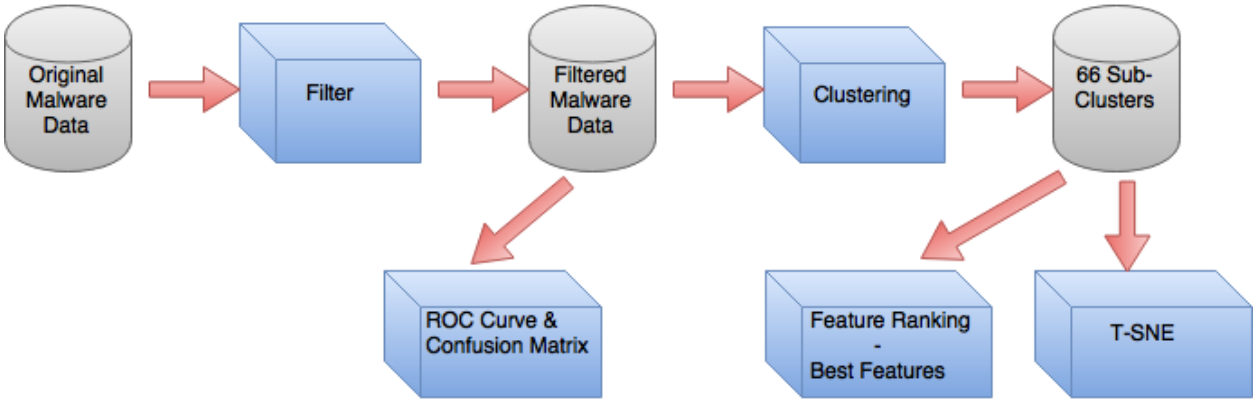
**Abstract:**
This report recorded the second phase of a course project done by graduate students from Applications of Advanced Analytics to Cybersecurity Course lectured by John Cavazos from University of Delaware, which implemented the whole Class Signature process to analyze a testset of 1100 malware to distinguish discriminative features of good predicted and bad predicted malware clusters. The process in second phase mainly include adding filter module, generating ROC curve / confusion matrix, developing a web-page for display and optimization from the first phase.

**Workflow:**




**Overview of Filtering Module:**
In the first phase of this project, we started Class Signature process directly from clustering module. As a result, the feature ranking and T-SNE modules displayed the visual analytic results of 66 sub-clusters simply generated from the 1100 original malware data via K-Means clustering algorithm.
However, the whole Class Signature process should be implemented based on a prediction model at the beginning. Because the experts who created Class Signature process used this technique as a tool to analyze prediction model. In the second phase of this project, we developed filter module, which consists of a decision-tree model and thresholds. This filter module made Class Signature process become meaningful in visual analytics of malware.

**Overview of Web-Page:**
In the first phase of this project, we generated the feature ranking lists and T-SNE pictures for 66 malware sub-clusters. In the second phase of this project, we optimized feature ranking lists as histogram and optimize T-SNE as same shape with different colors for 66 malware sub-clusters. Besides feature ranking and T-SNE, we also generated ROC curve and confusion matrix for 11 malware families. However, all of these visual results were saved separately in folder.

Thus, we designed and developed a friendly web-page to display all of the visual results together in grid style via HTML framework.

**Filtering:**

In filter module, we firstly imported 9,900 malware dataset to train a decision-tree model using sklearn open source library. Then, we imported the 1,100 malware test set to this model, which is the same original dataset in project phase one. After that, we can get the family label prediction probabilities of these 1,100 malwares.

In the second step, we set the thresholds (<=0.2 or >=0.8) to filter 1,110 malware test set based on prediction probabilities. The "<=0.2" threshold means the prediction probability of target family label is less than or equal 20%, which can represent the very bad predictions. The ">=0.8" threshold means the prediction probability of target family label is bigger than or equal 80%, which can represent the very good predictions. Through this way, we can filter out median predictions and only keep the very good predictions and the very bad predictions. Since our prediction model is very good, the majority predictions are very good. As a result, we still got 882 filtered malware dataset.

Finally, we use three CVS files to save all the information of these 882 filtered malware dataset. The first CVS file called "filtered_matrix.csv" saves all the feature information of filtered malware dataset. The second CVS file called "filtered_target.csv" saves all the ground truth of family labels of filtered malware dataset. And the third CSV file called "filtered_prediction.csv" saves all the prediction probabilities of filtered malware dataset. There are 11 family labels in total, so there are 11 columns of prediction probabilities in this CSV file. The first two CSV files serve as the input for clustering module. The rest process are same with project phase one, which are generating 66-sub clusters with feature ranking and T-SNE.

The third CSV file serves as the input for ROC curve, which is a new feature in project phase two.

**Web-Page:**

To design and develop a web-page, we use Bootstrap as framework. Bootstrap is the most popular HTML, CSS, and JS framework for developing responsive, mobile first projects on the web. We displayed all of our visual results in grid on the web-page with the title of "Malware Class Signature". There are 11 rows on the web-page, which represent for 11 families based on labels. For each row, there are 7 columns. The first column contains the ROC curve and confusion matrix for its family. The rest 6 columns contains 6 sets of T-SNE graphs and feature ranking histograms for 6 sub-clusters generated by clustering module.

**Conclusion:**

In project phase two, we finished the whole process of implementing Class Signature technique on malware dataset. Comparing with project phase one, we add new filter module, ROC curve/confusion matrix, a web-page and optimization for feature ranking from phase one.

The filter module consists of a decision-tree prediction model and thresholds. After filtering, we got the dataset of very good and very bad predictions. The filtered dataset serve as the input for clustering module, which is same with the process in phase one. The prediction probabilities can serve as the input for ROC curve for each malware family based on label name, which is a new feature in phase two. Besides that, we can also get the confusion matrix for each malware family. In feature ranking, we generate the histograms of top ten features for each sub-cluster.

In order to display all the visual results on a friendly GUI, we design and implement a web-page via html framework. There are 11 rows which represent 11 malware families. For each row, there are 7 columns. The first column contains the ROC curve and confusion matrix. The remaining 6 columns consist of 6 sets of T-SNE graphs and feature ranking histograms.