Improving intrusion detection using Deep-SVM technique

Nor Azman Mat Ariff[1,*], Muhammad Firdaus Janaludin[1], Mohd Zaki Mas'ud[1], Erman Hamid[1], Nazrulazhar Bahaman[1]

[1]Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia

[*]Corresponding e-mail: nazman@utem.edu.my

Keywords: Intrusion detection; SVM; Chi-Square;

ABSTRACT – Most conventional machine learning techniques have been demonstrated good predictions in classifying normal and intrusion network traffics. However, recently Deep Learning (DL) is the dominating solution, due to it's supremacy in terms of accuracy specifically when involve with large amount of data. Inspired from DL, many researchers introduced conventional algorithms trained in hierarchical fashion. In this paper, we applied Deep Support Vector Machine (Deep-SVM) using probability outputs. This approach using probability outputs from the previous layer as the input feature vectors of the current layer. The classification process is repeated until the stopping criteria is met. The experiment is performed on NSL-KDD dataset. The results show that the Deep-SVM via probability outputs has outperformed the traditional Bag-of-Word (BOW).

1. INTRODUCTION

As the number of connected devices over the Internet are rapidly increasing, it is vital to have an intelligent security mechanism to protect these devices from any attack. Previous researchers have proposed various techniques to prevent such penetration and create a safe environment over the Internet. However, as the technology keep evolving, these attackers also evolved to another level where their threats are more challenging to protect. Intrusion Detection System (IDS) is one of the method to detect intrusion. There are three main types of network analysis for IDS, namely signature-based, anomaly-based, and hybrid [1]. Signature-based detection gathers all known threats so that the threat can be identified in the future. Anomaly-based detection aims to identify unexpected events, which differ from the normal behaviour. Meanwhile, hybrid is combination of both, signature-based and anomaly-based detections. Most of machine learning approaches are considered as hybrid intrusion detection. In [2], J48 and Naïve Bayes classifiers were used in classifying intrusions. Overall, J48 outperformed Naïve Bayes in several evalution metrics. Comprehensive set of machine learning algorithm which included Random Forest, J48, Support Vector Method (SVM), CART and Naïve Bayes on the NSL-KDD intrusion detection dataset were studied in [3]. Nowadays, Deep Learning (DL), based on Deep Belief Network (DBN), is widely used approach in various fields. In [4], the researchers explored the capabilities of DL in intrusion detections through a series of experiments with the best accuracy achieved was 97.5%. Inspired by DL, [5] used Deep-SVM via kernel activations and train the training set into several layers. In this paper, Deep-SVM via probability outputs is proposed in identifying normal and intrusion traffics.

2. METHODOLOGY

In this section, we will describe our methodology, which involves two phases. Figure 1 shows the process flow of our methodology. First phase involves classification of Bag-of-Word (BOW) features. Since feature selection aims at finding a useful feature subset for classification and always give better performance, BOW that have been applied Chi-Square feature selection (CS) is included. In this phase, classification performance of BOW and CS are measured as the baseline accuracy. Two SVM kernels, namely Linear and RBF are deployed as these two kernels are widely used and efficient in classifying linear and non-linear separable problems. In this study, SVM classification will be fine-tuned to produce the class probability outputs.

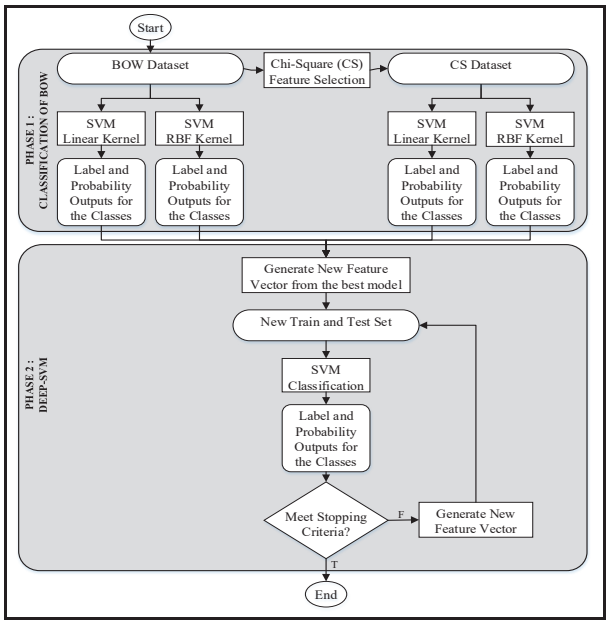

Figure 1 Process Flow of the proposed methodology

In second phase, Deep-SVM via class probability outputs approach is deployed to improve the classification performance in phase 1. We also extend this Deep-SVM approach by implementing Naïve Deep-SVM. The probability outputs from the best model in phase 1 will be used as a new train and test set for

Deep-SVM approach.This new dataset is considered as layer 1 dataset. Classification performance for this dataset is measured and if the performance result is not meet the stopping criteria, this process will be repeated. The probability outputs from layer 1 will be a train and test set for layer 2.

3. RESULTS AND DISCUSSION

The dataset used in this study is NSL-KDD dataset, which can be obtained from https://www.unb.ca/cic/datasets/nsl.html. The dataset contains 125,973 network traffics with 41 attributes where 67,343 are normal traffics and 58,630 are intrusion traffics. The evaluation using 100-repeated train-test split procedure where 1,000 instances are randomly selected for train set and test set. Table 1 shows the average classification accuracy (%) for BOW and CS models using linear and RBF kernels. Classification using RBF kernel outperformed linear kernel for both, BOW and CS. The best performance is 97.79% for BOW with RBF kernel.

Table 1 The average classification accuracy (%) of BOW and CS

Kernel	BOW	CS
Linear	95.35	95.35
RBF	**97.79**	97.77

Table 2 shows classification performances for our proposed methods, Deep-SVM and naïve Deep-SVM. RBF kernel was chosen as the SVM learning algorithm for Deep-SVM and naïve Deep-SVM since it gave the best classification performance for earlier experiments, BOW and CS. Both approaches outperformed BOW with RBF kernel, where Deep-SVM achieved the best performance with 97.95% and naïve Deep-SVM with 97.85%.

Table 2 The average classification accuracy (%) of D-SVM and Naïve D-SVM

Approach	Accuracy
Deep-SVM	**97.95**
Naïve Deep-SVM	97.85

Figure 2 provides the classification results of Deep-SVM for each layer. It is clearly shows that at layer 0 to the layer 12, the classification performances are slightly drop. However, the classification performances gradually increase from layer 13 to the upper layer and achieve the best performance at layer 98 and 99 with 97.95%. We stop the learning after layer 100 gave a lower accuracy with 97.94%. Naïve Deep-SVM gave the best classification performance at layer 2 with 97.85% and after that gradually decrease its accuracy until the learning stopped at layer 100 with 97.75%.

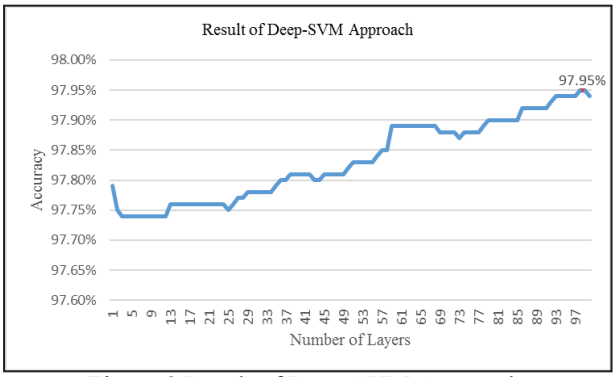

Figure 2 Result of Deep-SVM Approach

4. CONCLUSIONS

This research, applied a Deep-SVM via probability outputs approach to intrusion detections. The classification result shows that the approach can improve the classification performance of traditional BOW.

ACKNOWLEDGEMENT

Authors are grateful to InforsNet Research Group of Universiti Teknikal Malaysia Melaka (UTeM) for the support and special acknowledgement to Ministry of Education Malaysia for providing financial support through the Fundamental Research Grant Scheme (FRGS/2018/FTMK-CACT/F00391).

REFERENCES

[1] Y. Xin, L. Kong, Z. Liu, Y. Chen, and Y. Li, Machine Learning and Deep Learning Methods for Cybersecurity, *IEEE Access*, vol. 6, pp. 35365–35381, 2018.

[2] G. M. and R. R. Choudhary, A Review Paper on IDS Classification using KDD 99 and NSL KDD Dataset in WEKA, in 2017 International Conference on Computer, Communications and Electronics (Comptelix), 2017, pp. 553–558.

[3] S. Revathi and A. Malathi, A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection, vol. 2, no. 12, pp. 1848–1853, 2013.

[4] C. Yin, Y. Zhu, J. Fei, and X. He, A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks, *IEEE Access*, vol. 5, pp. 21954–21961, 2017.

[5] A. Abdullah, R. C. Veltkamp, and M. a. Wiering, An Ensemble of Deep Support Vector Machines for Image Categorization, 2009 Int. Conf. Soft Comput. Pattern Recognit., no. 2, pp. 301–306, 2009.