Poster: MORPH: Towards Automated Concept Drift Adaptation for Malware Detection

Md Tanvirul Alam*, Romy Fieblinger†*, Ashim Mahara*, and Nidhi Rastogi*
*Rochester Institute of Technology, Rochester NY 14623, USA
†University of Applied Sciences, 09648 Mittweida, Germany
{ma8235, rf7344, am7539, nidhi.rastogi}@rit.edu

***Abstract*—Malware evolution is a significant challenge since it can outsmart static detection models. This concept drift problem can cause performance degradation over time, leaving the system vulnerable to malware attacks. Existing research has primarily relied on selecting representative samples to update the model using active learning, however, self-training has recently emerged as a promising approach. In this poster abstract, we propose MORPH– an effective concept drift adaptation method that uses "pseudo-labels," specifically designed for neural-network-based malware detection models to improve the robustness of malware detection.**

I. Introduction

Malware threats continue to evolve and pose a critical challenge to cyber defenses, requiring adaptable solutions for automated analysis and identification of malicious patterns. However, the effectiveness of these approaches depends on the statistical similarities between the trained defense models and the real-world malware data. Concept drift refers to a shift in the underlying distribution of the test dataset that deviates from the train dataset (existing defense model), deteriorating model performance [3]. Active learning has recently shown promise for addressing this problem by choosing a subset of samples that are identified to be annotated by experts [3], [2]. However, frequent model updates with high-quality annotation can be costly. As a result, another orthogonal approach is to adapt the model using weak supervision or self-training. Notable works in this direction for malware detection are DroidEvolver [6] and DroidEvolver++[4]. DroidEvolver maintains a pool of five classification models and utilizes the ensemble prediction as the pseudo-label to identify aging models that deviate from the ensemble. However, the ensemble of linear models is ineffective at mitigating concept drift due to self-poisoning [4]. To overcome this limitation, we present **MORPH** (auto**M**ated c**O**ncept d**R**ift ada**P**tation algorit**H**m), a self-training model for concept drift adaption in malware detection. We address these **research questions (RQs)**:

RQ1:*Can pseudo-labeling enable automatic concept drift adaptation in neural network-based malware detection?*

RQ2:*Can pseudo-label-based adaptation reduce the frequency of annotation required in active learning?*

RQ3:*How does our proposed approach compare to prior automated concept drift adaptation methods?*

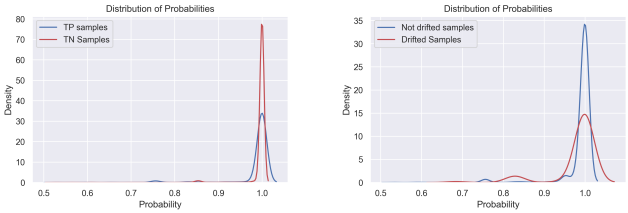


Fig. 1: Kernel Density Estimate plot for probability distribution on AndroZoo dataset for (left) TP and TN sample and (right) drifted vs not-drifted malware samples.

II. Proposed Methodology

The unique challenge of the concept drift approach in malware classification is its bias towards false negatives, where malware instances are erroneously classified as benign. As a result, when a model predicts a new sample as malware, the likelihood of being correct is higher due to the relatively low occurrence of false positives. So, even predictions with low confidence tend to be accurate. Conversely, low-confidence predictions may be incorrect in the case of benign predictions. This phenomenon is illustrated on the left half of Figure 1, where the malware predictions (true positives) have a broader probability distribution compared to benign predictions (true negatives). We propose a targeted pseudo-label strategy to build on this insight, as outlined in Algorithm 1. After the sample detector identifies relevant instances, we leverage semi-supervised learning to retrain the model by combining ground truth and pseudo-labeled samples. To achieve this, we form mini-batches comprising an equal number of original ground truth and pseudo-labeled data.

Additionally, we also combine MORPH with active learning model updates to reduce the frequency of annotation requirements. Self-training is useful in gradual domain adaptation scenarios, especially where the data distribution undergoes gradual changes over time (like in malware data) [5]. This phenomenon is effectively represented in Figure 2. We expect active learning to help the model adjust to severe shifts by including them in the training process while self-training

Network and Distributed System Security (NDSS) Symposium 2024
26 February - 1 March 2024, San Diego, CA, USA
ISBN 1-891562-93-2
https://dx.doi.org/10.14722/ndss.2024.23xxx
www.ndss-symposium.org

Algorithm 1: MORPH: Pseudo-labeling Algorithm

Input : Test samples
Output: Pseudo-labeled malware and benign samples

Calculate model prediction on test samples
$D_M \leftarrow$ samples predicted as malware
$D_B \leftarrow$ samples predicted as benign
Select N_M samples randomly from D_M with probability $> \tau_m$
Select top N_M samples with highest confidence from D_B with probability $> \tau_b$
Return the pseudo-labeled malware and benign samples selected

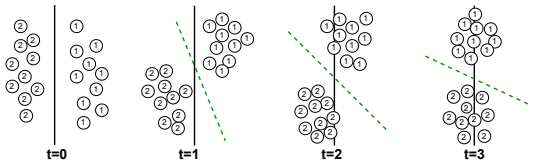


Fig. 2: Gradual adaptation to distribution shift for a binary classifier. Original model (solid line) is unable to classify samples in the target domain (time step t=3), but the model updated with pseudo labels (dotted line) can adjust its decision boundary.

adjusts to more gradual shifts afterward by leveraging the unlabeled data.

III. Experiments, Results and Conclusion

1) Datasets: In our experiments, we use two different datasets – AndroZoo [2] and Ember [1]. We found optimal hyperparameters using the AndroZoo dataset and used the same for Ember. The optimal value for $\tau_m = 0.6$. We can omit τ_b since we balance the number of benign and malicious samples, and the datasets have a lot more samples predicted as benign.

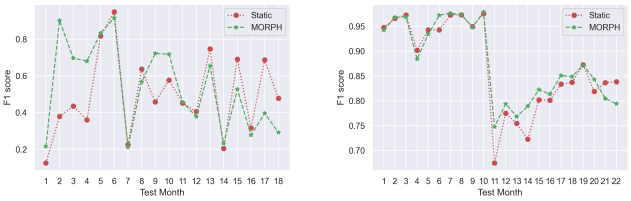


Fig. 3: F1 score on AndroZoo (left) and Ember (right) datasets for test months with MORPH and baseline neural network.

2) Utility of Pseudo Labels (RQ1): As demonstrated in Figure 3 MORPH can improve over baseline neural network, without being affected by self-poisioning. The adapted model can recover following a concept drift, which indicates that pseudo labels generated by neural networks can provide sufficient information for concept drift adaptation for malware detection.

3) Combination with Active Learning (RQ2): We show results in Figure 4 when we combine MORPH updates on unlabeled with active learning updates on a subset of labeled data. MORPH consistently enhances performance by leveraging the unlabeled samples instead of only performing active learning updates. When using a 100 annotation budget for active learning, MORPH improves the F1 score of the SoTA active learning method on the AndroZoo dataset by 3.46%. This signifies that MORPH can reduce the number of active learning updates required to maintain stable model performance.

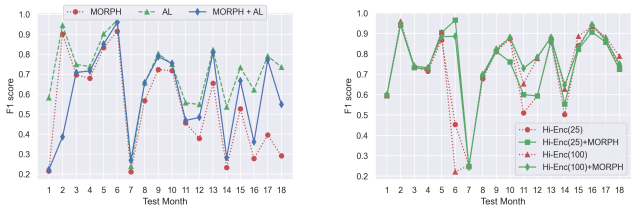


Fig. 4: F1 score on AndroZoo datasets for active learning with baseline neural network (left) and Hi-Enc (Hierarchical Contrastive Learning) [2] right.

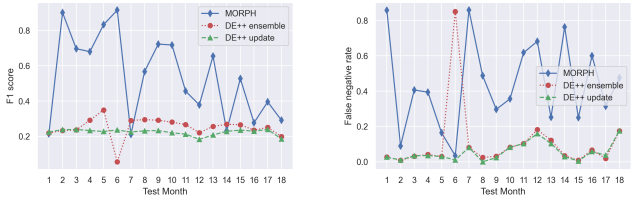


Fig. 5: F1 score (left) and FNR (middle) for DE++ ensemble model (with and without model updates), and MORPH.

4) Comparison With DroidEvolver++ (RQ3): We show our experimental results with the DE++ model update using ensemble-based pseudo labels in Figure 5. The result suggests that neural networks are more robust to concept drift than linear online learning models or an ensemble of such models.

IV. Conclusion

In this poster summary, we propose a novel self-training approach utilizing pseudo-labels on a neural network-based malware classifier, improving robustness against concept drift. Our method outperforms existing works and shows promising results when combined with active learning.

References

[1] H. S. Anderson and P. Roth, "EMBER: An Open Dataset for Training Static PE Malware Machine Learning Models," *ArXiv e-prints*, Apr. 2018.

[2] Y. Chen, Z. Ding, and D. A. Wagner, "Continuous learning for android malware detection," in *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*, 2023.

[3] R. Jordaney, K. Sharad, S. K. Dash, Z. Wang, D. Papini, I. Nouretdinov, and L. Cavallaro, "Transcend: Detecting concept drift in malware classification models," in *26th USENIX security symposium (USENIX security 17)*, 2017.

[4] Z. Kan, F. Pendlebury, F. Pierazzi, and L. Cavallaro, "Investigating labelless drift adaptation for malware detection," in *Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security*, 2021.

[5] A. Kumar, T. Ma, and P. Liang, "Understanding self-training for gradual domain adaptation," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5468–5479.

[6] K. Xu, Y. Li, R. Deng, K. Chen, and J. Xu, "Droidevolver: Self-evolving android malware detection system," in *2019 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2019, pp. 47–62.

MORPH: Towards Automated Concept Drift Adaptation for Malware Detection

Md Tanvirul Alam[1], Romy Fieblinger[12], Ashim Mahara[1], and Nidhi Rastogi[1]

[1]Rochester Institute of Technology, Rochester NY 14623, USA

[2]University of Applied Sciences, 09648 Mittweida , Germany

Contributions

- **Pseudo-label based concept drift** adaptation method for malware detection
- I**mprove over active learning** baseline for concept drift adaptation
- Outperforms prior work in Android malware detection

Motivations

Utility of self-training: Useful in gradual domain adaptation, especially where the data distribution undergoes gradual changes over time, such as malware data [1].

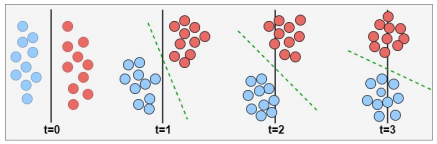


Figure 1. The original model (solid line) is unable to classify samples in the target domain (time step t=3), but the model updated with pseudo labels (dotted line) can adjust its decision boundary.

Unique characteristics of malware data: Concept drift contributes more to false negatives (malware predicted as benign), as a result we can treat benign and malware predictions differently when selecting pseudo-labeled samples.

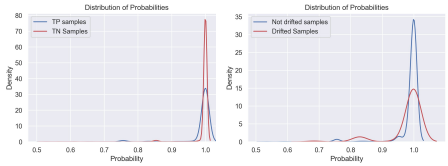
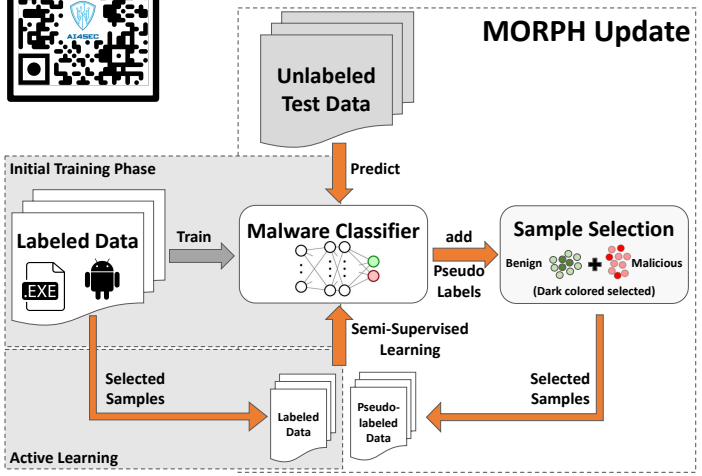


Figure 2. Kernel Density Estimate plot for probability distribution on Android malware dataset for (left) True Positive and True Negative sample and (right) drifted vs not-drifted malware samples.

Proposed Method: MORPH



Key steps in the model retraining phase:

1. Get model predictions on the unlabeled data
2. Select pseudo-labeled samples using Algorithm 1
3. Retrain the model using semi-supervised learning
4. For active learning, steps 1-3 are performed on the remaining unlabeled data after annotated samples are included in the training data

Sample Selection Algorithm

Algorithm 1: MORPH: Pseudo-labeling Algorithm

Input : Test samples
Output: Pseudo-labeled malware and benign samples

Calculate model prediction on test samples
$D_M \leftarrow$ samples predicted as malware
$D_B \leftarrow$ samples predicted as benign
Select N_M samples randomly from D_M with probability $> \tau_m$
Select top N_M samples with highest confidence from D_B with probability $> \tau_b$
Return the pseudo-labeled malware and benign samples selected

Datasets

- **Android – AndroZoo [2] : Train 12 months, Validation 6 months, Test 18 months**
- **Windows – Ember [3]: Train 1 month, Validation 1 month, Test 22 months**

Results

MORPH can improve over baseline neural network, without being affected by self-poisoning.

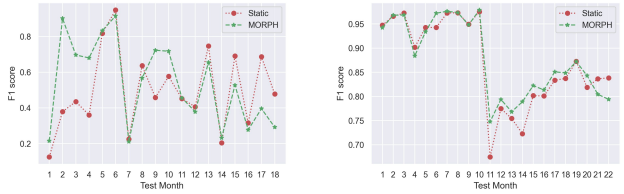


Figure 3. F1 score on AndroZoo (left) [2] and Ember (right) [3] datasets for test months with MORPH and baseline neural network.

MORPH consistently enhances performance by leveraging the unlabeled samples instead of only performing active learning updates. With 100 annotation budget, MORPH improves the F1 score of the SoTA active learning method by 3.46%.

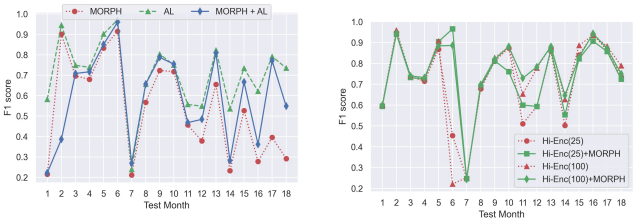


Figure 4. F1 score on AndroZoo datasets for active learning with baseline neural network (left) and Hi-Enc (Hierarchical Contrastive Learning [2] right.

Neural networks are more robust to concept drift than linear online learning models or an ensemble of such models and our proposed method is less susceptible to self-poisoning.

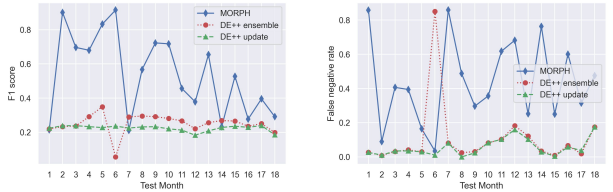


Figure 5. F1 score (left) and FNR (middle) for DroidEvolver++ [4] ensemble model (with and without model updates), and MORPH.

References

[1] A. Kumar, T. Ma, and P. Liang, "Understanding self-training for gradual domain adaptation," in International Conference on Machine Learning. PMLR, 2020, pp. 5468–5479

[2] Y. Chen, Z. Ding, and D. A. Wagner, "Continuous learning for android malware detection," in 32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023.

[3] H. S. Anderson and P. Roth, "EMBER: An Open Dataset for Training Static PE Malware Machine Learning Models," ArXiv e-prints.

[4] Z. Kan, F. Pendlebury, F. Pierazzi, and L. Cavallaro, "Investigating labelless drift adaptation for malware detection," in Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security, 2021