Int.J.HIT.TRANSC:ECCN.Vol.5: Issue 2A (2019) Page 9-16 ISSN: 0973-6875



Available Online at www.hithaldia.in/locate/ECCN
All Rights Reserved

ORIGINAL CONTRIBUTION

Efficient Intrusion Detection System based on SMOTE and PCA

[1]Angana Chakraborty and [2]Subhankar Joardar

[1]*Department of Computer Science & Engineering, Haldia Institute of Technology, Haldia, Purba Medinipur, West Bengal*

[1]*Department of Computer Science & Engineering, Haldia Institute of Technology, Haldia, Purba Medinipur, West Bengal*

ABSTRACT

Intrusion Detection System (IDS) is one of the determining factors of almost every security domain nowadays. Imbalanced class distribution and curse of dimensionality are the two factors that affect the efficiency of Intrusion Detection Dataset. In this paper, we have balanced the dataset using SMOTE (Synthetic Minority Oversampling Technique) and after balancing, PCA (Principal Component Analysis) has been implemented for feature extraction. The paper presents an efficient intrusion detection framework with high detection rate of individual classes for multi-class classification (Normal, IDS, Malware, Shellcode, Unknown) of Kyoto 2006+ dataset using Random Forest algorithm.

KEYWORDS: Intrusion Detection System (IDS), SMOTE (Synthetic Minority Oversampling Technique), PCA (Principal Component Analysis), Classification, Kyoto 2006+ dataset.

1. INTRODUCTION

The huge growth of Internet Technology has driven to an increasing number of cyberattacks day by day. Protecting data from vulnerabilities is a big challenge for the industries, companies and organizations. Complete protection of data can't be given with Firewall or Encryption only [1].Figure 1 represents minimum of one successful cyberattack percentage from 2016 to 2020 [2]. The graph points out the current alarming condition of network security. Intrusion Detection System (IDS) is one of the challenging research areas to deal with network vulnerabilities. Based on the data source, Intrusion Detection System is divided into two categories: 1)Host-based IDS 2)Network-based IDS. Operations on a single computer or host are examined by host-based IDS whereas individual packet flow through the network is evaluated in network-based IDS. IDS are categorized into two techniques based on the type of intrusions detection operations. They are Misuse Detection or Signature Detection and Anomaly Detection [3]. Misusedetection can't deal with novel or unknown attacks because they have a predefined known attack pattern. But Anomaly detection can detect unknown or novel attacks. High false alarm rate is one of the greatest disadvantages of anomaly detection. There are three categories of Anomaly detection system [4], named i) Statistical-based ii) Knowledge-based iii) Machine learning-based. Several machine learning techniques are used to deal with high detection rate and low false alarm rate. Machine learning-based approaches are divided into four types: 1) Supervised 2) Unsupervised 3) Semi-supervised 4) Reinforcement. Supervised machine learning techniques deal with labeled training data that consists of input vector pairs and class labels in contrast to unsupervised machine learning which does not consist of labeled training data. Unsupervised learning finds hidden patterns in the unlabeled training data. Support Vector Machine, Random Forest, etc. algorithms fall under supervised category whereas unsupervised supports K-Means clustering, C-Means etc. Semi-supervised machine learning takes minimum labeled training data with maximum unlabeled training data and falls under supervised and unsupervisedlearning. In reinforcement learning software agents achievethe goals by interacting with an environment.

*Corresponding Address: angana.chakraborty9@gmail.com

Our research goal is to build an efficient Network Intrusion Detection System with high individual detection rate. In our research, we have used SMOTE (Synthetic Minority Oversampling Technique) for dataset balancing and PCA (Principal Component Analysis) based feature extraction technique and after that, we have tested the proposed Intrusion Detection System using Random Forest algorithm. We have implemented binary and multi-class classification of Kyoto 2006+ dataset.

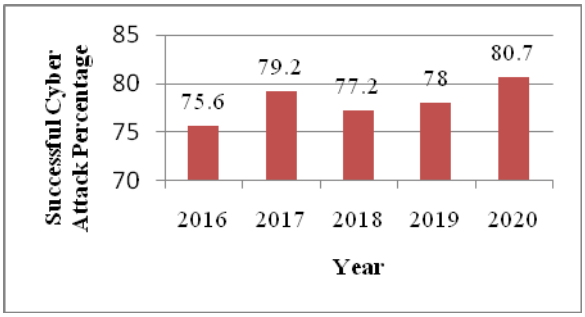


Figure 1: Minimum of one successful cyber attack percentage (Cyber Edge Group 2020 Cyberthreat Defense Report) [2]

2. LITERATURE REVIEW

Many types of research are going on Intrusion Detection System by applying different efficient approaches using different datasets like Kdd Cup 99, NSL-KDD dataset, Kyoto 2006+ dataset, etc. We have reviewed different research papers on IDS. Summary of our review work is explained in tabular form below:

We have surveyed that Patel et al. [1] have built a classifier that gives an accuracy of 99.56%, detection rate of 85.78%, false alarm rate of 0.45%, and error rate of 0.43 using NSL-KDD dataset. The proposed classifier showsbetter performance than other traditional classifiers (J48, Random forest, Random Tree). Authors have simulated NSL-KDD dataset which can't handle existing real network traffics.

Belavagi et al. [5] have compared performance of some supervised machine learning classifiers (Random Forest, Logistic Regression, Gaussian Naive Bayes, and SVM) using NSL-KDD dataset. Random Forest outperforms other classifiers concerning accuracy (99%), precision (0.99), recall (0.99) and f1-score (0.99) for binary classification. Authors have not implemented multi-class classification and they have not used any feature selection technique.

Revathi et al. [6] have investigated various machine learning algorithms (Random Forest, J48, SVM, CART, Naïve Bayes) performances usingNSL-KDD dataset and they have used CFS subset dimensionality reduction technique. Random Forest shows highest accuracy for all attack classes (99.8% accuracy for Normal, 99.1% accuracy for DOS, 98.9% accuracy for Probe, 98.7% accuracy for U2R, and 97.9% accuracy for R2L). Researchers have simulated NSL-KDD dataset which can't handle the aforesaid traffics.

Park et al. [7] have evaluated multi-class classification of Kyoto 2006+ dataset for different types of attacks (normal, unknown, shellcode, IDS+shellcode, malware, IDS). Good detection rate is obtained for malware (0.99), IDS+shellcode (0.91), and unknown attacks (1.00) but shellcode attack gives poor detection rate of 0.11. IDS and normal attack give average detection rate of 0.70 and 0.74 respectively. So, the result obtained by authors does not show good detection rate for all attacks.

Jabbar et al. [8] have built an ensemble classifier (RFAODE) for binary classification of Kyoto 2006+ dataset using Random forest, One-Dependence Estimator (AODE) algorithms. RFAODE shows highest accuracy of 90.51%, detection rate of 92.38 and lowest false alarm rate of 0.14 in comparison with other approaches (Random forest, Naïve Bayes, AODE). Authors have not evaluated multi-class classification.

Zhao et al. [9] have developed filter-based feature selection algorithm called redundant penalty between features(RPFMI) which achieves a higher accuracy of 99.772% on DOS class of KDD Cup 99 and higher accuracy of 97,749% for Kyoto 2006+ dataset for large samples. The algorithm obtained higher accuracy of 96.195 on U2R class and higher accuracy of 91.077% on R2L class for small samples.

Researchers have not evaluated accuracy of all classes in a single sample.

Sarkate et al. [10] have implemented SVM classifier with direct kernel method to obtain high accuracy, detection rate and low false positive rate on KDD Cup 99, NSL-KDD and Kyoto 2006+ datasets. KDD Cup 99 has highest accuracy of 99.87% compared to other datasets (NSL-KDD (98.81%) and Kyoto 2006+ (97.15%)). Highest detection rate, lowest false alarm rate is also obtained by KDD Cup 99 dataset. Auhors have not implemented multi-class classification and they have used KDD Cup 99 which contains duplicate records.

3. KYOTO 2006+ DATASET

NSL-KDD dataset [11] is the most publicly utilized intrusion detection dataset since 2000 [12], but the most important disadvantage of this dataset is that it can't deal with existing real networks traffics. For our research, we have used an efficient dataset called Kyoto 2006+ dataset [13] which was based on real traffic data from November 2006 to August 2009.14 statistical features of KDD Cup 99 [14] dataset were captured [15] from raw traffic data based on honey pot systems [16] of the Kyoto University. In addition to the 14 features, 10 more features [15] were also derived for investigating the intrusion detection system. Out of the 24 features of Kyoto 2006+ dataset, we have taken all 14 conventional features from KDD Cup 99 dataset plus feature no 10 ('Protocol') among the additional features of Kyoto 2006+ dataset for detecting the intrusions efficiently [8]. These features are written below in Table 1:

Table 1: Details of features selected for Kyoto 2006+ dataset

Feature No	Name of the Feature
1	Duration
2	Service
3	Source bytes
4	Destination bytes
5	Count
6	Same_srv_rate
7	Serror_rate
8	Srv_serror_rate
9	Dst_host_count
10	Dst_host_srv_count
11	Dst_host_same_src_port_rate
12	Dst_host_serror_rate
13	Dst_host_srv_serror_rate
14	Flag
Additional Feature (4)	Label
Additional Feature (10)	Protocol

In this paper, we have deployed multi-class classification to examine the performance of different attacks individually. Out of the 24 features of Kyoto 2006+ dataset, we have observed that additional feature no 1 'IDS_detection' represents IDS attack, additional feature no 2 'Malware_detection' represents malware attack, additional feature no 3 'Ashula_detection' represent shellcode and exploit codes attacks. We have classified total 5 attack types based on additional feature no 1, 2, 3. The attacks are categorized based on the following conditions:

1) 'Normal' when IDS_detection, Malware_detection, Ashula_detection features contain '0'
2) 'IDS' when only IDS_detection feature contains a value other than '0'
3) 'Malware' when only Malware_detection feature contains a value other than '0'
4) 'Shellcode' when only Ashula_detection feature contains a value other than '0'
5) 'Unknown' when all the above four conditions are false

Multi-Class Classification of the raw Kyoto 2006+ dataset is given in Table 2.

Table 2: Number of records of all attack types in the Multi-class classification of the raw Kyoto 2006+ dataset (Nov, 2015)

Attack types	Number of records
Normal	251365
IDS	7867
Malware	127
Shellcode	12
Unknown	286

4. PROPOSED FRAMEWORK

In our research, we have proposed an Intrusion Detection framework for binary and multi-class classification of Kyoto 2006+ dataset. The proposed framework is represented below in Figure 2.

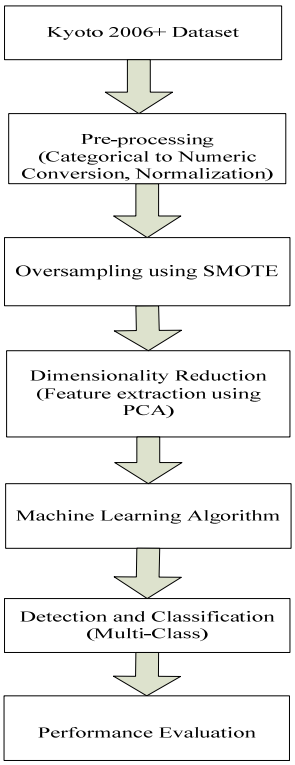


Figure 2: The Proposed Intrusion Detection Framework of Kyoto 2006+ dataset

Steps required to build the proposed framework are described below:

Step 1: Kyoto 2006+ dataset is imported.

Step 2: Pre-processing: Data pre-processing transform raw dataset to clean dataset. It is associated with dealing with missing values, removing duplicate values, conversion of categorical data from numeric data, and normalization of data.

a) Dealing with missing, duplicate values: Dealing with missing values is an important part of pre-processing. We have checked that there are no missing values in our dataset and we have removed duplicate records from the dataset.

b) Conversion of categorical data to numeric data: Among the selected features we found that feature no2('Service'), 14 ('Flag'), additional feature no 10 ('Protocol') have categorical data. So, we have converted those categorical data to numeric data. After that the dataset is split into two parts: i) 80% of the dataset for training ii) 20% of the dataset for testing.Multi-ClassClassification of the training dataset have been depicted in Table 3.

Table 3: Number of records of all attack types in the Multi-Class classification of the training dataset

Attack types	Number of records
Normal	201086
IDS	6304
Malware	102
Shellcode	10
Unknown	223

c) Normalization: Normalization or data scaling method is used to organize records or features in proper format. We have used Min-Max scaling for data normalization. In min-max scaling, all the features for training will be scaled within a fixed range from 0 to 1.

Min-Max scaling is illustrated in the equation below:

$$X_{norm} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \text{ ----------(i)}$$

In this equation, X_{norm} is the normalization result, X is the initial value before normalization, $X_{\max}$ represents maximum values of every feature; $X_{\min}$ represents minimum values of every feature.

Step 3: Oversampling using SMOTE: Balanced class distribution is a crucial task in data mining domain. Classification process complexity

increases due to imbalanced dataset and often the classifier is inclined towards the less frequent classes of data. Datasets with very less records of the minority class produce classifiers that give very higher detection rate for the majority class (es), but very lower detection rate for the minority class. Imbalanced class problem can be dealt by re-sampling (over sampling or under sampling) or synthesizing the infrequent classes. From Table 2 and Table 3, we have observed that Kyoto 2006+ dataset suffers from imbalanced binary and multi-class distribution (number of records of majority class is much higher than the minority classes). To overcome this issue, we have implemented a technique called Synthetic Minority Oversampling Technique (SMOTE)[17]. Minority classes' records are synthetically produced by SMOTE for dataset balancing. SMOTE is unique from all other oversampling methods because SMOTE doesn't produce duplicate records but similar records are produced by it. Duplication of some records increases the chance of overfitting problem. So, relevant approach is to generate records similar with existing records without duplicating the records [18]. After applying SMOTE, 228801 numbers of records have been considered for all attack types in the binary classification and 201086 numbers of records have been considered for all attack types in the multi-class classification of the training dataset.

Step 4: Feature extraction using PCA: Extracting the essential input featuresfrom dataset is an important step of intrusion detection. Feature extraction technique is also called dimensionality reduction where dimension of input features is reduced. Redundant and unnecessary features degrade the performance of intrusion detection system as well as it makes the database very complex. So, identifying important input features is very necessary. With the help of dimensionality reduction technique, speed of execution of machine learning algorithms becomes fast and memory requirement is also reduced [3]. We have used Principal Component Analysis (PCA) for dimensionality reduction.

Principal Component Analysis (PCA): PCA falls under the most popular dimensionality reduction techniques for statistical data analysis which transforms huge set of input features to a smaller set of input features (called Principal Component (PC)) while preserving the maximum information in the large set. For most of the datasets, first several PCs with large variances are chosen and the remaining PCs are omittedwith a minimum variance loss for dimensionality reduction [3].

Step 5: We have used Random Forest machine learning algorithm for classifying the network intrusions after oversampling and feature extraction.

Random Forest (RF): Random Forest is a supervised algorithm that supports classification, regression both. Multiple decision trees are constructed for training the model and results of predictions from the trees are combined to create the forest. Accuracy is directly dependent on the number of trees and thus overfitting can be avoided [19].

Step 6: Performance evaluation of results is discussed in Section 5.

5. EXPERIMENTAL RESULT ANALYSIS

This section provides an analysis multi-class classification of Kyoto 2006+ dataset using Random Forest algorithm after implementing SMOTE and PCA. The experiments are performed on Anaconda Python 3.7.1 Individual Edition (provides free and open-source Python distribution for scientific computing) using Jupyter Notebook Editor (consists of many python packages).

We have chosen different evaluation parameters to correctly implement the proposed Intrusion Detection System. The parameters are defined below:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \text{ ------(ii)}$$

Approach	Attack Types	Precision	Recall/Detection rate	F1-Score
PCA+Random Forest	Normal	0.99	1.00	0.99
	IDS	0.82	0.70	0.75
	Malware	0.96	1.00	0.98
	Shellcode	1.00	1.00	1.00
	Unknown	0.97	0.95	0.96
Proposed Approach(SMOTE+ PCA+ Random Forest)	**Normal**	1.00	**0.97**	0.98
	IDS	0.44	**0.86**	0.58
	Malware	0.96	**1.00**	0.98
	Shellcode	1.00	**1.00**	1.00
	Unknown	0.90	**0.97**	0.93

$$\text{Precision} = \frac{TP}{(TP + FP)} \text{ ------(iii)}$$

$$\text{Recall / Detection Rate} = \frac{TP}{(TP + FN)} \text{ ------(iv)}$$

$$\text{F1-Score} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} \text{ -------(v)}$$

At first, we have evaluated the precision, recall, andF1-score of all attack types for multi-class classification of Kyoto 2006+ dataset using PCA based feature extraction technique. After that, we have implemented the proposed approach using SMOTE for balancing the dataset and PCA for feature extraction. We have compared the proposed approach's performance without and with using SMOTE.

Table 4: Precision, Recall, F1-score for multi-class classification without and with using SMOTE (Kyoto 2006+ dataset)

From Table 4, we have observed that recall/detection rate of majority class ('Normal') is much higher (1.00) than the other minority classes (i.e. IDS, 0.70, Unknown, 0.95) without using SMOTE. But after implementing the proposed approach using SMOTE, detection rate of minority class improves (IDS, 0.86, Unknown, 0.97). So, the proposed approach shows balanced performance of all attack types.
We have compared our proposed approach's multi-class classification accuracy and weighted average performance of precision, recall, F1-score with the other approaches.

6. CONCLUSION

This paper represents an efficient intrusion detection framework based on binary and multi-class classification of Kyoto 2006+ dataset using Random Forest algorithm. First, we have evaluated the precision, recall, and F1-score of all attack types for binary, multi-class classification of Kyoto 2006+ dataset using PCA based feature extraction technique. But, we found that our dataset is imbalanced (biased towards majority classes). So, we have implemented our proposed approach using SMOTE (Synthetic Minority Oversampling Technique) for balancing the dataset and PCA for feature extraction. We have compared the proposed approach's performance without and with using SMOTE. After balancing the dataset using SMOTE and feature extraction using PCA, our proposed approach gives good detection rate for individual attacks in binary and multi-class classification. We have also observed that our proposed approach outperforms the other approach proposed by Park et al. [7].
In the future, we will again evaluate multi-class classification for better performance of all attacks under different training conditions. We will also implement genetic algorithm, particle swarm optimization, etc. techniques to obtain an optimized subset of features and improved accuracy.

References

[1] Patel, V., Shrivastava, A., Nagar, S.,(2019). An Efficient Technique for Classification of Intrusion Detection Dataset Using Machine Learning Technique. International Journal Of Innovative Trends In Engineering, ISSN. 2395-2946, Issue. 83, Vol. 59, Number 01

[2] https://www.comparitech.com/vpn/cybersecurity-cyber-crime-statistics-facts-trends

[3] B. N.,(2012).Classification of Intrusion Detection Dataset using machine learning Approaches.International Journal of Electronics and Computer Science Engineering, ISSN. 2277-1956

[4] Garcıa-Teodoroa, P., Dıaz-Verdejoa, J. Macia-Fernandeza, G., Vazquezb, E., (2009). Anomaly-based network intrusion detection: Techniques, systems and challenges. Computers & Security 28, pp. 18–28

[5] Belavagi, C, M., Muniyal, B.,(2016). Performance Evaluation of Supervised Machine Learning Algorithms for Intrusion Detection. Twelfth International Multi-Conference on information Processing, 117-123

[6] Revathi, S., Malathi, A.,(2013).A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection. International Journal of Engineering Research & Technology, Vol. 2, Issue. 12, ISSN. 2278-0181

[7] Park, K., Song, Y., Cheong, Y-G.,(2018).Classification of Attack Types for Intrusion DetectionSystems using a Machine Learning Algorithm.IEEE Fourth International Conference on Big Data Computing Service and Applications

[8] Jabbar, M A., Aluvalu, R., S, Reddy, Satyanarayana S.,(2017).RFAODE: A Novel Ensemble Intrusion Detection System.7th International Conference on Advances in Computing & Communications, 226-234

[9] Zhao, F., Zhao, J., Niu, X., Luo, S., Xin, Y.,(2018).A Filter Feature Selection Algorithm Based on Mutual Information for Intrusion Detection. Applied Sciences 8(9):1535

[10] Sarkate, Sopan, S., Qureshi, R.,(2017).Direct Kernel Method for Machine learning with Support Vector Machine. International Journal of Advanced Research in Computer and Communication Engineering, Vol. 6, Issue. 12

[11] Nsl-kdd data set for network-based intrusion detectionsystems,http://nsl.cs.unb.ca/KDD/NSLKDD.html

[12] Chae, H., Jo B., Choi, S., Park, T.,(2013). Feature Selection for Intrusion Detection using NSL-KDD.Recent Advances in Computer Science, pp. 184-187

[13] http://www.takakura.com/Kyoto data

[14] KDD Cup 1999, http://kdd.ics.uci.edu/databases/kddcup 99/kddcup99.html

[15] Song, J., Takakura, H., Okabe, Y., Description of Kyoto University Benchmark Data

[16] Song, J., Takakura, H., Okabe, Y., (2008).Cooperation of Intelligent Honeypots to DetectUnknown Malicious Codes. WOMBATWorkshop on Information Security Threat Data Exchange, The IEEE CS Press

[17] Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P.,(2002). SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, Vol. 16, pp. 321–357

[18] Hamid, Y., Sugumaran, M., Journaux L.,(2016).A Fusion of Feature Extraction and Feature Selection Technique for Network Intrusion Detection. International Journal of Security and Its Applications, Vol. 10, No. 8, pp.151-158

[19] Saranyaa,T., Sridevib,S., Deisyc, C.,Chungd, Duc, T., Khane, Ahamed, ,(2020).M.K.A.Performance Analysis of Machine Learning Algorithms in Intrusion Detection System: A review. Third International Conference in Computing and Network Communications, Procedia Computer Science 171, pp. 1251–1260

[20] Mitchell, T.,(1997). Machine learning. In: MacHraw Hill, Inc