Network Intrusion Detection Combined with Hybrid Sampling and Machine Learning

Md. Ismail Hossain

Md. Sadiqul Amin

I. Abstract

In network security Intrusion Detection System(IDS) plays an important role by discovering and preventing malicious activities.Because of the complexity and time-varying network environment, the network intrusion samples are submerged into a large number of normal samples, which is insufficient to train a model and detection result with high false detection rate. For this problem of data imbalance, we propose a network intrusion detection algorithm combined with hybrid sampling with machine learning. At first we use OSS (One Side Selection to reduce the noise samples in majority category,and then increase the Minority Samples by SMOTE (Synthetic Minority Over-sampling Technique). Using this hybrid sampling we can make a balanced data set and make the model fully learn the features of minority samples and greatly reduce the model training time. Secondly we use SVM (Support Vector Machine) to extract the feature and Carn Decision tree and also compare their result based on their accuracy and time complexity.The proposed network intrusion detection algorithm was verified by experiment on the NSL-KDD data set, and the classification accuracy can achieve 89.3

II. Introduction

The Internet Technology is continuously and rapidly developing, it's now become a inseparable tool for peoples daily life and great impact on changing people's life-style. With the increasing of the usage of the network technology the various network attack methods are also increasing, attack method are being updated, frequency of the attacks are increasing and network security issues are becoming serious. The main task of a NIDS (Network Intrusion Detection System) is to detect suspicious attacks and take corresponding measure to protect the network from attack and prevent the economic losses. Network intrusion detection is now becoming an important research content in the field of Network Security.

At present the commonly used intrusion detection system

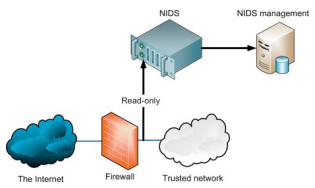

Fig. 1. An image of a NIDS

are divided into misuses detection and anomaly detection, but these two model have disadvantages of low accuracy and high false positive rate. Artificial intelligence based detection method i a hotspot in research of intrusion detection system. However there are some challenges in the implementation of the IDS:

(1) Dealing with large-scale, high-dimensional data points traditional NIDS approaches tends to apply dimension reduction to remove noise in measurements. It is likely to remove significant information when extracting features for intrusion detection rate.

(2) Another generic problem for building a NIDS using machine learning is imbalanced data. The imbalanced data will affect the performance of the model and leading to high false alarm rate and high miss rate of some minority classes samples. (3) The extraction of the feature data from network traffic data is sometimes complicated.

III. Literature Review

To build an efficient intrusion detection system, researchers use machine learning to detect various types of attacks[1]. Ünal Çavu³o§lu proposed that layered architecture is created by determining appropriate machine learning algorithms according to attack type [2]. The most widely studied algorithms are Support Vector Machine (SVM), Naive Bayes, Random Forest (RF) and other clustering algorithms. Zhao proposed the least squares support vector machine (LSSVM) algorithm based on hybrid kernel function, and each parameter of LSSVM is optimized using particle swarm optimization (PSO) algorithm[3]. Thaseen proposed to use Chi-square feature selection and SVM, Modied Naïve Bayes (MNB) and LPBoost integration to build an intrusion detection model. And the prediction of the class label was decided by a majority voting of SVM, MNB and LPBoost[4]. Sumaiya proposed an intrusion detection model using chi-square feature selection and multi class support vector machine which could get a better detection rate and reduced false alarm rate [5]. Tao proposed the feature selection, weight, and parameter optimization of support vector machine based on the genetic algorithm (FWP-SVM-GA) [6]. Compared with other SVM based intrusion detection algorithms, the detection rate is higher and the false positive and false negative rates are lower. Peng proposed a clustering method for IDS based on Mini Batch K-means combined with Principal Component Analysis (PCA), and the clustering method can be used for IDS over big data environment [7]. Farnaaz built a model for intrusion detection

system using RF classier, the model was efcient with lowfalse alarm rate and high detection rate [8].

IV. Proposed Detection Model

The proposed NIDS includes 4 parts. These are - (1) This article uses One Sided Selection for down-sampling the data-set to remove the nose while reducing the majority samples, and then to increase the minority samples, uses Synthetic Minority Over-sampling Technique [9]
(2) The symbolic feature attributes are converted to numeric and then normalized to get a standard format.
(3) SVM and CART are used to extract the features of the data to improve the accuracy of classification.
(4) After training, a model with good classification performance is obtained, and the model is used to classify the test set to obtain excellent classification results.

A. DATA PREPROCESSING

After applying heatmap we found that some feature are highly co-related [Fig-2].

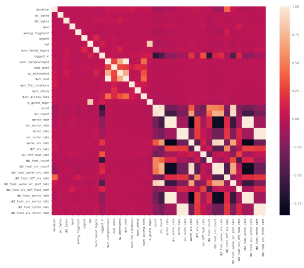

Fig. 2. An image of Heatmap representing co-relation among variables

We observe that srv_serror_rate, srv_rerror_rate, dst_host_srv_serror_rate, num_root, dst_host_serror_rate, dst_host_rerror_rate, dst_host_srv_rerror_rate, dst_host_same_srv_rate these variables are highly co-related. So, We droped these variable from our dataset for better performance and escape curse of dimesionality

B. CONSTRUCTION OF BALANCED DATA SET BASED ON HYBRID SAMPLING

Network traffic data is composed of a large amount of normal traffic and a small amount of abnormal traffic, which is a typical imbalanced data classification problem. In this case, when the overall error is minimized, although the prediction accuracy of some majority classes is improved, the prediction accuracy of minority classes is often very low.

One-side selection is an under-sampling method result ing from the application of Tomek links followed by the application of KNN.
SMOTE is an over-sampling method. Its main idea is to form new minority class samples by interpolating between several minority class samples that lie together. Thus, the overfitting problem is avoided. At the same time, it causes the decision boundaries for the minority class to spread fur-ther into the majority class space.
The imbalanced dataset is passed through OSS to minimize

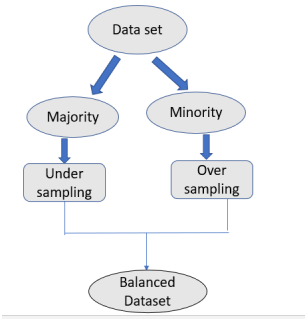

Fig. 3. An image of a Hybrid Sampling

the oversampled data. Then the under-sampled dataset is run through SMOTE to increase the number of minority class data. Hence, a balanced dataset is introduced.

C. DEEP NETWORK MODEL

To extract the feature from balanced dataset, Support Vector Machine is used.

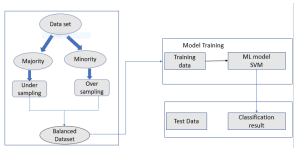

Fig. 4. An image of training and testing SVM model with balanced dataset

We also train a supervised learning method CART decision tree with same dataset. During the test we found that the CART decision tree even perform better. The accuracy jumped to 99.51

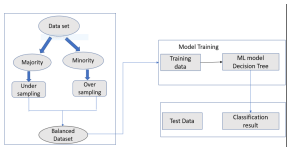

Fig. 5. An image of training and testing CART model with balanced dataset

V. Implementation

First the data is loaded from NSL-KDD dataset. Although most data colums are numeric, some string type columns are encoded to numeric value with OneHot encoder. Then heatmap is generated to find out highly correlated features. Correlation heatmap is graphical representaion fo correlation matrix represnting correlation between variables. As highly correlated features are not efficient to classify correctly, these features are dropped to make the classification process faster. The processed dataset is now ready to be train and tested. For model evaluation, SVM and CART decision tree is used. To

train these models, the data set is split into 67:33 ratio. Then SVM and CART is trained with same data from 67% of the data set. After that, each model is tested with remaining 33% data.

VI. Results

Fscore and Confusion matrix will be added later

Our model will be evaluated based on accuracy and training and detection time duration. Table I shows the training time, testing time.

TABLE I
SVM RESULT

Training Time	98.08
Testing Time	14.75
Accuracy	0.95

Table II contains performance of CART decision tree.

TABLE II
CART RESULT

Training Time	0.73
Testing Time	0.01
Accuracy	0.99

Table III shows comparative status of existing work with hybrid sampling, SVM and CART decision tree. All the works uses NSL-KDD dataset along with similar sampling technique, oversampled with SMOTE and undersampled with OSS. All these works shows different accuracy.

TABLE III
RESULT COMPARISON

Feature	**ExistingWork**	**SVM**	**CART**
Dataset	NSL-KDD	NSL-KDD	NSL-KDD
Sampling Technique	*SMOTE + OSS*	*SMOTE + OSS*	*SMOTE + OSS*
ML Model	*CNN-BiLSTM*	*SVM*	*CART*
Accuracy	*84.95%*	*95.39%*	*99.51%*

VII. Conclusion

In this paper a method for intrusion detection system based on the combination of hybrid sampling and machine learning is proposed and discussed. Firstly we used SMOTE and OSS to make a balanced dataset for model training. It can reduce the training time of the model and solves the common problem to some extend of inadequate training from imbalanced dataset. In addition, a network data pre-processing method is established for complex, multidimensional cyber threats, which is suitable for proposed deep hierarchical network model. We use this data to train SVM and CARN model.
NSL-KDD intrusion dataset is used to employed and evaluate our models. Based on statistical significance it could be concluded that our proposed approach outperform the CNN-BiLSTM model. The proposed method yields the superior result in term of accuracy and precision when validated in testing set.

References

[1] X. Wang, "Smote:"design of temporal sequence association rule-based intrusion detection behavior detection system for distributed network," *Modern Electron*.

[2] Ü.Çavu³o§lu, "A new hybrid approach for intrusion detection using machine learning methods," *Appl. Intell*.

[3] Z. Fuqun, ""detection method of lssvm network intrusion based on hybrid kernel function,"," *Modern Electron. Techn.,*.

[4] C. A. K. I. S. Thaseen and A. Ahmad, ""integrated intrusion detection model using chi-square feature selection and ensemble of classicers,"," *Arabian J. Sci. Eng.*

[5] C. A. K. I. S. Thaseen, ""intrusion detection model using fusion of chi-square feature selection and multi class svm,"," *J. King Saud Univ.*

[6] Z. S. P. Tao and Z. Sun, ""an improved intrusion detection algorithm based on ga and svm,"," *IEEE Access*.

[7] V. C. M. L. K. Peng and Q. Huang, ""clustering approach based on mini batch k-means for intrusion detection system over big data,"," *IEEE Access*.

[8] N. Farnaaz and M. A. Jabbar, ""random forest modeling for network intrusion detection system,"," *Procedia Comput. Sci.*

[9] M. Iftikhar, T. Singh, B. Landfeldt, and M. Caglar, "Smote: Synthetic minority over-sampling technique," *Journal Of Artificial Intelligence Research*, vol. 16, pp. 321–357, Mar 2002.