Transcending TRANSCEND: Revisiting Malware Classification in the Presence of Concept Drift - Summary - SidharthAnil

Presentation

The paper deals with revisiting and formalizing the paper 'Transcend: Detecting Concept Drift in Malware Classification Models'. The presentation starts off by explaining the different ways in which a data shift occurs. It can be due to the change in the frequency of features (covariate shift), change in the base rate of a class (prior probability shift), or change in the definition of a ground truth (concept drift). However, commonly, all 3 of these shifts are considered as concept drift and for the purpose of malware detection models this would suffice. As it was mentioned in the previous paper, the dissimilarity between the new sample and the existing samples forms the core of this technique and framework. Non Conformity Measure is used to calculate the distance between a new point from existing points and based on these values we calculate the p-value which is then compared to the threshold to make the decision whether the classification is reliable or not. The Transcend framework utilizes this measure using Conformal Evaluator to reject the samples that seem to be affected by concept drift, which could be analyzed further later on at a different part of the pipeline and/or utilized for retraining. The Conformal Evaluator produces two different metrics for a sample, Confidence, and Credibility. While confidence deals with the likelihood that a sample belongs to a certain class, credibility gives the reliability of this classification based on the training data (Statistics). The Evaluator has a phase of calibration where based on the p-values of a class, the threshold for each class is set. Once the Evaluator is classified it can then take in new samples and decide whether or not to reject them. There are different kinds of Non-Conformity Measure that are available for a user to choose from, and this choice can be made during calibration. Moreover, each rejection has a cost to it, whether it is in the resources spent to quarantine a sample until further analysis can be done on it, or the time taken to retrain a new model based on the rejected samples. Based on the different types of calibration techniques available, there are different Conformal Evaluators like Transductive Conformal Evaluation (TCE), Approximate Transductive Conformal Evaluation (Approx-TCE), Inductive Conformal Evaluation (ICE), Cross Conformal Evaluation (CCE).

Using all of these systematic approaches, an improved version of TRANSCEND was created called TRANSCENDENT. This was done utilizing CCE and the dataset was restricted to DREBIN instead of mixing DREBIN and MARVIN as it was done in their previous framework. The underlying classification model was a linear SVM and the quality metrics compared between were Confidence, Credibility, and Probability. These choices were made by testing out and comparing the results given by different combinations. TRANSCENDENT was also compared with past work in the field such as DroidEvolver and CP-REJECT. While CP-REJECT was found to fail to recognize drifting points in comparison to TRANSCENDENT, DroidEvolver suffered from poor labeling on the incoming sample over time. It was also shown that TRANSCENDENT works well with other models in different domains, such as GBDT

for Windows PE malware and RF for PDF malware. This points to the robustness of the framework. Even though the framework was found to be less affected by the underlying model a suitable NCM must be selected for its performance. Comparing different calibration methods, CCE was found to perform better and between evaluation metrics, credibility proved to be the best fit to decide the thresholds. Moreover, rising rejection rates must be interpreted as due to the deterioration of the classifier.

Discussion

- The discussion started off with a reiteration of the difference between confidence and credibility (which was a core theme in the last discussion too), where confidence uses probability whereas credibility uses statistics. It was also mentioned how the p-value is utilized to accept a null hypothesis which essentially means that the classification of a sample is under suspicion and should hence be quarantined and reevaluated.
- Context was given to this paper by explaining how it was an attempt to formalize and standardize the contents presented in the previous papers by the author. Also, some of the mistakes they made were rectified such as using DREBIN alone instead of both DREBIN and MAVIN. The presence of biases in their previous paper would lead to a loss of trust in the conclusions they drew using their experiments. "Science is about isolating variables", which they did not achieve in their previous paper.
- It was also mentioned that the metric of credibility is not widespread yet, as the paper is fairly recent. However, it is expected that the trend in the malware detection domain would lead to the acceptance of credibility as a valuable criterion in creating the models.
- A section that added value to the paper was where they went over the place of this tool inside a pipeline. Because a pipeline has the capacity to accommodate various independent techniques, this is what is used by most AV companies, and hence elaborating how TRANSCENDENT would fit into a pipeline makes it more practical.
- Another section that was useful is the cost that a rejection which was mentioned during the presentation. This cost directly translates into False Positive Rate and reducing the rejection cost works along the same trajectory as reducing the FPR. Spam detection was also mentioned as another area where FPR is a metric of large consequences and hence high value.
- An interesting point that was mentioned was the advantage in resources a big commercial company would have compared to a model built as a personal project. In the context of spam detection, a company like Google would have access to a large database of emails, which they can compare to see how many users got a certain mail. This would be a technique that would easily increase the suspicion score of an email as spam even before it reached any machine learning model. Commercial companies have advantages similar to this in the domain of malware detection too.
- Another point of discussion was how the thresholds were determined. It could either be done through a technique such as Grid Search that would exhaustively search through different combinations. It could also be done randomly which is proven to be more effective at times, especially considering the cost of computing resources and time. Moreover, the way thresholds are set could be by ensuring that each day around 20 samples would be flagged as undergoing

concept drift, and hence be evaluated by humans. This might be because that was the limit of human resources, that is, the maximum amount of samples a team of analysts could analyze in a day. This shows how the limitations in other factors such as human resources decide the threshold.

- Here, the discussion turned toward the attacker's side of things. It was mentioned how there were Russian-origin malware files, that were designed not to attack systems that had the Russian Language in them. This is an example of some levels of detection that the malware attempts to do before execution. Most malware nowadays has an anti-vm layer that detects virtualization tools or monitoring tools to ensure that the malware cannot be run in a sandbox environment for dynamic analysis. Even script kiddies can make use of this technique because such anti-vm detections are given out as services with which the amateur attackers can pack their basic payload.
- This led to a very brief discussion on packers, downloaders, and server-side polymorphism which are methods that the attackers employ to make the defender's tasks more difficult.
- Another point of discussion was on the usage of PDFs as a common means of spreading malware. This is because PDF is not textual even though it looks so. PDF files actually run instructions to generate the content, and this instruction-running capability of PDF files can be exploited by malware creators to run malicious instructions. A certain extent of this is possible in Office file formats too using Macros, which is why they are disabled by default.
- Since the features change from one file format to another, AVs normally have different models for different file formats. And since it is not possible to cover all the file formats that are available in the world, AV companies normally create models for the commonly used formats. This part is taken advantage of by attackers where they achieve the same goals through less-used file formats.
- This also leads to the fact that when a new file format comes out, it is only after some users get infected that an AV company would be able to build a model to cover this format.
- Another point of discussion was whether it is possible to create an AV model using the low-level commands as it might be immune to the obfuscations done at a surface level. It was stated that it could be theoretically possible
- Another point of discussion was if it is possible to infect a user with zero clicks. This is possible in case the user has any services with vulnerabilities in it. For example, if he's using an outdated version of a browser then a user who goes to a website could get infected with zero clicks. Another example is the iOS SMS parsing vulnerability that was exploited by NSO. However, this is rare and the majority of attacks would be phishing and similar methods.
- Concept drift happens in files like PDF and Doc files too but not to the extent of a PE File. This is because, in a PE file, there is more flexibility that can be used by an attacker whereas the working of files like PDF and Doc are more strict and hence there is not enough space for drift to happen at the scale it does in PE files.
- The possibility of malware that uploads all the files from a system was also mentioned. However, it was made clear that the more common type of malware is a Remote Access Trojan with Command&Control Centers.