Effective Intrusion Detection Using Data Augmentation with Generative Adversarial Network

Uneneibotejit J. Otokwala – School of Computing, Robert Gordon University, Aberdeen – UK

Introduction

Effective detection of cyber-attacks are always hampered by insufficient data which often leads to overfitting and biasness during classification. While some data augmentation strategies have been used to increment the size of the minority class(es) in an imbalanced dataset; often the structure of the generated data points does not follow the underlying distribution of the original dataset and thus does not improve classification. Using Sort-Augment and Combine (SAC) technique, we used Generative Adversarial Network (GAN) to generate synthetic data values from the class subsets which were then combined with the individual subsets to form the new augmented training dataset and thereafter fitted a machine learning model.

Objectives

The objectives of this research are to:

- To use Sort, Augment and Combine algorithm to perform the data augmentation
- Use Generative Adversarial Network (GAN) to generate synthetic data from the from classes
- Apply Machine Learning classifiers on the augmented training dataset after combining the augmented classes
- Compare the output of the GAN augmented model with output of other models

Contribution

- A data augmentation strategy for class imbalance in datasets that can be used with both binary and multiclass datasets.
- A synthetic data that is of high perceptual quality and that has the same data distribution as the original data.
- To demonstrate the SAC technique's effectiveness using performance metrics such as sensitivity, specificity, and overall accuracy

Original data vs Synthetic data

The plot in Figure 1 shows the comparison of the original data versus the synthetic data.

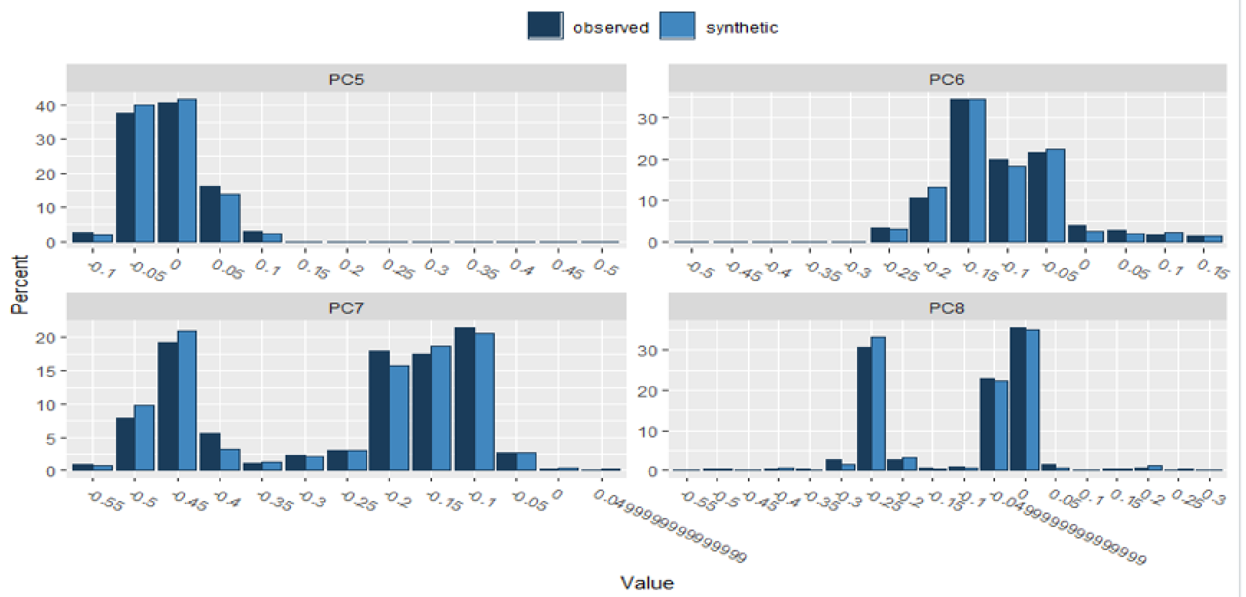

Figure 1. A compare plot of original vs generated data

Algorithm: Sort-Augment-Combine

```
1: Load dataset
2: split dataset into subsets of classLabels (Xi , i+1, n)
3 : repeat
4 : for i ⟵ 1: ncol (Xi , i+1, n) do
5:        Load Xi
6:        apply synthetic function generator to generate (Xi)
7 :       end for
8 : Combine(Xi + Xi)
9 : repeat step 3 : step 8 for classLabels (Xi+1)
10: until newClassLabels are formed
11 : Group (newTraining ⟵ (allClassLabels))
12: Return (newT rainingdataset)
```

Result

Table 1 showing the output Random Forest model with 5 fold cross validation.

		Sensitivity									
	Overall accuracy	Analysis	Backdoor	DoS	Exploit	Fuzzers	Generic	Normal	Reconnaissance	Shellcode	Worm
Original Data	84	6	0	42	67	56	97	97	67	25	9
Augmented Data	89.7	95	94	77	70	76	97	93	89	98	99

Table 2 showing a comparison of out of SAC vs other augmentation approaches

	Overall accuracy	Sensitivity	Specificit y
Original data	97	98	71
ROSE Augmented	94	96	90
SAC Augmented	93	91	94
SAC + ROSE	98	98	98

Conclusion

The augmentation of dataset or the augmentation of the minority class(es) using Sort-Augment-and Combine technique has proved to be very effective in generalization and classification of attacks in multiclass and binary class datasets. The technique has also proved to be effective in combination with ROSE for optimal classification in the binary dataset. This is significant and critical in intrusion detection as the overall accuracy, sensitivity, and specificity has improved.

References

[1] Lata, K., Dave, M., KN, N. (2019, February). Data augmentation using generative adversarial network. In Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE)

[2] Lunardon, N., Menardi, G., & Torelli, N. (2014). ROSE: A Package for Binary Imbalanced Learning. R journal, 6(1)

Acknowledgment

I acknowledge the contributions of
Dr. Andrei Petrovski
Dr. Harsha Kalutarage