Normal distribution with mean zero and unit variance. The statistical normalization is defined as:

$$\mathrm{Data}_i = \frac{x_i - \mu}{\sigma} \tag{ii}$$

Where, x_i = original data of the feature or attribute
μ= mean of data value

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{iii}$$

σ=standard deviation

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)} \tag{iv}$$

However, using statistical normalization, the data set should follow a Normal distribution, that is, the number of sample n should be large according to central limit theorem. The statistical normalization does not scale the value of the attribute into [0,1].

iii. Frequency Normalization

Frequency normalization is to normalize an attribute by considering the proportion of a value to the summed value of the attribute. It is defined as:

$$x_i = \frac{x_i}{\sum_i x_i} \tag{v}$$

Frequency normalization also scales an attribute into [0,1].

B. **Feature Selection**

Once pre-processing is applied, the pre-processing Module creates the Feature Vector matrix of dataset that represents in which each row i represents the instances and j represents the packet attributes [18].

The aim Feature selection phase is to further select only those features from the database which are relevant for proper classification of the dataset and consequently reduces the feature space dimension so as to reduce complexity by removing irrelevant data. In this research work for feature selection Correlation Analysis is performed using Pearson, Spearman and Kendall coefficients which are explained below.

Pearson Correlation Analysis

Pearson correlation coefficient ρ is calculated by the formula as given below:

$$\rho = \frac{E[AD] - E[A]E[D]}{\sqrt{E[A^2] - (E[A])^2}\sqrt{E[D^2] - (E[D])^2}} \quad \text{(vi)}$$

Where:
A stands for the Attribute Vector
D stands for the Decision Vector
E[A] stands for the sum of the elements in A
Spearman Correlation Analysis
Spearman Correlation coefficient σ is calculated by the formula mentioned below:

$$\sigma = 1-(6\Sigma d_i^2)/n(n^2 - 1)$$

(vii)
Where,
$d_i$stands for the difference between the ranks of variables P and Q
n stands for the sample size
Kendall Correlation Analysis
Kendall Correlation coefficient τ is calculated by the formula as given below:
$\tau = (n_c - n_d)/(1/2n(n - 1))$ (viii)
Where,
$d_i$stands for the difference between the ranks of variables P and Q
n stands for the sample size
After doing Pearson Correlation, Spearman Correlation and Kendall-rank Correlation, we get a list of attributes that satisfy the respective correlation criteria. After obtaining the three individual results which reduces the number of features using Algorithm discussed below:
Attribute Selection after Correlation
procedure ATTRIBUTESELCTION(Dataset)
rows ← nrows(Dataset)
cols ← ncols(Dataset)
pearsonVector ← pearson(Dataset)
spearmanVector ← spearman(Dataset)
kendallVector ← kendall(Dataset)
for each i in 1:cols do
ifpearsonVector[i]>0 AND spearmanVector[i]>0 AND kendallVector[i]>0 then
Selection ← true
else
Selection ← false
end if
end for
return dataset[,Selection]
end procedure

C. Intrusion Detection Phase

For intrusion detection or classification dataset multilevel classifier is used. In this research work three multilevel classifier performance is analyzed i.e. Multilevel SVM, SVM-ELM-SVM-SVM classifier and SVM-ELM-SVM-ELM classifier are used. In figure 1, multilevel classifier is illustrated that consists of four levels.
For Multilevel SVM classifier at all level classifier support vector machine (SVM) algorithm is applied i.e. DOS, Probe, U2R, R2L and Normal are classified using SVM algorithm (as shown in figure 3).

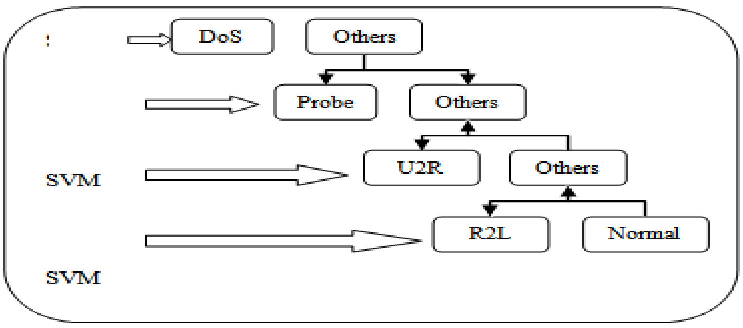

Figure 3: Multilevel SVM Classifier

Whereas in Multilevel SVM_ELM at four levels of classifier support vector machine (SVM) and extreme learning machine (ELM) is used alternately i.e. DOS and U2R are classified using SVM as well as Probe and R2L is classified using ELM(as illustrated in figure 4).

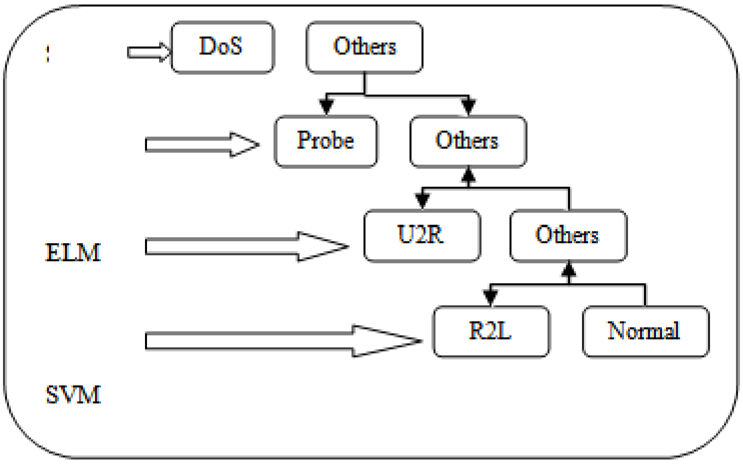

Figure 4: Multilevel SVM_ELM Classifier

Whereas in MultiSVM_ELM classifierat four levels of classifier support vector machine (SVM) and extreme learning machine (ELM) is used i.e. DOS, U2R and R2L are classified using SVM as well as Probe is classified using ELM (as illustrated in figure 5).

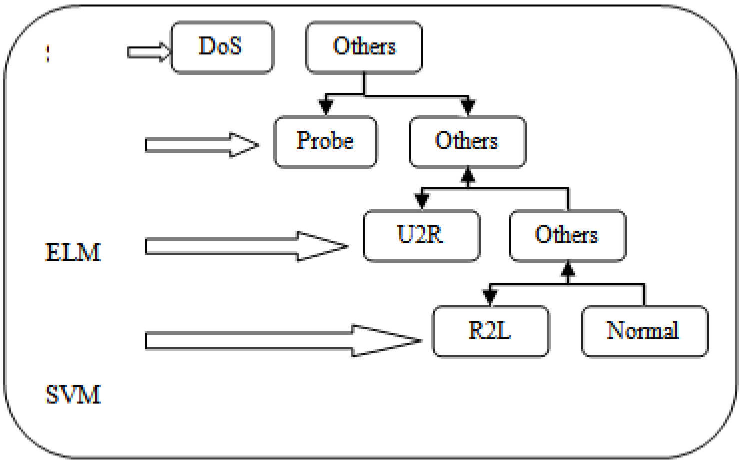

Figure 5: MultiSVM_ELM Classifier

III. SIMULATION RESULTS

A. Dataset Description

The KDD Cup 1999 dataset was used for the Third International Knowledge Discovery and Data Mining Tools Competition. Each connection instance is described by 41 attributes (38 continuous or discrete numerical attributes and 3 symbolic attributes). Each instance is labeled as either normal or a specific type of attack. These attacks fall under one of the four categories: DoS, Probe, U2R, and R2L. KDD Cup 1999 provided both the training and testing datasets, which are called 10% KDD and corrected dataset, respectively. The 10% KDD dataset contains 22 types of attacks, whereas the corrected dataset features the same 22 types of attacks, along with 17 additional attack types [19].

B. Performance Parameters

To evaluate the proposed algorithm, it is concentrated on three indications of performance: detection rate and False Alarm Rate [20].

If one sample is an anomaly and the predicted label also stands anomaly, then it is called as true positive (TP).

If one sample is an anomaly, but the predicted label stands normal, then it is called as false negative (FN).

If one sample is a normal and the predicted label also stands normal, then it is true negative (TN).
If one sample is normal, but the predicted label stands anomaly, then it is termed as false positive (FP).
TP stands the number of true positive samples, FN stands the number of false negative samples, FP stands the number of false positive samples, and TN stands the number of true negatives.
The accuracy and detection rate are calculated as:

Detection Rate =TP/(TP+FN)*100 (ix)

False Negative Rate (FNR) = FN/(FN+TP) *100 (x)

False Positive Rate (FPR) =FP/(FP+TN) *100 (xi)

False Alarm Rate (FAR)= (FPR+FNR)/2 (xii)

C. Result Analysis

For performance evaluation, multilevel hybrid classifiers are used. The performance evaluation is performed using normalized feature based multilevel classifiers. By applying normalization techniques over KDD-99 dataset it has been observed that best result is obtained by using Multilevel classifiers.

Table I: Performance Analysis of Normalization Techniques

Parameter	**SVM_ELM_SVM_SVM**			**SVM_ELM_SVM_ELM**			**MULTI_SVM**		
	Mean Range Normalization	Statistical Normalization	Frequency Normalization	Mean Range Normalization	Statistical Normalization	Frequency Normalization	Mean Range Normalization	Statistical Normalization	Frequency Normalization
Detection Rate	98.932	97.8513	82.0346	**99.5545**	99.553	60.8309	99.1323	99.0224	76.1827
FPR	0.185	0.0139	0.0345	**0.1652**	0.1645	0.0035	0.0122	0.1488	0.0582
FNR	1.068	2.1487	17.9654	**0.4455**	0.447	39.1691	0.8677	0.9776	23.8173
FAR	0.6265	1.0813	8.9999	**0.3054**	0.3058	19.5863	0.44	0.5632	11.9377

Table I shows the performance evaluation of multilevel classification algorithms over dataset. From the result analysis it has been analyzed that detection rate and false alarm rate of Multilevel SVM_ELM classification achieved best result with mean range normalization.

IV. CONCLUSION

This research work proposes a multi-level hybrid classification intrusion detection system. The normalization technique is used to pre-process training dataset and provides high accuracy and detection rate as compared to existing work. The performance measures illustrate than multilevel classification algorithms outperform better with mean range normalization. From the result analysis it has been analyzed that detection rate and false alarm rate of Multilevel According to simulation on KDD-99 dataset, the proposed algorithm achieved approx. 99% detection rate as well as 0.3% False alarm rate.

REFERENCES

[1]. Garcia-Teodoro, P., “Anomaly-based network intrusion detection: techniques”, systems and challenges. Comput. Security vol. 28. issue, pp. 18–28, 2009.

[2]. Sufyan T Faraj Al-Janabi, HadeelAmjedSaeed, “A neural network-based anomaly intrusion detection system”, IEEE, 2011.

[3]. J. Ryan, M. Lin, and R. Miikkulainen, “Intrusion Detection with Neural Networks,” Conference in Neural Information Processing Systems, 943–949.

[4]. A. K. Ghosh and A. Schwartzbard, “A Study in Using Neural Networks for Anomaly and Misuse Detection,” Conference on USENIX Security Symposium, Volume 8, pp. 12–12, 1999.

[5]. P. L. Nur, A. N. Zincir-heywood, and M. I. Heywood, “Host-Based Intrusion Detection Using Self-Organizing Maps,” in Proceedings of the IEEE International Joint Conference on Neural Networks, pp. 1714–1719, 2002.

[6]. Sharma, R.K., Kalita, H.K., Issac, B., “Different firewall techniques: a survey”, International Conference on Computing, Communication and Networking Technologies (ICCCNT), IEEE, 2014.

[7]. Meng, Y.-X., “The practice on using machine learning for network anomaly intrusion detection”, International Conference on Machine Learning and Cybernetics (ICMLC), vol. 2, IEEE, 2011.

[8]. SumaiyaThaseenIkram, Aswani Kumar Cherukuri, “Intrusion detection model using fusion of chi-square feature selection and multi class SVM”, Journal of King Saud University –Computer and Information Sciences, 2016.

[9]. Manjula C. Belavagi and BalachandraMuniyal, “Performance Evaluation of Supervised Machine Learning Algorithms for Intrusion Detection, Procedia Computer Science”, Elsevier, 2016.

[10]. Saad Mohamed Ali Mohamed Gadal and Rania A. Mokhtar, “Anomaly Detection Approach using Hybrid Algorithm of Data Mining Technique”, International Conference on Communication, Control, Computing and Electronics Engineering, IEEE, 2017.

[11]. Ibrahim, H. E., Badr, S. M., &Shaheen, M. A., “Adaptive layered approach using machine learning techniques with gain ratio for intrusion detection systems”, International Journal of Computer Applications, vol. 56, issue 7, pp. 10–16, 2012.

[12]. Wen Feng, Qinglei Zhang, Gongzhu Hu, Jimmy Xiang Huang, “Mining network data for intrusion detection through combining SVMs with ant colony networks”, Elsevier, Vol 37, pp 127-140, 2014.

[13]. Nutan Farah Haq, MusharratRafni, AbdurRahmanOnik, Faisal Muhammad Shah, Md. Avishek Khan Hridoy and Dewan Md. Farid, “ Application of machine Learning Approaches in Intrusion Detection System : A Survey”, (IJARAI) International Journal of Advanced Research in Artificial Intelligence, Vol. 4, No. 3, 2015.

[14]. Kuang, F., Xu, W., & Zhang, S., "A novel hybrid KPCA and SVM with GA model for intrusion detection", Applied Soft Computing Journal, vol. 18, pp. 178–184, 2014.
[15]. PrasantaGogoi, D.K. Bhattacharyya, B. Borah1 and Juga, K. Kalita, "MLH-IDS: A Multi-Level Hybrid Intrusion Detection Method", The Computer Journal, Vol. 57 issue 4, pp. 602-623, 2014.
[16]. TaeshikShon "A Machine Learning Framework for Network Anomaly Detection using SVM and GA", IEEE, 2005.
[17]. YadigarImamverdiyev "Anomaly detection in network traffic using extreme learning machine", IEEE, 2016.
[18]. AthanasiosTsiligkaridis "Anomaly Detection In Transportation Networks Using Machine Learning Techniques", IEEE, 2017.
[19]. WathiqLaftah Al-Yaseen, Zulaiha Ali Othman, MohdZakree Ahmad Nazri, "Multi-Level Hybrid Support Vector Machine and Extreme Learning Machine Based on Modified K-means for Intrusion Detection System", International Journal in Expert Systems With Applications, Elsevier, 2017.
[20]. HebatallahMostafaAnwer, Mohamed Farouk, Ayman Abdel Hamid, "A Framework for Efficient Network Anomaly Intrusion Detection with Features Selection", IEEE, 2018.
[21]. Dharmendra Kumar, Ravi Singh Pippal, "A comprehensive review on intrusion detection system and techniques," In: Proceedings of the Conference on Contemporary Technological Solutions towards fulfillment of Social Needs, August 31, 2018, Bhopal, India, pp. 133-137.

Ravi Singh Pippal "Analysis Of Data Normalization With Multilevel Classifiers For Intrusion Detection "International Journal of Computational Engineering Research (IJCER), vol. 08, no. 09, 2018, pp 61-67