Transcend: Detecting Concept Drift in Malware Classification Models - Part1 - Summary - SidharthAnil

Presentation

The paper mainly deals with a method to detect concept drift in machine learning models. The presentation started off with an introduction to the concept of concept drift and some visual aids to help understand the topic. Concept drift is a change in the statistical properties of an object in unforeseen ways. In the context of malware, the two main reasons for concept drift are malware evolution and new malware families. While the former focuses on the way malware evolves in response to the malware detection systems, the latter focuses on completely new malware families developed. A major engineering choice is when to retrain a model in response to concept drift. Since retraining takes resources, it does not make sense to retrain for small magnitudes of concept drifts.

The paper's contribution is a statistical evaluation technique, Conformal Evaluator, to quantify the drift in a model's classification and a framework called Transcend that utilizes CE. While the framework and its implementation are towards the second part of the paper, the first part and hence this presentation focuses on the Conformal Evaluator. The core idea is to use statistics to see how well a sample fits into the class it was classified into. For this reason, they use the P-value.

The instinctual way to understand P-value for a sample is to see how many samples in a class are more dissimilar to the class average than the sample under focus. This would mean, that if the P value of a sample reduces, it indicates that it is more dissimilar compared with the rest of the class. Using the P-values, a Non-Conformity Measure is derived that is used to detect the samples that are drifting. Moreover, the threshold value is computed separately for each class using the p-values of training samples. The higher you raise your thresholds, the more reliable your predictions/classifications are going to be.

Using this Non-Conformity Measure a case study was done on the domain of Android malware. A classifier was trained on the Drebin dataset, and then the presence of concept drift was checked using the Marvin dataset as a test set. It was observed that the classifier trained on Drebin was misclassifying a large amount of the Marvin test set. A subset of the samples that were drifting was identified using the Non-Conformity Measure, and this subset was fed back into a retraining phase of the classifier. This retrained classifier performed much better with the data confirming the presence of concept drift as the reason for the weak performance earlier.

This statistical technique provides a lot of added functionality or advantages over other existing methods. It is algorithm agnostic, as it can be applied in the same way irrespective of the model that forms the base of the classifier. Moreover, it uses statistics as its measure as opposed to probability, which provides an analysis along a different dimension. It shows how reliable the model's classification is, which could be adjusted quantitatively by adjusting the threshold value. This whole technique, therefore, provides a way to identify concept drift and rectify it before the model is affected too negatively.

Discussion

- The first point of discussion was how this measure of unreliability regarding the model's classification was done through statistical methods and not by creating a different machine learning model or anything like that. This would mean that a process of testing how reliable the model's classification is, could be run in parallel to the classification model with very little cost.
- The method to calculate the p-value was reiterated over so that the class understood clearly how it relates to the sample being dissimilar compared to the rest of the samples in the class. The technical definition of the value was also mentioned where it showed the likelihood of the sample occurring under the null hypothesis. Another point mentioned regarding this was that it was not likely that each and every sample would be put through this statistical measure in a malware detection pipeline. It was much more likely that a subset of the samples would undergo this process, and that would be enough to detect if the model is suffering from drift.
- One shortcoming of the paper that was mentioned in class was the way they set the threshold values to reflect good results. This is not indicative of the real world and hence doesn't hold a lot of practical significance. However, it was also mentioned that this method is one of the more recent advancements in the area of detecting concept drift. And the mathematical details that are presented in the paper would help a reader who is intending to create such drift detection systems or try to validate the conclusions drawn from such techniques.
- A major difference that this method focuses on is how it relies on statistics and not probability. This leads the discussion to the difference between the two at least in the context of the paper, or malware detection in general. The way to understand this is that statistics deals with the data you already have while probability makes predictions about the data in the future.
- This kinda difference at a mathematical level between the NCM and the regular evaluators would mean that the NCM method provides a metric along a different dimension. While the confidence level of the model predicts how confident the model is about its classification, the NCM works in parallel, irrespective of the model, and tells how similar the new sample is compared to the files that the model has already seen. These are two different measures, where one deals with confidence while the other deals with similarity. Because these two measures deal in different axes, there are 4 different possible combinations:
    1. Files classified as malicious and similar to the malicious files seen before
    2. Files classified as malicious and dissimilar to the malicious files seen before
    3. Files classified as benign and similar to the benign files seen before
    4. Files classified as benign and dissimilar to the benign files seen before
- Building from the previous point, it should be observed that these two measures provide a different perspective on the same data. They are both derived from the same data, but it shows two points of

view which is very important. This aspect of the two points of view, is very similar to the way accuracy and precision deal with different aspects of the same data.

- A strategy could be utilized where both probability and statistics are used in combination with each other. Since statistics deals with the present data, it is more short-term but more aggressive. On the other hand, probability deals with future data, it is long-term and less aggressive. For this reason, one combination that was proposed was that probability could be used as the first layer, and when it fails to provide a clear-cut decision, statistics could be utilized as the more aggressive second layer.
- One major drawback of the paper was that they mixed two datasets, DREBIN and MARVIN. This means that the entire research would fall under the Sampling Bias pitfall. This would also mean that Data Snooping would be present which would take away the credibility of their results. Even though they removed the duplicate entries from DREBIN, there are still scenarios possible that a certain feature that appeared on a certain timestamp on MARVIN appeared earlier on DREBIN and hence was learned by the model. This would be mistakenly inferred as the model dealing with concept drift effectively when in reality it was just because the training dataset contained a peek towards the future that allowed it to learn unrealistically.
- Another point mentioned was that concept drift is a phenomenon that also happens in benign files, but not to the extent it happens in malicious files. This is because in benign files it has to occur due to a natural shift, whereas, in the case of a malicious file, there is a factor of the malware creators trying to move away from the current data distribution in order to make their files surprising and evasive.