CSCI-B 565 DATA MINING
Project Report - Microsoft Malware Classification Challenge
Morning Class
Computer Science Core
Spring
Indiana University,
Bloomington, IN

Nayana Charwad (ncharwad@umail.iu.edu)
Amritanshu Joshi (amrijosh@umail.iu.edu)

April 28 2015

All the work herein is solely by authors.

1 Company

Kick It Out Data Protection Systems: Kick It Out or KIO as it is known in the data management market is a leading data protection service provider. Brainchild of founders Amritanshu Joshi and Nayana Charwad, KIO came into existence in spring of 2010, when the age of big data was in a nascent stage. Focusing mainly on data protection services, it grew from day to day. Today KIO boasts an international clientele ranging from big honchos like Microsoft and Apple to various startups. The workforce has grown manifold from 20 to around 2000 employees in 5 years. Since data is an entity which keep on growing, efforts from KIO to keep the data free from malwares and protect sensitive information has been ever-present. As a result, KIO has been the recipient of many awards. The 2014 Big Data Protection Force Award was the latest gem in KIOs collection. Next section will walkthrough profiles of our founders in detail.

1 **Amritanshu Joshi:** Graduated from Manipal University in 2012, Amritanshu was hired by Ericsson, the telecom major. He worked there for 2 years as a Software Developer before opting for higher studies. In fall of 2014, Amritanshu joined Indiana University in Bloomington to pursue Master of Science degree in Computer Science. Having an immense interest in data security and the advent of big data, Amritanshu thought of starting his own data security firm. Wasting no chance he started KIO with Nayana Charwad, a fellow Masters student in Indiana University.

2 **Nayana Charwad** Graduated from Pune University, Nayana has experience working with IBM for five years. Working as technical consultant for Enterprise Resource Planning systems Nayana developed great interest in big data, data mining and data security. Her areas of interest also consists of cloud computing and object oriented software developments. As a result, she joined Indiana University in Bloomington to pursue Master of Science degree in Computer Science. Since Nayana always wanted to start her own firm, when opportunity came, she immediately joined Amritanshu to start KIO.

Next sections of this report will explain detail project requirements, approaches KIO explored to achieve best results and future enhancements for this project.

2 Project Details and Requirement

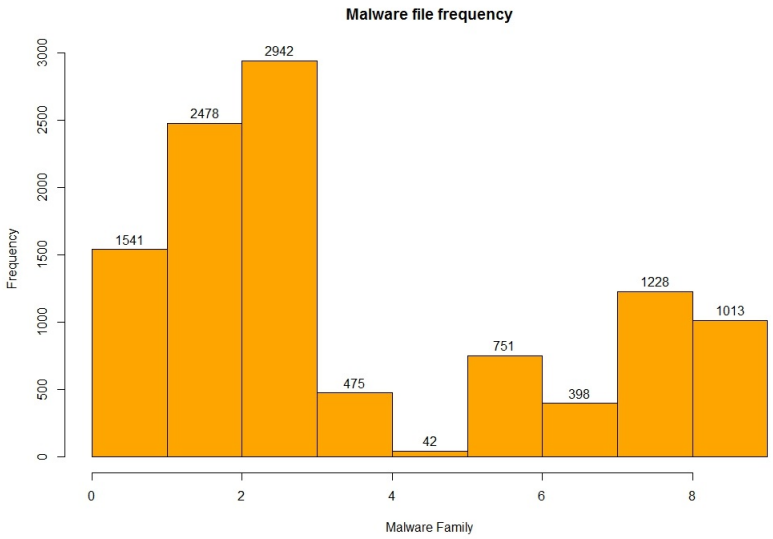
1 Contractor Acme Dynamic General Systems

Our contractor Acme Dynamic General Systems is a data processing and data management company. It manages data for around 200 zoos all over the United States. The workforce which makes this possible is around 4000 employees. Highly respected organization in the field of data, it recently acquired a number of zoos in the international market and will be providing them services from the next financial year.

2 Brief high-level outline

Acme handles a lot of data on a regular basis. On an average in a day it handles around 10 TB of data. Mostly this data comprises day to day activities in all the zoos in the United States. This activity reporting is in form of bytes files which are created using assembly language code using their top secret compiler. The creation of assembly language code is handled by each zoo which sends it to a centralized server where it is compiled and converted to bytes and thus all the data gets stored. Last month Acme reported a breach in their firewall when an outsider uploaded malicious assembly language files onto their server. Although they had a backup of their data stored securely, this activity affected their systems and they had to face a lengthy downtime. To avoid such a cataclysmic situation in the future Acme hired us to create a solution for their problem.

3 Problem Description We were given 10868 files each of assembly language code(.asm) and bytes code(.bytes) as part of model training data and 10873 files each of .asm and .bytes as part of test data. Our task was to create a prediction mechanism wherein our model predicts the family of malware which a .asm or a .bytes file belongs to. Th training data(which was around 250GB in size) was to be used to create the prediction model and the test data(also 250GB) was put into that trained model to predict the family of each file in test data. The final output was expected to be a comma separated values(.csv) file containing the file name and probability of each file against the 9 different malware families which it can belong



3 Executive Summary

(a) Solution

In this project we implemented below steps for data reduction, classification and prediction.

1 **Feature Selection**

Various feature selection methods such as bag of words, term frequency-inverse document frequency(tf-idf) were used to identify important feautures from given dataset.

2 **Feature Reduction**

Since number of features generated were considerably high, correlation matrix and PCA were then used for feature reduction. This helped in reducing prediction error due to highly correlated features.

3 **Classification and Prediction:**

Next step was to classify the data and predict the test data based on this model. Reasons for choosing random forest over other classification and prediction methods are outlined in the table below:

Algorithm	Accuracy	Outcome
Naive Bayes	57.32%	Very low accuracy
Support Vector Machine	74.63%	Better than Naive Bayes but still less accurate
Random Forest	91.59%	Best accuracy obtained before applying feature selection algorithm

We chose to go ahead with random forest as it gave us better accuracy than the other two algorithms. Selected features were then passed onto the algorithm for classification and prediction.

Technology Used

We implemented Java programs for calculating bag of words and tf-idf. R was used for initial data analysis, data visualization, implementation of random forest, finding correlation between features and PCA.

Below table represents log loss comparison of various implemented approaches for malware classification. Next section of this report explains each approach in detail. Public score represents log loss score on 30 percent test data whereas private score represents log loss score on remaining 70 percent test data.

[Log loss scores are referenced from Microsoft Malware Classification challenge at www.kaggle.com]

Number	Approach	Public Score	Private Score
1	Frequency of Asm and Bytes combined after manually removing highly correlated features (mtry = default, ntree = 2400)	0.068625947	**0.059689639**
2	Frequency of Asm and Bytes combined(mtry = default, ntree = 1200)	0.068973453	**0.059921639**
3	Frequency of asm opcodes only(mtry = 6, ntree = 1200)	0.086469677	0.066620868
4	Deleted 34 correlated attributes. default mtry, ntree = 1200	0.071617142	0.066683227
5	Frequency of asm opcodes and byte codes removing 80 correlated features	0.081970783	0.073906011
6	Frequency of asm opcodes and byte codes removing 50 correlated features	0.082204854	0.073932519
7	Frequency of asm opcodes only(default mtry, ntree = 1200)	0.081627125	0.076444187
8	Frequency of byte codes only(mtry = 6 and ntree = 800)	0.10535592	0.097550866
9	Frequency of byte codes only (mtry = 6, ntree=1200)	0.104922648	0.10455553
10	Frequency of byte codes only(mtry = 4, ntree = 1200)	0.110988767	0.106149539
11	Frequency of byte codes only (First Submission)	0.110000475	0.110222416
12	PCA with only ASM files included	0.217618833	0.182090641
13	TFIDF with all features included	1.038320615	1.034270105
14	TFIDF with removal of features below certain score	1.051373076	1.07103843
15	PCA with all features	1.655404655	1.656669458
16	PCA with manually selected features	2.238051168	2.24474315

1 Bag of words model

In this model occurence of every byte code in byte file and occurence of every operation code in asm file is used as feature. Frequency of all occurences of each byte or operation code is computed for each file.

File Type	byte files	asm files
Features selected	256	90

As we can see total number of features selected with bag of words model is considerably high and hence correlation analysis is done using R to find out correlation between features and remove highly correlated features.

Algorithm to remove highly correlated features

1. Get correlation matrix from R.
2. Remove the elements on and below diagonal in correlation matrix.
3. In remaining matrix, fetch all elements having correlation value more than 0.6
4. Based on selected elements, remove one feature from two highly correlated features.

Features extracted after feature selection and feature reduction are then passed to random forest decision tree algorithm for classification and prediction.

Observations

1. Best accuracy and lowest log loss **0.0596** was achieved with this approach when we combined features of both asm and bytes files and removed 120 highly correlated features
2. Features extracted from asm files were more meaningful as compared to features extracted from byte files.
3. Increase in number of trees for random forest achieved more accuracy. However accuracy remained constant after we increased number of trees to 2400.
4. Accuracy increased when we selected greater values of mtry during implementation of random forest. mtry represents number of features randomly sampled as candidates at each split

2 Principal Component Analysis

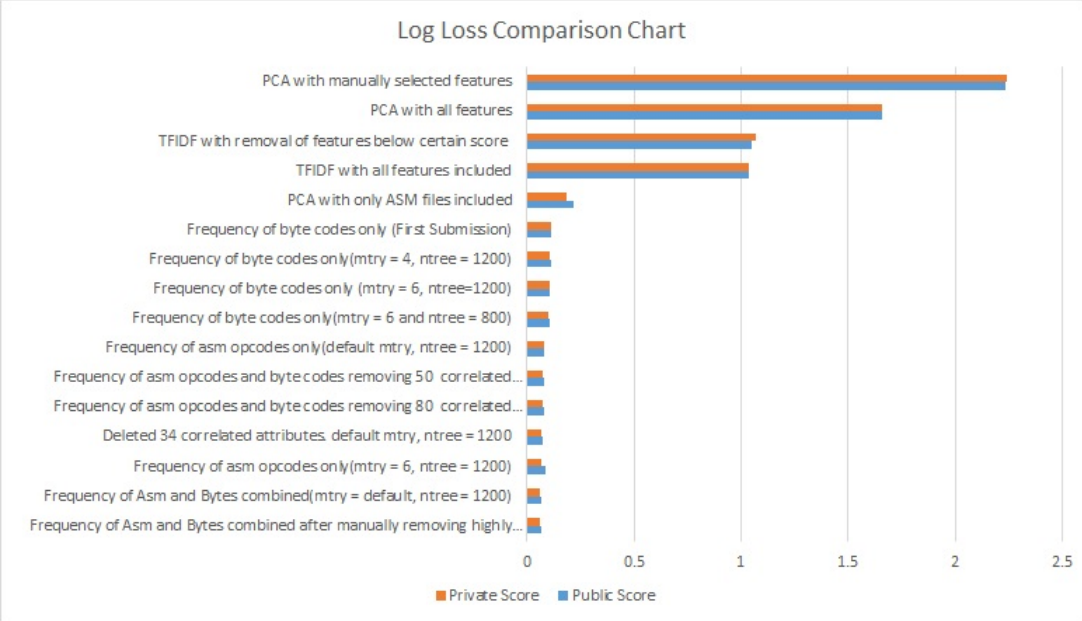
Principal component analysis creates orthogonal pricipal components in existing dataset. Each principal coponents have high variance with other principal components. Since top principal components have very less correlation with each other we selected best features based on top principal components from PCA. As per paper [1] below algorithm is used to extract best features from existing dataset.

Algorithm to extract features using PCA

1. Prominent principal components are selected based on the variance of data they represent.
2. Eigen vectors for each principal component are sorted based on significant values of attributes.
3. Top k attributes are selected from each selected principal component.
4. Selected attributes were appended to final list of features used for classification and prediction.

Observations

1. Log loss with feature selection approach using PCA was higher as compared to bag of words method.
2. As paper [1] suggests we may need to combine minimum redundancy maximum relevance method to extract best features.
3. Features extracted from asm files were more meaningful as compared to features extracted from byte files.
4. Using only PCA does not work well for categorical data as compared to continuous data.



3 Term Frequency-Inverse Document Frequency(tf-idf)

tf-idf is numerical computation method that represents importantance of a particular word in a given document as compared to rest of the documents in corpus. With this approach, words which occur frequently in all documents are given lower priority as compared to words which occur in specific documents. We implemented Java program to extract best features from asm and bytes file using tf-idf.

Term frequency for each word is calculated with formula :

$$TermFrequency(tf) = 0.5 + \frac{(0.5 * f(t,d))}{f(w,d)}$$

f(t,d) = Raw frequency of word in document

Inverse Term frequency for each word is calculated with formula :

f(w,d) = Maximum raw frequency of any word in document

$$InverseDocumentFrequency(idf) = log + \frac{N}{d}$$

N = Total number of documents

d = Number of documents where given term is present

tf-idf for each word is then calculated with belwo formula

$$tf - idf = tf * idf$$

Observations

1 Log loss with feature selection approach using tf-idf was higher as compared to bag of words method.

3 Features extracted from asm files were more meaningful as compared to features extracted from byte files.

(b) Outcome

As discussed in earlier sections of report, random forest decision tree algorithm delivers promising results for categorical data as compare to Support vector machines. Asm files extract more

meaningful features as compare to byte files and features from both files can be combined to deliver highest accuracy. Also after analyzing multiple approaches for feature selection and feature reduction methods we conclude that bag of words model combined with removal of highly correlated features is best suited for features selection and feature extraction. Currently we used manual method to remove highly correlated features but future work should include dynamic programming to extract best features from given set of features.

4 Future

We think existing accuracy can be improved further by implementing below future enhancements.

1 Sophisticated frequency algorithm

Since asm files give more accurate features as compared to byte files, information gathered from asm files can be used in more meaningful ways. For example, operation code counts inside loops can be increased based on value of loop counter.

2 Dynamic algorithm for feature selection

Dynamic algorithm can be implemented to remove highly correlated features and extract more meaningful features.

3 Feature Selection using asm images

As suggested at research project [2] asm files can be converted to images and then top features can be selected based on the newly formed image file.