IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Special Issue 1, April 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.118

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology
An ISO 3297: 2007 Certified Organization Volume 11, Special Issue 1, April 2022
9th National Conference on Frontiers in Communication and Signal Processing Systems (NCFCSPS '22)
28 th -29 th April 2022
Organized by
Department of ECE, Adhiyamaan College of Engineering, Hosur, Tamilnadu, India

Intrusion Detection of Imbalanced Network Traffic Based on Machine Learning and Deep Learning

Mr.Martin Joel Rathnam[1], P.Mukesh[2], K.Rajesh[3], P.Raj Kumar[4], R.Saran Kumar[5]

Assistant Professor, Department of Electronics and Communication Engineering, Adhiyamaan College of Engineering, Krishnagiri, Tamil Nadu, India [1]

U.G Students, Department of Electronics and Communication Engineering, Adhiyamaan College of Engineering, Krishnagiri, Tamil Nadu, India [2,3,4,5]

ABSTRACT: In imbalanced network traffic, malicious cyber-attacks can often hide in large amounts of normal data. It exhibits a high degree of stealth and obfuscation in cyberspace, making it difficult for Network Intrusion Detection System (NIDS) to ensure the accuracy and timeliness of detection. This paper researches machine learning and deep learning for intrusion detection in imbalanced network traffic. It proposes a novel Difficult Set Sampling Technique (DSSTE) algorithm to tackle the class imbalance problem. First, use the Edited Nearest Neighbour (ENN) algorithm to divide the imbalanced training set into the difficult set and the easy set. Next, use the Means algorithm to compress the majority samples in the difficult set to reduce the majority. Zoom in and out the minority samples' continuous attributes in the difficult set synthesize new samples to increase the minority number. Finally, the easy set, the compressed set of majority in the difficult, and the minority in the difficult set are combined with its augmentation samples to make up a new training set.

KEYWORDS: IDS, imbalanced network traffic, machine learning and deep learning

I.INTRODUCTION

With the rapid development and wide application of 5G, IoT, Cloud Computing, and other technologies, network scale, and real-time traffic become more complex and massive, cyber-attacks have also become complex and diverse, bringing significant challenges to cyberspace security. As the second line of defense behind the firewall, the Network Intrusion Detection System (NIDS) needs to accurately identify malicious network attacks, provide real-time monitoring and dynamic protection measures, and formulate strategies. James Anderson first proposed the concept of intrusion detection in 1980, and then some scholars applied machine learning methods in intrusion detection .However, due to the limitation of computer storage and computing power at that time, machine learning failed to attract attention. With the rapid development of computers and the emergence and promotion of Artificial Intelligence (AI) and other technologies, many scholars have applied machine learning methods to network security. They have achieved certain results. In real cyberspace, normal activities occupy the dominant position, so most traffic data are normal traffic; only a few are malicious cyber attacks , resulting in a high imbalance of categories.

II. RELATED WORKS

INTRUSION DETECTION SYSTEM (IDS) In the research of network intrusion detection based on machine learning, scholars mainly distinguish normal network traffic from abnormal network traffic by dimensionality reduction, clustering, and classification, to realize the identification of malicious attacks. Pervez proposed a new

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology
An ISO 3297: 2007 Certified Organization Volume 11, Special Issue 1, April 2022
9th National Conference on Frontiers in Communication and Signal Processing Systems (NCFCSPS '22)
28 th -29 th April 2022
Organized by
Department of ECE, Adhiyamaan College of Engineering, Hosur, Tamilnadu, India

method for feature selection and classification merging of multi-class NSL-KDD Cup99 dataset using Support Vector Machine (SVM) and discussed the classification accuracy of classifiers under different dimension features. Shiraz studied some new technologies to improve CANN intrusion detection methods' classification performance and evaluated their performance on the NSL-KDD Cup99 dataset**.** The result shows the CANN detection rate and reduces the failure the alert rate is improved or provides the same performance. Bhattacharya proposed a machine learning model based on hybrid Principal Component Analysis (PCA)-Firefly.

CLASS BALANCING METHODS In the field of machine learning, the problem of category imbalance has always been a challenge. Therefore, intrusion detection also faces enormous challenges in network traffic with extremely imbalanced categories. Therefore, many scholars have begun to study how to improve the intrusion recognition accuracy of imbalanced network traffic data. Piyasak proposed a method to improve the accuracy of minority classification. This method combines the Synthetic Minority Over-sampling Technique (SMOTE) and Complementary Neural Network (CMTNN) to solve imbalanced data classification. Experiments on the UCI dataset show that the proposed combination technique can improve class imbalance problems.

III. METHODOLOGY

Faced with imbalanced network traffic, we propose the Difficult Set Sampling Technique (DSSTE) algorithm to compress the majority samples and augment the number of minority samples in difficult samples, reducing imbalance in the training set that the intrusion detection system can achieve better classification accuracy. We use Random Forest, SVM, XGBoost, LSTM, MiniVGGNet, and AlexNet as classifiers for classification models. We proposed the intrusion detection model shown in Figure 1. Data pre-processing first performed in our intrusion detection structure, including duplicate, outlier, and missing value processing. Then, partitioning the test set and the training set, and the training set processed for data balancing using our proposed DSSTE algorithm.

DSSTE ALGORITHM In imbalanced network traffic, different traffic data types have similar representations, especially minority attacks can hide among a large amount of normal traffic, making it difficult for the classifier to learn the differences between them during the training process. In the similar samples of the imbalanced training set, the majority class is redundant noise data. The number is much larger than the minority class, making the classifier unable to learn the distribution of the minority class, so we compress the majority class. The minority class discrete attributes remain constant, and there are differences in continuous attributes. Therefore, the minority class's continuous attributes are zoomed to produce data that conforms to the true distribution. Therefore, we propose the DSSTE algorithm to reduce the imbalance.

MACHINE LEARNING AND DEEP LEARNING ALGORITHMS: In the classifier's design, we use Random Forest, SVM, XGBoost, LSTM, AlexNet, and Mini-VGGNet to train and test, which are detailed in the following part.

RANDOM FOREST Leo Breiman proposed random Forest in 2001. Random Forest is an excellent supervised learning algorithm that can train a model to predict which classification results in a certain sample type belong to based on a given dataset's characteristic attributes and classification results. Random Forest is based on a decision tree and adopts the Bagging(Bootstrap aggregating) method to create different training sample sets. The random subspace division strategy selects the best attribute from some randomly selected attributes to split internal nodes. The various decision trees formed are used as weak classifiers, and multiple weak classifiers form a robust classifier, and the voting mechanism is used to classify the input samples. After a random forest has established a large number of decision trees according to a certain random rule when a new set of samples is input, each decision tree in the forest makes a prediction on this set of samples separately, and integrates the prediction results of each tree, get a final result.

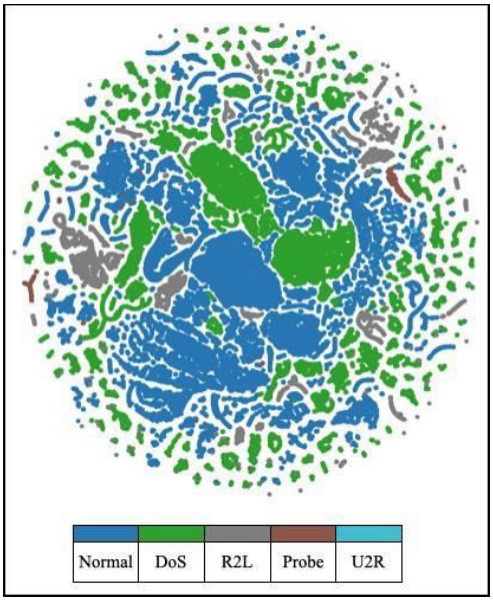
ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology
An ISO 3297: 2007 Certified Organization Volume 11, Special Issue 1, April 2022
9th National Conference on Frontiers in Communication and Signal Processing Systems (NCFCSPS '22)
28 th -29 th April 2022
Organized by
Department of ECE, Adhiyamaan College of Engineering, Hosur, Tamilnadu, India

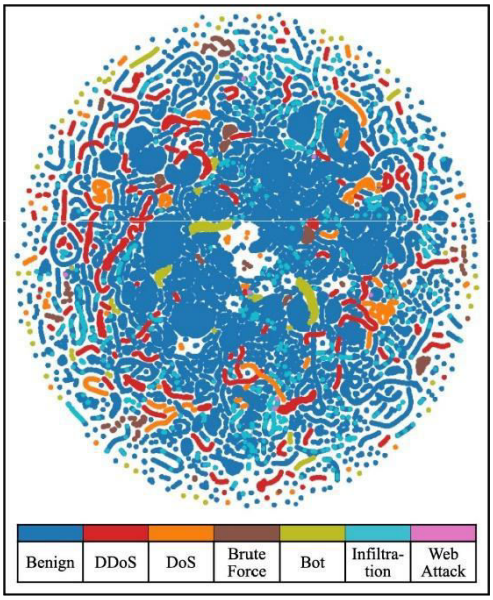
SUPPORT VECTOR MACHINE Coretes and Vapink first proposed support Vector Machine (SVM) in 1995. It shows many unique advantages in a small sample, nonlinear, and high-dimensional pattern recognition and can be extended to other functions such as function fitting Machine learning problems. Before the rise of deep learning, SVM was considered the most successful and best-performing machine learning method in recent decades. The SVM method is based on the Vapnik Chervonenkis(VC) dimension theory of statistical learning theory and the principle of structural risk minimization. Its basic idea is to find a separation hyperplane between different categories, so that different category can be better separated. The SVM method believes that when deciding to separate the hyperplane, only the sample point closest to the hyperplane, as long as the support vector is found, the hyperplane can be determined.

XGBoost It is a parallel regression tree model that combines the idea of Boosting, which is improved based on gradient descent decision tree by Chen and Guestrin . Compared with the GBDT(Gradient Boosting Decision Tree) model, XGBoost overcomes the limited calculation speed and accuracy. XGBoost adds regularization to the original GBDT loss function to prevent the model from overfitting. The traditional GBDT performs a first-order Taylor expansion on the calculated loss function and takes the negative gradient value as the residual value of the current model. In contrast, XGBoost performs a second-order Taylor expansion to ensure the accuracy of the model

IV. EXPERIMENTAL RESULTS



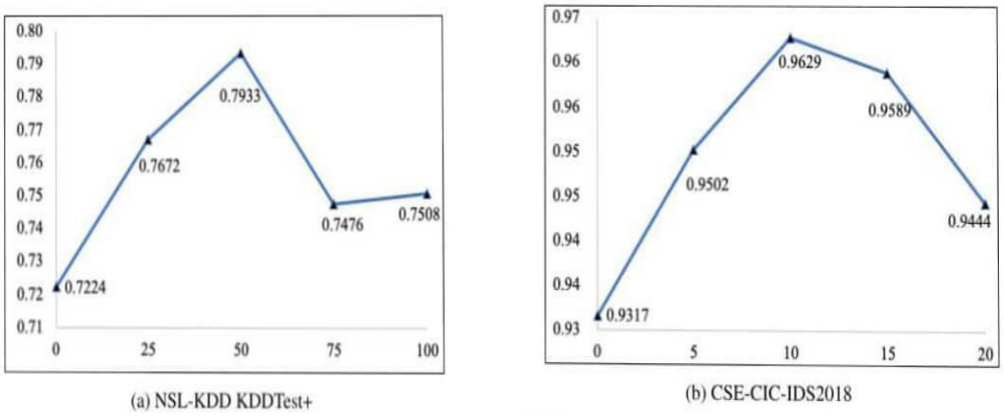
Use t-SNE to visualize NSL-KDD(a)



CSE-CIC-IDS2018(b)

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology
An ISO 3297: 2007 Certified Organization Volume 11, Special Issue 1, April 2022
9th National Conference on Frontiers in Communication and Signal Processing Systems (NCFCSPS '22)
28 th -29 th April 2022
Organized by
Department of ECE, Adhiyamaan College of Engineering, Hosur, Tamilnadu, India



(a) NSL-KDD KDDTest+ (b) CSE-CIC-IDS2018

F1-Score of DSSTE algorithm with different scaling factor K.

V. CONCLUSION

As network intrusion continues to evolve, the pressure on network intrusion detection is also increasing. In particular, the problems caused by imbalanced network traffic make it difficult for intrusion detection systems to predict the distribution of malicious attacks, making cyberspace security face a considerable threat. This paper proposed a novel Difficult Set Sampling Technique(DSSTE) algorithm, which enables the classification model to strengthen imbalanced network data learning. A targeted increase in the number of minority samples that need to be learned can reduce the imbalance of network traffic and strengthen the minority's learning under challenging samples to improve the classification accuracy. We used six classical classification methods in machine learning and deep learning and combined them with other sampling techniques. Experiments show that our method can accurately determine the samples that need to be expanded in the imbalanced network traffic and improve the attack recognition more effectively.

REFERENCES

1. D. E. Denning, ''An intrusion-detection model,'' IEEE Trans. Softw. Eng., vol. SE-13, no. 2, pp. 222–232, Feb. 1987.
2. N. B. Amor, S. Benferhat, and Z. Elouedi, ''Naive Bayes vs decision trees in intrusion detection systems,'' in Proc. ACM Symp. Appl. Comput. (SAC), 2004, pp. 420–424.
3. M. Panda and M. R. Patra, ''Network intrusion detection using Naive Bayes,'' Int. J. Comput. Sci. Netw. Secur., vol. 7, no. 12, pp. 258–263, 2007.
4. M. A. M. Hasan, M. Nasser, B. Pal, and S. Ahmad, ''Support vec- tor machine and random forest modeling for intrusion detection system (IDS),'' J. Intell. Learn. Syst. Appl., vol. 6, no. 1, pp. 45–52, 2014.
5. N. Japkowicz, ''The class imbalance problem: Significance and strate- gies,'' in Proc. Int. Conf. Artif. Intell., vol. 56, 2000, pp. 111–117.
6. Y. LeCun, Y. Bengio, and G. Hinton, ''Deep learning,'' Nature, vol. 521, no. 7553, pp. 436–444, 2015.
7. Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, ''Deep learning for visual understanding: A review,'' Neurocomputing, vol. 187, pp. 27–48, Apr. 2016.
8. T. Young, D. Hazarika, S. Poria, and E. Cambria, ''Recent trends in deep learning based natural language processing [review article],'' IEEE Comput. Intell. Mag., vol. 13, no. 3, pp. 55–75, Aug. 2018.
9. N. Shone, T. N. Ngoc, V. D. Phai, and Q. Shi, ''A deep learning approach to network intrusion detection,'' IEEE Trans. Emerg. Topics Comput. Intell., vol. 2, no. 1, pp. 41–50, Feb. 2018.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology
An ISO 3297: 2007 Certified Organization Volume 11, Special Issue 1, April 2022
9th National Conference on Frontiers in Communication and Signal Processing Systems (NCFCSPS '22)
28 th -29 th April 2022
Organized by
Department of ECE, Adhiyamaan College of Engineering, Hosur, Tamilnadu, India

10. D. A. Cieslak, N. V. Chawla, and A. Striegel, ‘‘Combating imbalance in network intrusion datasets,’’ in Proc. IEEE Int. Conf. Granular Comput., May 2006, pp. 732–737.
11. M. Zamani and M. Movahedi, ‘‘Machine learning techniques for intru- sion detection,’’ 2013, arXiv:1312.2177. [Online]. Available: http://arxiv. org/abs/1312.2177
12. M. S. Pervez and D. M. Farid, ‘‘Feature selection and intrusion classifi- cation in NSL-KDD cup 99 dataset employing SVMs,’’ in Proc. 8th Int. Conf. Softw., Knowl., Inf. Manage. Appl. (SKIMA), Dec. 2014, pp. 1–6.
13. H. Shapoorifard and P. Shamsinejad, ‘‘Intrusion detection using a novel hybrid method incorporating an improved KNN,’’ Int. J. Comput. Appl., vol. 173, no. 1, pp. 5–9, Sep. 2017.

BIOGRAPHY

MR. MARTIN JOEL RATHNAM
ASSISTANT PROFESSOR,
DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING,
ADHIYAMAAN COLLEGE OF ENGINEERING,
ANNA UNIVERSITY.

MUKESH .P
DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING,
ADHIYAMAAN COLLEGE OF ENGINEERING,
ANNA UNIVERSITY.

RAJESH .K
DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING,
ADHIYAMAAN COLLEGE OF ENGINEERING,
ANNA UNIVERSITY.

RAJ KUMAR .P
DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING,
ADHIYAMAAN COLLEGE OF ENGINEERING,
ANNA UNIVERSITY.

SARAN KUMAR .R
DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING,
ADHIYAMAAN COLLEGE OF ENGINEERING,
ANNA UNIVERSITY.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.118

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details