Weakly Supervised Learning for Cybersecurity Attack From Only Unlabeled Data

Feng Lin
School of Information Technology and Electrical Engineering
The University of Queensland, Qld., 4072, Australia

Abstract

Cyber security problems consistently existed, and it is necessary to enhance the cyber protection system [1]. Currently, there are many machine learning protection systems that possess reliable performances. However, most of the machine learning models are supervised learning, and labelling training data would be a considerable burden to the experts [2]. Theoretically, weakly supervised learning is likely to be a good way to solve this problem [3]. In this paper, several cyber attack datasets with different characteristics are selected, and Unlabeled-Unlabeled (UU) model will be mainly investigated in this paper with the input of cyber security datasets. Empirical risk minimization (ERM) is an effective evaluation of how the model output performed [4]. UU method has the merit that it has the capability to make binary classification without any label in the dataset. Also, imbalance is another common feature that exists in the cyber security dataset [5], so the imbalanced dataset adjustment will also be investigated in this paper.

1. Introduction

Nowadays, all the industries tend to be intelligent, so the internet combinations are indispensable [6]. Thus, many important systems, for instance, the internet of things (IoT), bank transactions, and server protection system, must be protected safely [7]. Machine learning is one of the effective ways to reinforce the protection system, and it has credible learning results on cyber security data [8]. But there are some issues that are difficult to be solved. Cyber technologies have increased rapidly because the hardware infrastructures have been evolving to mature, and cyber attacks are continuously evaluated [6]. Therefore, in terms of enhancing the cyber protection system, the exponentially growing cyber security data collections are collected, and it is inevitable to make a great effort to label all the data [1]. Some novel methods, such as active-learning, and self-supervised learning, performed well on labelling data, but they still need expert to label them [9].

ERM has been proved that it can be appropriately used in weakly supervised learning based on rigorous math derivation, and many weakly supervised learning methods have been invented, such as Positive-Unlabeled (PU) learning model, UU learning model and etc [3]. In terms of the previous research, weakly-supervised learning methods are invented that the model needs less training information but has a reliable performance. ERM can be used as the output loss estimation because it has a good convergence during the iterations in the model [10]. UU learning is one of the typical weakly-supervised learning that it can do binary classification without any label in the training process, and it has remarkably reliable performance on the benchmark dataset, such as MNIST [3]. It only requires an estimation of the prior of the original dataset and makes 2 subsets from the original dataset with the estimated priors. Based on the previous research on benchmark datasets, the more differences between 2 training sets, the better training result occurs [3]. In terms of UU training process doesn't need any label, which means the output layer doesn't need ground truth labels to do the comparison, so the output needs to apply a classification boundary to distinguish the classification results. For example, a threshold of 0 could be set to classify the output result when testing the trained model. Detailed cases of this output classifier will be investigated in the experiment.

Generally, the majority of cyber security attack datasets are imbalanced because the cyber attacks often appear rarely and consistently [11]. Thus, imbalanced data will cause many training defects, such as the biased classification to the majority clusters and good results on testing errors but worse performance on minority data recognition [12]. UU learning has some unique drawbacks in terms of the mechanism: Rated training subsets need adequate data on both sides, which has been proved in the experiment in this paper. Several solutions will be discussed in this paper to alleviate

the issue caused by the imbalance: Oversampling, under-sampling and re-weight the training class (Modify UU risk function). Oversampling and under-sampling are two most common ways to figure out the imbalance issue in machine learning. While referring to the 'No free lunch' rule, either method has its own advantage and disadvantage [12]. Comparisons between these methods will be investigated in the experiment in this paper.

2. UU learning method

2.1. PN problem assumption

Now suppose there is a binary classification problem: a big dataset contains positive and negative data, and the prior of this dataset is π. Consider making two subsets from the big dataset, and 2 subsets have priors of θ and $\theta^{'}$. Intuitively, the density distributions of 3 datasets are the same because all the data are from the big dataset, which is equation

$$p_{tr}(x|y) = p_{tr}^{'}(x|y) = p(x|y)$$

Normally, loss function in the output layer assess the difference between the ground truth and the training output value, and risk demonstrates the expectations of the loss, while ERM estimates the minimized mean loss in the output, which can be written as

$$R_{emp} = min\frac{1}{n}\sum_{i=1}^{n} L(y_i, g(x_i)) \tag{1}$$

g is a binary classifier, $g(x_i)$ represent the model output and y_i indicate the ground truth value. Loss function have to be surrogate loss in UU [3]. Then the PN risk function can be written as

$$\widehat{R}_{pn}(g) = \frac{\pi_p}{n}\sum_{i=1}^{n} L(g(x_i)) + (\frac{1-\pi_p}{n^{'}})\sum_{j=1}^{n^{'}} L(-g(x_j^{'})) \tag{2}$$

while the x_i and x_j indicates the positive and negative joint density.

2.2. UU learning risk function re-write

Now assume exist a unlabeled big dataset which has prior π_p, and then building a training subset from the big dataset which has approximate prior θ. The risk function of this subset is

$$R(g) = E_{p_{tr}}[\bar{L}(g(X_{tr})] \tag{3}$$

where $\bar{L}(g(X_{tr}) = aL(g(x)) + bL(-g(x))$ [13]. This function with single training subset is not possible to do the risk estimation because θ is free [3]. Therefore, another subset from the big dataset with prior of $\theta_{'}$ will be added to do risk estimation, the risk function is shwon as

$$R(g) = E_{p_{tr}}[\bar{L}_+(g(X_{tr})] + E_{p_{tr}^{'}}[\bar{L}_-(-g(X_{tr})] \tag{4}$$

where loss functions $\bar{L}_+ = aL(g(x)) + bL(-g(x))$ and $\bar{L}_- = cL(g(x)) + dL(-g(x))$. The derivation has been proved that [3]

$$a = \frac{(1-\theta^{'})\pi_p}{\theta - \theta^{'}},\ b = -\frac{(1-\pi_p)\theta^{'}}{\theta - \theta^{'}},$$
$$c = \frac{(1-\pi_p)\theta}{\theta - \theta^{'}},\ d = -\frac{(1-\theta)\pi_p}{\theta - \theta^{'}}$$

and let the coefficient into the risk function, then the final risk function of UU can be written as

$$\begin{aligned}\widehat{R}_{uu}(g) = \frac{1}{n}\sum_{i=1}^{n}(&\frac{(1-\theta^{'})\pi_p}{\theta - \theta^{'}}L(g(x_i)) \\ &- \frac{(1-\pi_p)\theta^{'}}{\theta - \theta^{'}}L(-g(x_i))) \\ &+ \frac{1}{n^{'}}\sum_{i=1}^{n^{'}}(\frac{(1-\pi_p)\theta}{\theta - \theta^{'}}L(g(x_j^{'})) \\ &- \frac{(1-\theta)\pi_p}{\theta - \theta^{'}}L(-g(x_j^{'})))\end{aligned} \tag{5}$$

2.3. UU learning risk function re-weight for imbalanced input

Re-weight the risk function from imbalanced input in PU learning has obviously better performance [14]. Assume there is a balanced big dataset that the prior $\pi_p^{'}$ is 0.5, and the joint density of the balanced $P_{balance(x,y)}$ and imbalance data $P(x,y)$ are the same, then the risk of the balanced data is

$$\begin{aligned}R_{balance}(g) &= E_{P_{balance(x,y)}}[L(Yg(X))] \\ &= \pi_p^{'}E_{P_{balance(x|y=1)}}[L(g(X))] \\ &+ (1-\pi_p^{'})E_{P_{balance(x|y=-1)}}[L(-g(X))] \\ &= \pi_p^{'}E_{P_{(x|y=1)}}[L(g(X))] \\ &+ (1-\pi_p^{'})E_{P_{(x|y=-1)}}[L(-g(X))]\end{aligned} \tag{6}$$

Based on the previous $R_{uu}(g)$ derivation, the re-weighted risk function only modify the π_p to $\pi_p^{'}$ in $R_{uu}(g)$.

3. Experiment

3.1. Dataset

There are many different types of cyber attacks, such as Distributed Denial of Service Attacks (DDOS), SQL injection attacks, Vulnerability Exploitation, etc [7]. In terms of UU learning only make binary classification, so that all types of attacks in the dataset have to be converted into one category. In the experiment, four cyber-attack datasets are selected to establish investigations, including two balanced datasets and two imbalanced datasets. Also, two dataset contains mixed attacks and datasets contains single attack type.

- Kddcup99 dataset is a mixed attacks dataset which has 41 features, and it covers 396743 attack samples, and 97277 normal access samples [15].
- UNSW-NB15 dataset is a mixed attacks dataset which has 49 features, and it covers 22215 attack samples, and 677786 normal access samples [16].
- CIC-IDS2017 contains one DDOS attacks dataset, which has 78 features, and it covers 128027 attack samples and 97718 normal access samples.
- CIC-IDS2017 contains one 3 SQL attacks dataset, which has 78 features, and it covers 2180 attack samples and 168186 normal access samples.

CICds-2017 can be found in the URL: `https://www.unb.ca/cic/datasets/ids-2017.html`

3.2. Model

Based on the previous research, the MNIST dataset has been tested by the UU risk function on a Fully Connected model with five depth, and the result illustrates UU learning has outstanding performance in doing binary classification [3]. Hence, in this paper, five depth FC MLP, 512 batch size, RELU activation function, sigmoid loss function and SGD optimizer will be selected in the UU model investigations. Referring to the reasonable training set scale, the Kddcup99 dataset and UNSW-NB15 dataset select 150000 as the size of one subset, and two datasets in CIC-IDS2017 are selected 60000 as one subset size.

In terms of making comparisons, a same structure supervised learning FC MLP model with sigmoid loss function will be used in the experiment, and training sizes are 300000 and 120000, corresponding to the dataset in UU.

3.3. Experiment process

In the experiment, base case of UU model are: Batchsize = 512, $\theta \& \theta^{'}$ = 0.9&0.1, loss function = sigmoid loss, optimizer = SGD.

Imbalance cyber dataset investigation The imbalanced cyber attack dataset problem will be investigated by applying several diverse solutions, including:

- Over-sampling
- Under-sampling
- Re-weight the minority cluster in the risk function

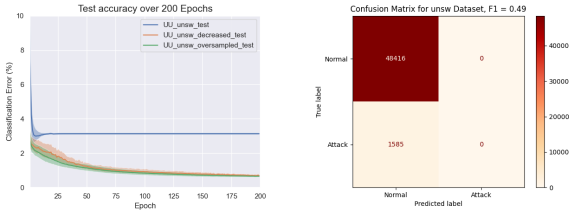
In terms of the UU mechanism, setting up 2 rated training subsets requires sufficient samples. While the imbalance cyber attack data often contains inadequate attack samples. Thus, before doing the Re-weight method, oversampling need to be applied to make the training sets.

Then the best performance solution on the Imbalanced dataset will be compared with the normal FC MLP model result.

UU learning in parameters investigation Two balanced datasets will be investigated in this section. Experiments consist of the diverse value of UU parameters in the UU model.

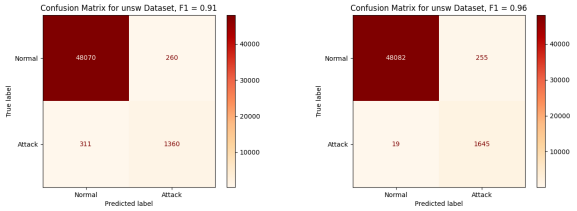
- Comparing with normal MLP
- Convert surrogate losses
- Priors variation

3.4. Result analysis



(a) (b)

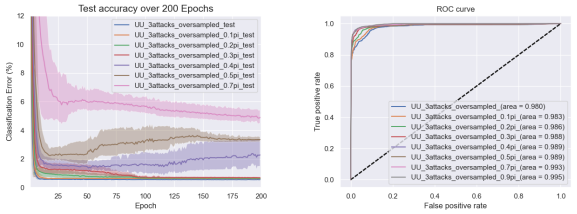
Figure 1. Comparison between classification errors after training UU model with different adjusted UNSW dataset, including confusion matrices



(a) (b)

Figure 2. Over-sampled and under-sampled method confusion matrix results

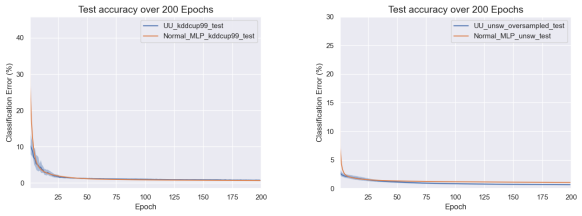
Imbalanced dataset with insufficient input has a terrible effect on UU model, while after decrease the input size or oversampling the training set fix this issue.



(a) (b)

Figure 3. over-sampled input subsets risk function Re-weight on 3 SQL attack dataset.

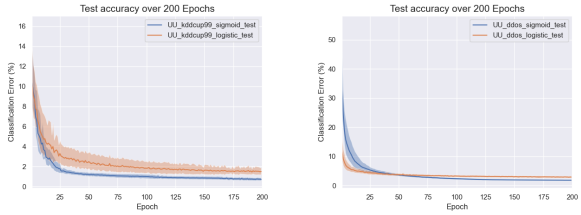
Re-weight the risk function doesn't result in a expected performance. Figure 3 shows when prior increase from 0.01 to 0.7, the result keep getting worse when testing on 3 SQL attack dataset. Whatever over-sample the data from the big dataset or scale the data when creating 2 training sets, the result are all not better than the result that only apply over-sample method.



(a) (b)

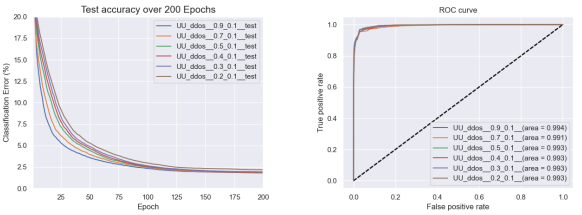
Figure 4. Classification error between normal MLP and UU learning model.

Figure 4 indicates that UU learning and normal MLP has the same level ability when classify both DDOS and Kddcup99 datasets, and Figure 5 demonstrate that sigmoid loss has a slightly better performance than 0-1 loss, which is same as the previous research [].



(a) (b)

Figure 5. Sigmoid loss and 0-1 loss results on Kddcup99 dataset.



(a) (b)

Figure 6. Close one training set prior to another on DDOS dataset.

In Figure 6, reducing the high training set prior from 0.9 to 0.2, there are very few variation between each result. When 2 training priors are equals to 0.1, the final recognition error raised about 0.2 %, which is extremely subtle. Based on the rated training set size, 0.1 covers 15000 positive attack samples, which is sufficient to make the model familiar with the attack data features. Also it might caused by the data characteristics, such as the diversity between attack samples and normal access samples are distinctive.

4. Conclusion

Overall, UU learning has a reliable performance on distinguishing cyber attacks without input labels. Imbalanced dataset can efficiently be solved by oversampling the minority cluster, while re-weight method on imbalanced UU training process is not effective. UU learning model has the similar level binary classification capability with the normal MLP model on cyber security datasets, which indicates UU learning is possible to avoid the enormous labelling cost on binary classification in non-sequential cyber security dataset. In UU model intrinsic aspect, sigmoid loss function has better result than the 0-1 loss. Also, approaching one

training subset prior to another prior cause a slight defect because the fundamental training set contains abundant samples even it takes up 10% proportion.

References

[1] D. Kamenov, "Intelligent methods for big data analytics and cyber security," *Information & Security*, vol. 39, no. 1, pp. 255–262, 2018.

[2] L. F. Sikos and K.-K. R. Choo, *Data science in cybersecurity and cyberthreat intelligence*. Springer, 2020.

[3] N. Lu, G. Niu, A. K. Menon, and M. Sugiyama, "On the minimal supervision for training any binary classifier from only unlabeled data," *arXiv preprint arXiv:1808.10585*, 2018.

[4] S. R. Kulkarni and G. Harman, "Statistical learning theory: a tutorial," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 3, no. 6, pp. 543–556, 2011.

[5] X. A. Larriva-Novo, M. Vega-Barbas, V. A. Villagrá, and M. S. Rodrigo, "Evaluation of cybersecurity data set characteristics for their applicability to neural networks algorithms detecting cybersecurity anomalies," *IEEE Access*, vol. 8, pp. 9005–9014, 2020.

[6] C. Sandvig, "The internet as infrastructure," in *The Oxford handbook of internet studies*, 2013.

[7] B. Von Solms, "Corporate governance and information security," *Computers & Security*, vol. 20, no. 3, pp. 215–218, 2001.

[8] C. Aravindan, T. Frederick, V. Hemamalini, and M. V. J. Cathirine, "An extensive research on cyber threats using learning algorithm," in *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, pp. 1–8, IEEE, 2020.

[9] J. Z. Bengar, J. van de Weijer, B. Twardowski, and B. Raducanu, "Reducing label effort: Self-supervised meets active learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1631–1639, 2021.

[10] M. C. Du Plessis, G. Niu, and M. Sugiyama, "Clustering unclustered data: Unsupervised binary labeling of two datasets having different class balances," in *2013 Conference on Technologies and Applications of Artificial Intelligence*, pp. 1–6, IEEE, 2013.

[11] M. Zamani and M. Movahedi, "Machine learning techniques for intrusion detection," 2013.

[12] D. Ramyachitra and P. Manikandan, "Imbalanced dataset classification and solutions: a review," *International Journal of Computing and Business Research (IJCBR)*, vol. 5, no. 4, pp. 1–29, 2014.

[13] N. Quadrianto, A. J. Smola, T. S. Caetano, and Q. V. Le, "Estimating labels from label proportions.," *Journal of Machine Learning Research*, vol. 10, no. 10, 2009.

[14] G. Su, W. Chen, and M. Xu, "Positive-unlabeled learning from imbalanced data," in *Proceedings of the 30th International Joint Conference on Artificial Intelligence, Virtual Event*, 2021.

[15] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the kdd cup 99 data set," in *2009 IEEE symposium on computational intelligence for security and defense applications*, pp. 1–6, Ieee, 2009.

[16] N. Moustafa and J. Slay, "Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set)," in *2015 military communications and information systems conference (MilCIS)*, pp. 1–6, IEEE, 2015.