Empirical Investigation of Resampling Techniques in an Intruder Detection System

Arjun Puri and Manoj Kumar Gupta

Abstract

The intruder detection system plays a fundamental function in recognizing assaults in networks. To design an intelligent intruder detection system invites researchers from the machine learning domain to work in this area. With the availability of KDD99 datasets, some researchers encounter a class imbalance problem in it. This article performs a detailed empirical investigation of various resampling techniques to mitigate the effect of class imbalance. The study is performed on NSL-KDD multi-class datasets using fivefold cross-validation with G-Mean and AUC as evaluation metrics considering the decision tree as a classifier. The study inferred that the SMOTE technique performs well compared with the rest of the art.

Keywords

Class Imbalance • Oversampling • Undersampling • Intruder Detection System

1 Introduction

With the coming of innovation and the network of PCs, it likewise expands the danger of assaults. These assaults in networks affect the integrity, confidentiality, and availability of data on the Internet. The widely spread of network increases attacks in various sectors, which either act like eavesdrop or sometimes raise denial of service attacks (Intisar et al., 2019). These attacks may result in the theft of sensitive information on the Internet. There is always a need for an intelligent intruder detection system to secure data from these different attacks. The primary objective of the intruder detection system is to detect intrusion over the Internet. The intruder detection system is needed to be applied at the correspondence level to screen network activities and connections (Zhang et al., 2020). The intruder detection system's design process is divided into two subsections: Signature-based and anomaly-based intruder detection systems (Bedi et al., 2020). Signature-based identify intruders based on the previous attack pattern lead to a problem when a system cannot recognize any new attack. This problem of signature-based intruder detection system needs to address, so anomaly-based system comes into existence. In an anomaly-based intruder detection system, previous as well as new attacks in networks are easily detected.

Anomaly-based intruder detection systems are nowadays improving using different machine learning techniques (Aldweesh et al., 2020). These machine learning techniques are based on probabilities and mainly rely on the distribution of datasets. When there is an imbalance among different classes in datasets, it may result in a class imbalance problem. This problem is often seen in real-time datasets when there is a need to study class with less probability. Traditional classifier's behaviors in imbalanced class datasets are biased toward the majority class.

While studying intruder detection datasets such as KDD'99 and NSL-KDD datasets, the researcher found a class imbalance problem. These datasets contain normal network traffic in the majority than attack instances. To design an effective intruder detection system, researchers try to improve traditional machine learning classification techniques. For improvement, researchers try to use different class imbalance handling techniques to make predictions of traditional classifiers more precise. Many researchers work to handle the class imbalance problem in the intelligent designing of the intrusion detection system. Some

A. Puri (✉) · M. K. Gupta
School of Computer Science and Engineering, Shri Mata Vaishno Devi University, Jammu and Kashmir, Katra, 182320, India
e-mail: Puri.arjun7@gmail.com

M. K. Gupta
e-mail: Manoj.gupta@smvdu.ac.in

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2021

P. K. Singh et al. (eds.), *Innovations in Information and Communication Technologies (IICT-2020)*,
Advances in Science, Technology & Innovation,
https://doi.org/10.1007/978-3-030-66218-9_31

researchers try to use resampling techniques, while others rely on an algorithmic modification to deal with an imbalanced class dataset.

This article provides a comprehensive empirical study of different data level class imbalance handling techniques in intruder detection datasets. This thorough investigation may result in further answering the following research questions.

- Does class imbalance affect the recognization of different attacks using traditional decision tree classifiers?
- Which existing resampling technique is suitable for handling class imbalance in intrusion detection datasets?

The remaining sequences of sections of the article are as follows: related work followed by an experimental framework, then results and discussion, and at the end, the conclusion and also contain future work.

2 Related Work

For effective designing an intruder detection system, the class imbalance problem plays a vital role. In (Gonzalez-Cuautle et al., 2020), authors develop a technique to identify intruders in the intruder detection system using SMOTE with a grid-search considering different machine learning algorithms' optimization procedures. This article work authors deal with the tuning of different techniques and try to find the optimal solution for class imbalance in intruder detection system design. Another article (Rodda and Erothi, 2016), where authors consider Naïve Bayes, BayesNet, decision tree, and random forest classifier for analysis to observe their imbalanced intruder detection behavior and observe the mentioned techniques, shows poor performance. In (Abdulhammed et al., 2018), authors developed a technique for intruder detection systems using class imbalance handling techniques and show that voting, stacking, random forest, and DNN techniques perform well. In another article (Telikani and Gandomi, 2019), authors develop technique based on cost-sensitive learning called cost-sensitive symmetric autoencoder classifier to deal with class imbalance and classification problem in the intruder detection datasets. They show comparison with symmetric autoencoder (SAE) and non-symmetric deep autoencoder (NDAE) and show that CSSAE technique performs better than other.

For class imbalance handling, various techniques have developed so far. Based on the researcher's, class imbalance handling techniques are categorized into two main subsections: Data level techniques and algorithmic techniques. Data level techniques deal with change in the distribution of datasets among classes where algorithmic techniques improve existing algorithms and make them robust to handle class imbalance problem in datasets. To overcome class imbalance, extensive work is done on data resampling techniques. In (Chawla et al., 2002), the researchers proposed a synthetic minority oversampling technique to deal with the imbalanced problem. The proposed is suffered from two primary subproblem random instance selections for creating synthetic instances and sometimes suffers from overfitting problems. To remove short comes of SMOTE, various variants of SMOTE are developed to deal with it. Like in (Sáez et al., 2015), authors develop SMOTE with an iterative partitioning filtering technique to deal with the class imbalance and random noise generation by SMOTE instance. In another article (Batista et al., 2004; Puri and Gupta, 2019), the authors try to combine the SMOTE technique with ENN and TomekLink undersampling technique to overcome the SMOTE problem.

However, some authors try to develop techniques under the category of undersampling. In (Seiffert et al., 2010), authors use random undersampling (RUS) technique, which reduces majority instances at random. This random removal of instances may sometimes result in loss of potential information. So, to remove this abnormality in the RUS technique, many researchers develop different techniques like the edited nearest neighbor technique (ENN) (Alejo et al., 2010), cluster center-based undersampling technique (Lemaitre, 2016–17) used for handling class imbalance using the undersampling approach.

3 Experimental Framework

This section contains detailed descriptions of a dataset, i.e., NSL-KDD datasets, class imbalance handling techniques, classifier used for comparison and evaluation metrics, and experimental design.

3.1 Datasets

NSL-KDD (McHugh, 2000; Tavallaee et al., 2009) is an improved version of KDD'99 datasets. It contains 22 types of attack categories under DOS, Probe, U2R, and R2L where DOS contains Back, Land, Neptune, pod, Smurf, teardrop; Probe category contains Ipsweep, nmap, portsweep, satan; R2L contains ftp_write, guess_password, Imap, Multihop, phf, spy,warezclient,warzmaster; and U2R contains Load_-module, buffer_overflow, rootkit, perl. This dataset contains 41 feature sets where six features are categorical, and the rest are numerical. The detailed category of attacks, along with distribution, is shown in Fig. 1.

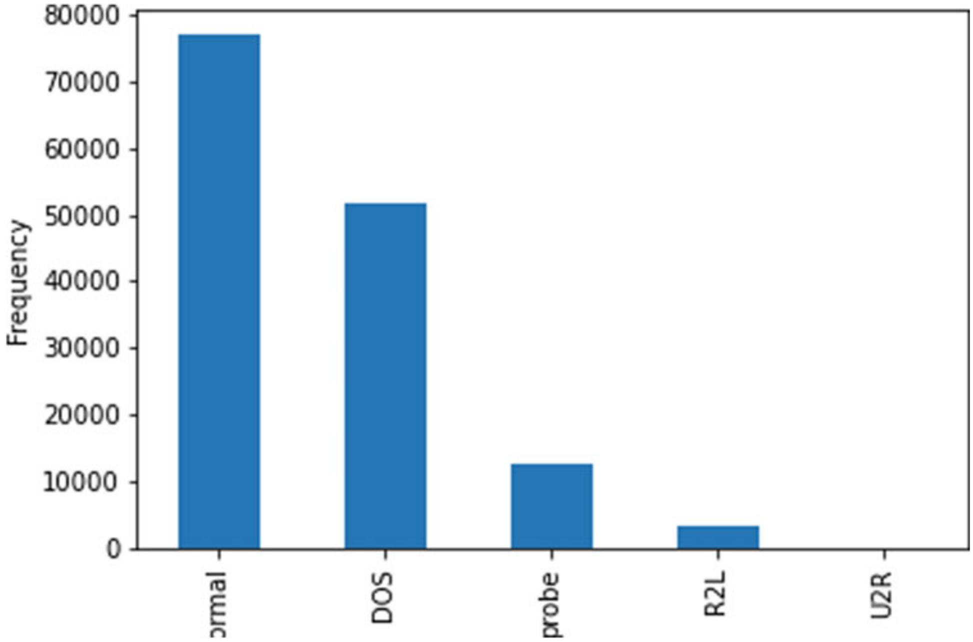

Fig. 1 NSL-KDD dataset description

3.2 Class Imbalance Handling Technique and Classifier

For detail comparative analysis of intruder detection system, we consider SMOTE, SMOTE-ENN, SMOTE-TomekLink, SMOTE-IPF as oversampling techniques RUS, ENN, cluster centroid-based undersampling technique as undersampling technique and also consider decision tree (Safavian and Landgrebe, 1991) as a classifier for classification purpose. The detailed description of the category of class imbalance and techniques used is shown in Fig. 2.

Random Undersampling (RUS) (Seiffert et al., 2009) is used to delete the majority of instances in the datasets randomly. This technique may lose potential information as it selects an instance to be deleted at random. The whole process may lead to underfitting in the classification of instances.

The cluster-based undersampling technique (Cluster) is another approach where a cluster of majority instances having similar behavior are clustered using clustering technique, and undersampling is done on this majority clustered instances so that it will be considered equivalent to minority instances.

Edited Nearest Neighbor (Wilson, 1972) (ENN) technique also developed to deal with borderline and noisy instances in the datasets. This algorithm may delete instances from majority or minority instances to make a clear decision boundary and make datasets balanced.

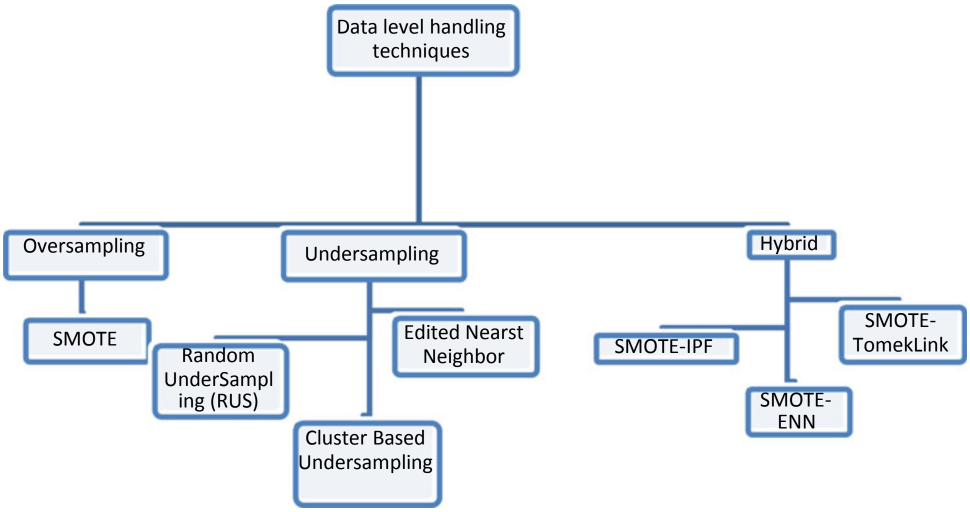

Fig. 2 Overview of data level class imbalance handling techniques

SMOTE (Chawla et al., 2002) belongs to the oversampling techniques category. Working with this technique creates artificial minority sample distribution-wise equivalent to majority instances. The algorithm considers minority instances at random. This algorithm results in noisy instances during artificial instance creation and finally results in a disturbance in decision boundary creation.

SMOTE-ENN (Batista et al., 2004) technique advancement over the oversampling technique falls under the hybrid technique, where SMOTE acts as an oversampling technique, and ENN acts as noisy removal in the overall dataset.

SMOTE-TomekLink (Batista et al., 2004; Puri and Gupta, 2019) technique is another variant of the SMOTE technique that falls under hybrid techniques where SMOTE creates noisy instances are removed using the TomekLink undersampling technique.

SMOTE IPF (Sáez et al., 2015) is known as the synthetic majority oversampling technique combined with the iterative partitioning filtering technique where the iterative partitioning filtering technique is used for noisy removal.

3.3 Performance Metrics

For effective measurement of resampling technique in intruder detection datasets, geometric mean (Fernández et al., 2018) and area under the ROC curve (AUC) (Bekkar et al., 2013) as performance metrics because they are sensitive toward the imbalanced class. Geometric mean (G-Mean) is composed of the accuracy of class raise to the root of m, where m is the number of classes. G-Mean is mathematically represented as follows:

$$\mathrm{G-Mean} = (\text{Accuracy of class})^{1/\mathrm{m}} \tag{1}$$

Where as AUC is also considered as the right metric for imbalanced class datasets.

3.4 Experimental Design

This subsection contains a detailed description of the experimental design and describes the overall formulation of an experiment. For the experiment, we consider Python 3.6 as a simulation tool. The whole process is represented in Fig. 3.

The first dataset derived from public sources is divided into training and test parts. We first combine these training and test datasets into one dataset. The processed dataset contains many problems to deal with, like categorical features, which may reduce classifier efficiency and these categorical features need preprocessing and must be converted into the desired form. After that, all features need to have standard represented to scalar normalization to make it independent of scale.

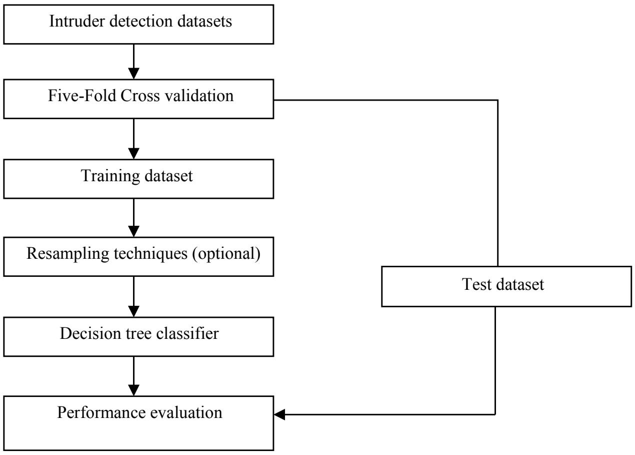
Intruder detection datasets
Five-Fold Cross validation
Training dataset
Resampling techniques (optional)
Decision tree classifier
Performance evaluation
Test dataset

Fig. 3 Experimental flow chart

After all preprocessing, we consider fivefold cross-validation for building a model where five times a given dataset is divided into training and testing datasets—considering training datasets as a learning step for machine learning algorithms and testing dataset act as a test for the same algorithm. Sometimes, to improve the decision boundary of given algorithms, we consider different class imbalance handling techniques as resampling techniques in training datasets because if we consider resampling at training, dataset will not cause any biased nature of our experiment.

The overall performance of fivefold cross-validation is collected using the average score of performance metrics.

4 Results and Discussion

This section contains numerical values as the performance of different resampling techniques on NSL-KDD datasets using different data level class imbalance handling techniques with decision tree classifier. Moreover, overall lesson learned for addressing different research questions mentioned in Sect. 1.

Figure 4 shows the results of different resampling techniques to handle class imbalance using G-Mean on the test dataset of the intruder detection system. The experiment is performed using cross-validation and shows the following results.

- The experiment concluded that applying the cluster-based undersampling technique to deal with the datasets' imbalanced problem is not worthy.

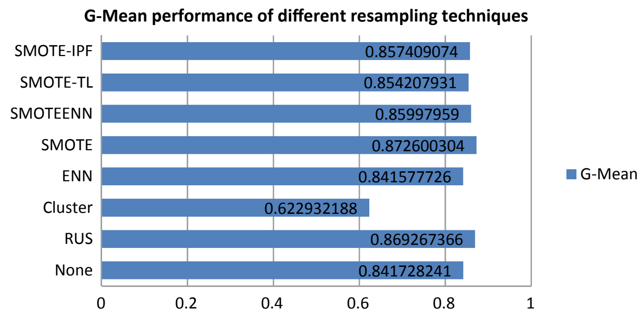

Fig. 4 G-Mean performance of different resampling techniques on test intruder detection datasets

- Further results on comparing ENN with none (without resampling) show similar results; moreover, none ahead of the ENN resampling technique.
- Finally, we also conclude that the SMOTE technique outperforms hybrid techniques like SMOTEENN, SMOTE-TomekLink, and undersampling techniques.
- However, the second-best technique comes from the undersampling technique known as the random undersampling technique.

To further confirm the best resampling technique, we consider AUC as metrics in Fig. 5 and collect the following inferences.

- From the results of this metric, we came up with another side of inferences, where we observe that the hybrid resampling technique like SMOTE ENN and SMOTE-TomekLink and SMOTE techniques show almost similar results.
- Based on observation, we also infer that the cluster-based undersampling technique is the worst choice to deal with the class imbalance in the dataset. Further results confer that none compared with ENN, and RUS shows almost similar results.

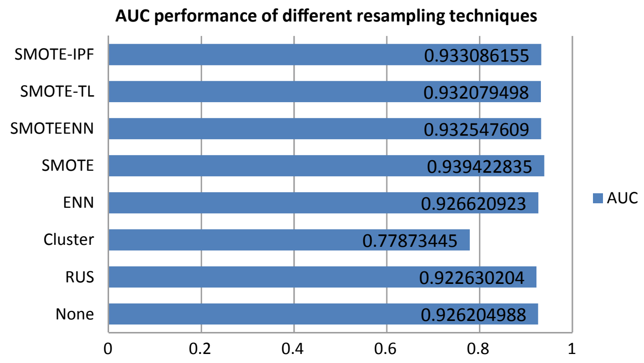

Fig. 5 AUC performance of different resampling techniques on test intruder detection datasets

4.1 Lesson Learned

Based on the experiment performed, we answer the above-mentioned research question.

Q1: Does class imbalance affect the recognization of different attacks using traditional decision tree classifier

Ans. The analysis shown in Figs. 4 and 5 shows that class imbalance creates a significant effect on the recognization of different attacks in intruder detection datasets. From the analysis, we concluded based on the non-effectiveness of the traditional classifier while dealing with a class imbalance in the intruder detection dataset.

Q2: Which existing resampling technique is suitable for handling class imbalance in intrusion detection datasets?

Ans. From experiment either by using AUC or G-Mean, we say that SMOTE with a decision tree is an effective solution to deal with class imbalance problem in intrusion detection datasets.

5 Conclusion and Future Work

This article performs a study on multi-class intruder detection system datasets using class imbalance handling techniques with a decision tree as a classifier. From the study, we conclude that intruder detection datasets suffer from imbalanced class distribution, and some resampling techniques show promising results to cope with class imbalance problem. From experimental results, we also conclude that either by using G-Mean or AUC metrics, it is clear that the SMOTE technique outperformance than rest.

This article shows a portion of the study in context with one classifier. Further, this study may be extended with multiple classifiers and also multiple handling techniques for class imbalance.

References

Abdulhammed, R., Faezipour, M., Abuzneid, A., & AbuMallouh, A. (2018). Deep and machine learning approaches for anomaly-based intrusion detection of imbalanced network traffic. *IEEE Sensors Letters, 3*(1), 1–4.

Aldweesh, A., Derhab, A., & Emam, A. Z. (2020). Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues. *Knowledge-Based Systems, 189,* 105124.

Alejo, R., Sotoca, J. M., Valdovinos, R. M., & Toribio, P. (2010). *Edited nearest neighbor rule for improving neural networks classifications.* Paper presented at the International Symposium on Neural Networks.

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM Sigkdd Explorations Newsletter, 6*(1), 20–29.

Bedi, P., Gupta, N., & Jindal, V. (2020). Siam-IDS: Handling class imbalance problem in intrusion detection systems using siamese neural network. *Procedia Computer Science, 171,* 780–789.

Bekkar, M., Djemaa, H. K., & Alitouche, T. A. (2013). Evaluation measures for models assessment over imbalanced data sets. *Journal of Information Engineering and Application, 3*(10).

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16,* 321–357.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). Performance measures *Learning from Imbalanced Data Sets* (pp. 47–61): Springer.

Lemaitre, G., Nogueira, F., Oliveira, D., Aridas, C. (2016–17). From https://imbalanced-learn.readthedocs.io/en/stable/generated/imblearn.under_sampling.ClusterCentroids.html.

Gonzalez-Cuautle, D., Hernandez-Suarez, A., Sanchez-Perez, G., Toscano-Medina, L. K., Portillo-Portillo, J., Olivares-Mercado, J., & Sandoval-Orozco, A. L. (2020). Synthetic minority oversampling technique for optimizing classification tasks in botnet and intrusion-detection-system datasets. *Applied Sciences, 10*(3), 794.

Intisar, S., Guan, L., & Edirisinghe, E. (2019). *Investigating the Effective Use of Machine Learning Algorithms in Network Intruder Detection Systems.* Paper presented at the Advances in Information and Communication Networks: Proceedings of the 2018 Future of Information and Communication Conference (FICC).

McHugh, J. (2000). Testing intrusion detection systems: A critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory. *ACM Transactions on Information and System Security (TISSEC), 3*(4), 262–294.

Puri, A., & Gupta, M. K. (2019). *Comparative Analysis of Resampling Techniques under Noisy Imbalanced Datasets.* Paper presented at the 2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT).

Rodda, S., & Erothi, U. S. R. (2016). *Class imbalance problem in the network intrusion detection systems.* Paper presented at the 2016 international conference on electrical, electronics, and optimization techniques (ICEEOT).

Sáez, J. A., Luengo, J., Stefanowski, J., & Herrera, F. (2015). SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. *Information Sciences, 291,* 184–203.

Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics, 21*(3), 660–674.

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2009). RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics-Part a: Systems and Humans, 40*(1), 185–197.

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2010). RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics-Part a: Systems and Humans, 40*(1), 185–197.

Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). *A detailed analysis of the KDD CUP 99 data set.* Paper presented at the 2009 IEEE symposium on computational intelligence for security and defense applications.

Telikani, A., & Gandomi, A. H. (2019). Cost-sensitive stacked auto-encoders for intrusion detection in the Internet of Things. *Internet of Things*, 100122.

Wilson, D. L. (1972). Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics* (3), 408–421.

Zhang, H., Huang, L., Wu, C. Q., & Li, Z. (2020). An Effective Convolutional Neural Network Based on SMOTE and Gaussian Mixture Model for Intrusion Detection in Imbalanced Dataset. *Computer Networks*, 107315.