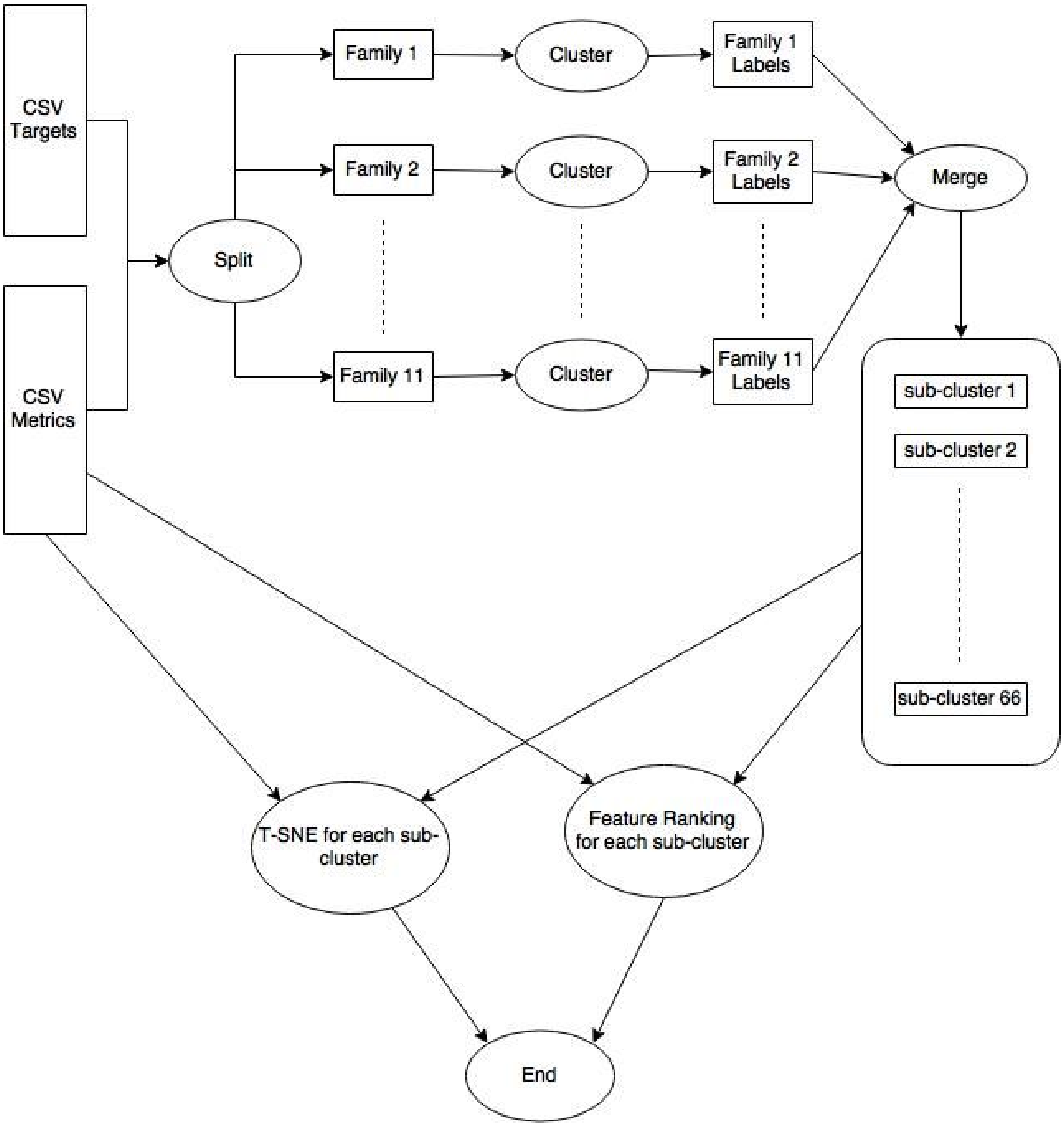
Malware Class Signature (Phase 1)

Clustering: Split & Merge Module

Wanxin Li

April, 2017

Project Pipeline



Clustering

As the project pipeline shows, the first part of this project is to generate sub-clustered files as inputs for feature ranking and T-SNE. In project 1, we choose the subset c from course Github repository [1] as the original input data for clustering. This subset contains 1,100 malwares which are all recorded in two CSV files. The first CSV file ”Metrics” records malware ID and matching features. There are 347 feature columns in total. The second CSV file ”Targets” records family label in matching order. There are 11 different family labels in total. And each family contains 100 malwares.

The goal for clustering part is to generate 6 sub-clusters for each family based on K-Means algorithm [2]. And we will get 66 sub-clusters in total. We divide this clustering part into three modules: Split, Cluster, Merge. All three modules are written in a same python script “cluster.py” [3].

For the first module, we need to split 1,100 malwares into 11 groups based on family labels. Our solution is to import all values from “Targets” and “Metrics” into a python two-dimensional list data structure. Each row contains a malware’s ID, 347 features and matching family label. Then we sort the 1,100 malwares based on family label in alphabetical order. After that, we can split the 1,100 malwares into 11 groups, which are saved in 11 two-dimensional list data structure. In the meanwhile, the family labels are deleted before saving into these 11 new sub-lists. Because the family label are not the expected input for K-means algorithm. The new sub-lists are in the same format with the 1,100 malwares list. The only difference is that the new sub-list contains 100 rows of malwares.

After applying K-means algorithm for each family, we generate 11 clustered label lists, each list has two column: malware ID and clustered label value. These clustered labels are represented in integer from 0 to 5, which are 6 different sub-clusters for a specific family. For the third module, we need to merge these 11 clustered label lists and output 66 label files in CSV or TXT format as the inputs for feature ranking and T-SNE part.

For module 3, we firstly sum up all the 11 clustered label lists into one list which contains the whole 1,100 malwares. Obviously, this list has a wider range of clustered label values from Integer 1 to 66. Next, we use this list to generate 66 sub-cluster label files. The file format is CSV. Each label file has 1,100 rows and two columns: malware ID and label value. The label value contains only integer 1 or 0. The integer 1 represents the sub-cluster which we want to display as highlight in T-SNE. The integer 0 represents the other 65 sub-clusters which we want to display as the background in T-SNE.

Future Work

In project 1, we import malware subset c [1] as the original input for class signature process, which contains 1,100 malwares. For next step, we will import the whole malware dataset into the class signature process, which will contain 11,000 malwares in total. In clustering part, we will handle tenfold malwares data comparing with project 1. The output will still be the 66 sub-clustered files. However, each file will contains 11,000 rows instead of 1,100 in project 1.

Besides that, we also need to develop a prediction model, which will import nine malware subsets as training set and the rest one as testing set. Through this way, we can get the confusion matrix, FPR-TPR, Recall-Precision, Threshold-Accuracy graphs.

Reference

[1] Course Github Repository:
https://github.com/cavazos-lab/spring-2017-CISC850-project/blob/master/dataset.md
[2] K-Means Clustering: https://en.wikipedia.org/wiki/K-means_clustering
[3] cluster python script:
https://github.com/cavazos-lab/spring-2017-CISC850-visual-analytics-1/blob/master/scripts/cluster.py