Comenius University, Bratislava
Faculty of Mathematics, Physics and Informatics

Comparison of machine learning algorithms for classification of algorithmically generated domains

Master's Thesis

2020
Bc. Frederik Koľbík

Comenius University, Bratislava
Faculty of Mathematics, Physics and Informatics

Comparison of machine learning algorithms for classification of algorithmically generated domains

Master's Thesis

Study programme:	Computer science
Study field:	Computer science
Department:	Department of Computer Science
Supervisor:	Mgr. Jakub Daubner, PhD.

Bratislava, 2020
Bc. Frederik Koľbík

Comenius University in Bratislava
Faculty of Mathematics, Physics and Informatics

THESIS ASSIGNMENT

Name and Surname:	Bc. Frederik Koľbík
Study programme:	Computer Science (Single degree study, master II. deg., full time form)
Field of Study:	Computer Science
Type of Thesis:	Diploma Thesis
Language of Thesis:	English
Secondary language:	Slovak

Title: Comparison of machine learning algorithms for classification of algorithmically generated domains.

Annotation: - examine related work on using machine learning for detection of algorithmically generated domains
- analyze various features of algorithmically generated domains and identify the most signifcant ones
- experimentally compare the accuracy of various machine learning algorithms on several datasets and on real data

Supervisor:	Mgr. Jakub Daubner, PhD.
Department:	FMFI.KI - Department of Computer Science
Head of department:	prof. RNDr. Martin Škoviera, PhD.
Assigned:	06.04.2020

Approved: 06.04.2020 prof. RNDr. Rastislav Kráľovič, PhD.
Guarantor of Study Programme

.....
Student

.....
Supervisor

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Frederik Koľbík
Študijný program: informatika (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: informatika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: anglický
Sekundárny jazyk: slovenský

Názov: Comparison of machine learning algorithms for classification of algorithmically generated domains.
Porovnanie algoritmov strojového učenia pre klasifikáciu algoritmicky generovaných domén.

Anotácia: - preskúmajte doterajšiu prácu pri použití strojového učenia na detekciu algoritmicky generovaných domén
- analyzujte rôzne črty algoritmicky generovaných domén a určte najvýznamnejšie
- experimentálne porovnajte presnosť rôznych algoritmov strojového učenia na viacerých datasetoch, aj na reálnych dátach

Vedúci: Mgr. Jakub Daubner, PhD.
Katedra: FMFI.KI - Katedra informatiky
Vedúci katedry: prof. RNDr. Martin Škoviera, PhD.

Dátum zadania: 06.04.2020

Dátum schválenia: 06.04.2020 prof. RNDr. Rastislav Kráľovič, PhD.
garant študijného programu

.....
študent

.....
vedúci práce

Abstrakt

Mnoho malvéru v posledných rokoch začalo používať algoritmy na generovanie domén (DGA, z angl. Domain Generation Algorithm) pri komunikácii s riadiacim serverom. Tieto algoritmy generujú veľké množstvo domén, ale len malá časť je naozaj použitá pri komunikácii so serverom. Za posledné roky boli otestované viaceré spôsoby detekcie takýchto domén. Prístupy založené na strojovom učení sa stali veľmi populárnymi a úspešnými. V tejto práci skúmame rôzne typy DGA algoritmov a to, ako ich detegovať a poskytujeme porovnanie a vyhodnotenie piatich algoritmov strojového učenia s učiteľom pre klasifikáciu DGA domén s použitím viacerých množín charakteristických čŕt. Počas našich testov sme zistili, že najlepšie túto úlohu spĺňajú algoritmy založené na rozhodovacích stromoch. Takisto sme analyzovali ťažko detegovateľné DGA algoritmy a domény, ktoré generujú.

Kľúčové slová: malvér, algoritmus na generovanie domén, strojové učenie, klasifikácia

Abstract

In recent years, a lot of malware has started to use domain generation algorithms (DGAs) in communication with command-and-control servers. These algorithms generate a large number of domains, but only a small portion of them are actually used in C&C communication. Over the years, there have been numerous ways of detecting these kinds of domains tested. The approaches based on machine learning have become very popular and successful. In this thesis we look at different types of DGAs and how to detect them and provide a comparison and evaluation of five supervised machine learning algorithms for DGA domain classification using multiple sets of features. During our tests, we have found that decision tree-based algorithms perform the best. We have also analyzed hard-to-detect DGAs and the domains they generate.

Keywords: malware, domain generation algorithm, machine learning, classification

Contents

List of Tables

Introduction

A lot of malware needs to communicate with its command-and-control servers, botnets particularly. At first, botnets used a static list of domains that they needed to establish connection with. Static lists of domains can be easily blocked or blacklisted, that is why malware authors have come up with domain generation algorithms (DGAs). These algorithms dynamically generate a large number of domains which are then resolved by malware in order to get an IP address of the C&C server. The authors have to register only a very small portion of the generated domains.

Naturally, a need to detect these kinds of domains have arisen in security research. Machine learning algorithms have proven to be quite useful for this task. Many supervised and unsupervised learning algorithms have been tested, recently, deep learning algorithms have become popular. Our focus is on supervised learning methods and on arithmetic-based and hash-based DGAs that often generate domains that look like a random cluster of letters and digits.

In this thesis we provide a comparison of five supervised learning algorithms for classification of algorithmically generated domains: Random Forest, Gaussian Naive Bayes, Logistic Regression, Gradient Boosting Classifier and Support Vector Machine. We train these models with five different sets of features extracted from the domain names and evaluate them by performing two types of experiments. We also analyze hard-to-detect domains of DGAs used by various malware families.

The thesis is structured as follows: chapter 1 contains basic information about domain generation algorithms and examples of various types of DGAs and the domains they generate. In chapter 2 we describe machine learning algorithms that we use in our experiments, in chapter 3 we provide an overview of related work of DGA domain detection. There is a description of experiments that we perform and the data we use in chapter 4 and in chapter 5 there is a list of features we use to train the models. Finally, in chapter 6 we examine the results and provide an analysis of hard-to-detect domains.

Chapter 1

Domain generation algorithms

In this chapter we provide basic information about domain generation algorithms and how malware uses them, and list examples of different types of DGAs.

1.1 Malware and domain generation algorithms

A lot of malware seen in the world nowadays needs to communicate with its command-and-control (C&C) server. As the name suggests, through it the malware authors can send commands to malware instances on victims' computers and control them. Alternatively, the connection with the server can be used to exfiltrate data from the victim such as screenshots, logged keys or even whole files. The C&C infrastructure is particularly useful for controlling botnets - large networks of infected devices (bots), which can be used for denial-of-service attacks, stealing data, distributing spam or mining cryptocurrencies. The C&C communication is illustrated in figure 1.1.

First botnets used centralized C&C servers, the bots queried a predefined C&C domain name, which was resolved to an IP address of the C&C server and then, a connection could be established. This means that the domain name was either hard-coded or obfuscated in some way in the malware. This introduces a single point of failure - the malware can be reverse-engineered, the domain name can be extracted and then blocked or blacklisted. In some cases the entire C&C server can be taken down, in which case the botnet operators lose control over the entire botnet.

To mitigate this single point of failure, malware authors have developed domain generation algorithms (DGAs) [30] [2], which dynamically generate a list of random domains (algorithmically generated domains - AGDs, we also refer to them as "DGA domains"). Only a small number of domains in this list is used for C&C communication. If the currently used domain is discovered and blocked, the authors can register another domain from the list and their operation continues without much interruption.

In more detail, the DGAs work like this. The malware periodically runs the domain

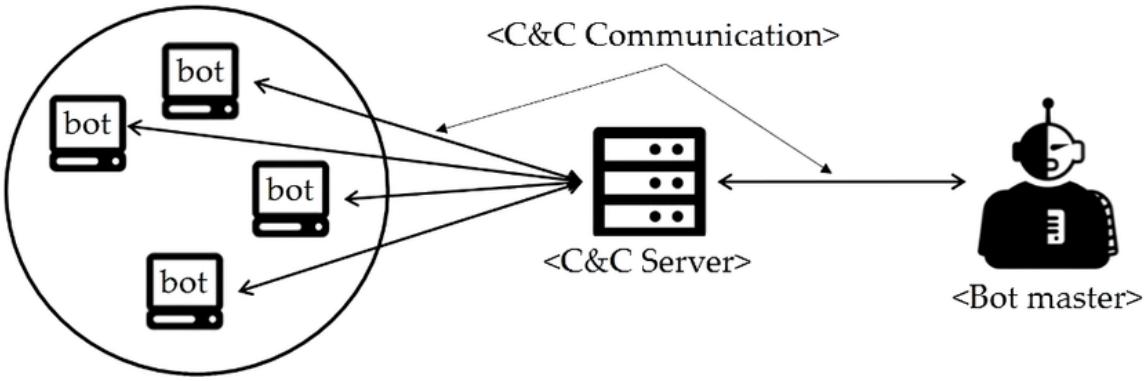

Figure 1.1: C&C communication. Jeon, Jaewoo & Cho, Youngho. (2019). Construction and Performance Analysis of Image Steganography-based Botnet in KakaoTalk Openchat. Computers. 8. 61. 10.3390/computers8030061.

generation algorithm, which produces a list of domains. The malware sends DNS queries trying to resolve generated domains, until one domain successfully resolves to an IP address of the C&C server. The aim of the DGAs is that they should be as unpredictable for security researchers as possible, but at the same time predictable for malware authors so that they know what domains are generated at a given time and so that they can register some of them. To achieve this kind of predictability, the DGAs generate domains based on a shared secret (seed), which is known both to malware and its authors. The typical seeds used are numerical constants or current time or date.

The two most significant properties of seeds are time-dependence and determinism. The DGA is time-dependent if it incorporates a time source for generating domains, resulting in a limited period of time during which the generated domains are valid. The time source can be time or date (of the infected machine, of the HTTP response from C&C server...) or something else, for example a trending topic on Twitter.

This brings us to the second important property of seeds - determinism. For majority of known DGAs, the DGA parameters are known beforehand, which means that all domains can be generated at any time, even domains which will not be valid until some point in the future. On the other hand, there are a few exceptions - the Torpig [34] malware family uses the aforementioned Twitter trends as a seed and Bedep [32] family uses foreign exchange reference rates published daily by European Central Bank as a seed. This leaves only a limited time for the malware authors to register generated domains once the seeds become available.

Based on time-dependence and determinism seeding properties, malware using DGAs can be divided into four categories: time-independent and deterministic (TID), time-dependent and deterministic (TDD), time-dependent and non-deterministic (TDN) and time-independent and non-deterministic (TIN). According to Daniel Plohmann et al. [30], there was no known malware in the last category at the time of their research.

Other thing worth mentioning is the use of top-level domains (TLDs). While some malware families use only one TLD in their DGA, others use varieties of TLDs. The reason for this can be quick availability and low fees for domain registration. Also, for some TLDs like ".xyz" or ".top", the registration can be anonymous and automated.

1.2 DGA types

Based on domain generation scheme that malware uses, there are four known types of DGAs [30]:

1.2.1 Arithmetic-based DGAs

These are the most common DGAs used by malware. There are two approaches of how to use arithmetics to generate domains, either directly calculating ASCII values of characters or using an array of characters and calculating an offset in this array. We illustrate these two approaches on the following examples:

DGA used by Ranbyus malware family

The first example is the DGA used by Ranbyus malware family. This DGA uses the current date and a hard-coded seed to generate ASCII values of characters, thus generating whole domain name.

```
for i = 0 to 13:
    day = (day >> 15) ^ 16 * (day & 0x1FFF ^ 4 * (seed ^ day))
    year = ((year & 0xFFFFFFF0) << 17) ^ ((year ^ (7 * year)) >> 11)
    month = 14 * (month & 0xFFFFFFFE) ^ ((month ^ (4 * month)) >> 8)
    seed = (seed >> 6) ^ ((day + 8 * seed) << 8) & 0x3FFFF00
    int x = ((day ^ month ^ year) % 25) + 'a'
    domain[i] = x
```

Example 1: Pseudo code of DGA of Ranbyus. Reversed and reimplemented by Johannes Bader [5].

Examples of generated domains:

```
hcfoopojnuqxho.su
undrdsbhivryqn.tw
dkehliueofdued.net
mpuakxjqpscfpj.com
eelolbwmfmtkae.pw
```

DGA used by Simda malware family

The second example is the DGA used by Simda malware family. In this DGA the characters are taken alternately from arrays of consonants and vowels based on a key and a base. The length, TLD and key change in different malware samples.

```
length = 7
tld = "com"
key = "1676d5775e05c50b46baa5579d4fc7"
base = 0x45AE94B2

consonants = "qwrtpsdfghjklzxcvbnmv"
vowels = "eyuioa"

step = 0
for m in key:
    step += ord(m)

for nr in range(1000):
    domain = ""
    base += step

    for i in range(length):
        index = int(base/(3+2*i))
        if i % 2 == 0:
            char = consonants[index % 20]
        else:
            char = vowels[index % 6]
        domain += char

    domain += "." + tld
    print(domain)
```

Example 2: Python code of DGA of Simda. Reversed and reimplemented by Johann Bader [6].

Examples of generated domains:

```
gatyfus.com
lyvyxor.com
vojyqem.com
qetyfuv.com
puvyxil.com
```

1.2.2 Hash-based DGAs

These DGAs use hash digest or a part of it as a generated domain. Hashing functions such as MD5 or SHA256 are usually used. An example is listed below.

DGA used by Dyre malware family

The DGA used by Dyre malware family calculates SHA256 hash of the current day and a number from some range. The first byte in the hash is then replaced with a byte derived from this number. This edited hash is taken as a domain name, which is finally concatenated with a TLD chosen from a list.

```
def dyre_dga(num)
        date_str = '{0.year}-{0.month}-{0.day}'.format(date.today())

        tlds = ['.cc', '.ws', '.to', '.in', 'hk', '.cn', '.tk', '.so']
        hash = sha256('{0}{1}'.format(date_str, num)).hexdigest()[3:36]
        replace_char = chr(0xFF & ((num % 26) + 97))

        return '{0}{1}{2}:443'.format(replace_char, hash, tlds[num % len(tlds)])

todays_domains = [dyre_dga(i) for i in range(333)]
```

Example 3: Python code of DGA of Dyre. Reversed and reimplemented by Talos [18].

Examples of generated domains:

```
bd9b9c8ca02a67700b45839adb1f37e736.ws
d66e28de33bcabb213a1de204887f5fa04.in
ga871a3a9443a3ba7be89c6d5be85d9868.cc
oe937eef24f4685daa2d86c39e38bee34b.hk
t9824d95a91ac868deea12a247fa3cd55e.cn
```

1.2.3 Wordlist-based DGAs

Wordlist-based DGAs use a concatenation of words from some list as a generated domain. The wordlists are usually hard-coded, but the words can also be taken from a publicly accessible source (such as the American Declaration of Independence used by Gozi malware family [17]). The resulting domain names look a lot less random and more like made-up by a human, thus they are harder to detect.

DGA used by Suppobox malware family

The Suppobox DGA concatenates two words from a given wordlist (not listed here) and ".net" TLD to generate a domain.

```
def generate_domains(time_, word_list):
    with open("words{}.txt".format(word_list), "r") as r:
        words = [w.strip() for w in r.readlines()]

    if not time_:
        time_ = time.time()
    seed = int(time_) >> 9
    for c in range(85):
        nr = seed
        res = 16*[0]
        shuffle = [3, 9, 13, 6, 2, 4, 11, 7, 14, 1, 10, 5, 8, 12, 0]
        for i in range(15):
            res[shuffle[i]] = nr % 2
            nr = nr >> 1

        first_word_index = 0
        for i in range(7):
            first_word_index <<= 1
            first_word_index ^= res[i]

        second_word_index = 0
        for i in range(7,15):
            second_word_index <<= 1
            second_word_index ^= res[i]
        second_word_index += 0x80

        first_word = words[first_word_index]
        second_word = words[second_word_index]
        tld = ".net"
        print("{}{}{}".format(first_word, second_word, tld))
        seed += 1
```

Example 4: Python code of DGA of Suppobox. Reversed by Jason Geffner [16] and reimplemented by Johann Bader [3].

Examples of generated domains:

```
increaseinside.net
wouldinstead.net
rememberinstead.net
wouldexplain.net
rememberexplain.net
```

1.2.4 Permutation-based DGAs

Permutation-based DGAs generate all possible permutations of some initial domain name. Currently, there is only one known malware family using this type of DGA [30] and that is VolatileCedar:

DGA used by VolatileCedar malware family

This DGA takes some initial value and generates permutations of it.

```
domain_list = []
domain_list.append(initial_value)
current_domain = list(initial_value)

while True:
    for i in range(0, len(current_domain)-1):
        tmp = current_domain[i+1]
        current_domain[i+1] = current_domain[i+0]
        current_domain[i] = tmp
        domain_list.append("".join(current_domain))

    if current_domain == list(initial_value):
        break
```

Example 5: Python code of DGA of VolatileCedar. Reversed and reimplemented by Checkpoint [9].

Examples of generated domains:

```
xploreredotnte.info
oreredotntexpl.info
ntexploreredot.info
exploreredotnt.info
loreredotntexp.info
```

Chapter 2

Machine learning

In this chapter we provide a brief introduction to machine learning and describe classifying algorithms that we use in our tests.

Machine learning [1] studies algorithms that improve with experience. The algorithms learn from past experiences, they build (train) a mathematical model based on input (training) data and use it to make predictions on new (testing) data. Machine learning algorithms are used in a variety of areas, such as natural language processing, computer vision, email filtering, customer evaluation or cyber security.

Machine learning algorithms can be divided into three basic categories:

1. supervised learning - algorithms learn from a training set where all inputs and desired outputs are known, the goal is to find a mapping function (or a close approximation of it), which is then used for predicting the output from the input data

2. unsupervised learning - input data is unlabeled, the goal of the algorithms is to find a structure or patterns in the input data

3. reinforcement learning - algorithms are goal-oriented, they use software agents that learn in an interactive environment based on rewards and punishments

Common tasks, where machine learning is used, include:

- classification - an instance of supervised learning where a mapping function is approximated from input variables to a discrete output variable, i.e. input data is split into two or more categories (classes)

- regression - an instance of supervised learning where a mapping function is approximated from input variables to a continuous output variable

- clustering - an instance of unsupervised learning where input data is split into categories based on some measure of similarity or distance

Focus of this work is on supervised classifiers.

2.1 Classifiers

We use the following classifiers in our work: Naive Bayes, Random Forest, Gradient Boosting Classifier, Logistic Regression and Support Vector Machine.

2.1.1 Naive Bayes

Naive Bayes classifiers [24] are a collection of algorithms based on Bayes' theorem (or rule) in probability theory:

$$P(A|B) \;=\; \frac{P(B|A)P(A)}{P(B)}$$

$P(A|B)$ represents conditional probability, i.e. how likely is event A going to occur if event B is true, similarly for $P(B|A)$. $P(A)$ and $P(B)$ are marginal probabilities of events A and B. If the events A and B are independent, the conditional probability is equal to a product of the marginal probabilities of events A and B, i.e. $P(A|B) = P(A)P(B)$.

Now, given class variable y and vector of features $X = (x_1,\ \ldots,\ x_n)$ of size n, we can apply Bayes' theorem in the following way:

$$P(y|X) \;=\; \frac{P(X|y)P(y)}{P(X)} \tag{a}$$

The classifiers based on Bayes' theorem use a naive assumption that all features are independent of each other. Working with this assumption, we can rewrite the equation above in the following ways:

$$P(y|x_1,\ \ldots,\ x_n) \;=\; \frac{P(x_1,\ \ldots,\ x_n|y)P(y)}{P(x_1,\ \ldots,\ x_n)}$$

$$P(y|x_1,\ \ldots,\ x_n) \;=\; \frac{P(x_1|y)\ldots P(x_n|y)P(y)}{P(x_1)\ldots P(x_n)}$$

$$P(y|x_1,\ \ldots,\ x_n) \;=\; \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1)\ldots P(x_n)}$$

Because the values of the feature vector are known, the denominator is a constant for given input, so the left side of the equation and the numerator are proportional:

$$P(y|x_1,\ \ldots,\ x_n) \;\propto\; P(y)\prod_{i=1}^{n} P(x_i|y)$$

Now, we can calculate the probability for all possible values of y for given input and choose the value with the highest probability:

$$y \;=\; argmax_y P(y)\prod_{i=1}^{n} P(x_i|y)$$

The value $P(y)$ for some value of y can be estimated from the data in the data set - $P(y)$ = number of samples with class y / total number of samples. The differences between the Naive Bayes classifiers are regarding assumptions of the distribution of probabilities $P(x_i|y)$. For discrete features the most popular are Bernoulli NB and multinomial NB, for continuous features it is Gaussian NB:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

where the parameter μ is the mean of the Gaussian and the σ^2 is the variance. These parameters can be estimated with the training data, one way how to choose them is to maximize the likelihood of the model generating the data, this is called the Maximum Likelihood Estimate (MLE). The MLE of the mean and variance for the Gaussian distribution is:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

where n is the number of samples.

In our work we use the Gaussian Naive Bayes (GNB) classifier.

2.1.2 Random Forest

Random Forest (RF) classifier [7] is a decision tree-based model consisting of a collection of individual decision trees that operate as an ensemble. Ensemble learning methods use a combination of predictions of several base models in order to improve predictive performance, robustness and generalizability over a single model.

Decision tree

The decision tree model uses input variables to predict the value of a target variable. Each internal node in the decision tree is labeled with some input feature and each leaf is assigned a class, meaning that the data point has been assigned a particular class.

The splitting process is how a tree is built. The source data set is split into subsets based on rules that rely on classification features. Each edge in the tree is labeled with the possible value derived from its parent's split parameter.

Bootstrap aggregating

The random forest algorithm applies bootstrap aggregating (bagging) technique in training. For each tree that is to be built, a subset of the training data set is sampled

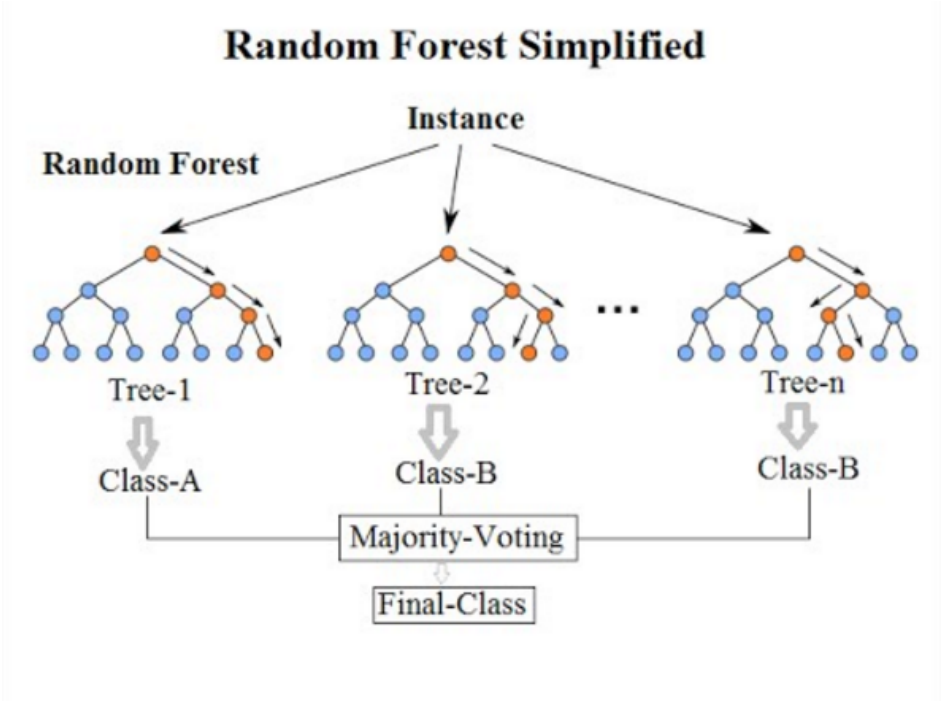

Figure 2.1: Random Forest with majority voting prediction.
Venkata Jagannath / CC BY-SA - `https://community.tibco.com/wiki/random-forest-template-tibco-spotfirer-wiki-page`

with replacement, of size of the training set. Then a classification tree is trained on this sampled data. This process is repeated a number of times. Random forests also use feature bragging technique where a random subset of features is used in each node splitting, the size of this subset is often chosen as square root of all features count.

There are two ways how to make a final prediction from the predictions of the individual decision trees. Either the decision trees vote for a single class and the majority vote is then taken as a final prediction or the decision trees output probability of classification and the average of these probabilities is taken as a final prediction.

Classification using random forest is illustrated in figure 2.1.

2.1.3 Gradient Boosting Classifier

Gradient Boosting Classifier (GBC) [15] is a model based on gradient boosting algorithm. Boosting algorithms in machine learning convert weak learners to stronger ones. Weak learner is defined as a classifier that can label samples only slightly better than by random guessing. Boosting technique is a sequential ensemble method, where subsequent learners learn from the mistakes made by previous learners.

One of the first algorithms to leverage boosting technique was AdaBoost [14]. It uses short decision trees as weak learners and assigns weights to instances during training. Instances that are hard to classify have more weight assigned. When new trees are

added, they are assigned these difficult instances and all weights are adjusted. The predictions are then made by the majority voting of the weak learners.

Gradient boosting also uses a set of decision trees as weak learners. The difference is in handling the weights of instances, in gradient boosting, when a new weak learners is added, the weights of previous trees are not changed. The goal of the gradient boosting classifier is to minimize the loss or the difference between the actual and predicted class value of a sample. For classification the logarithmic loss is usually used.

To minimize the loss function the gradient descent algorithm is used. Gradient descent is an optimization algorithm for finding a local minimum of a differentiable function. To do that we iteratively move in the direction of the steepest descent by using the negative of the gradient of the function. The gradient of a function at some point is a vector of partial derivatives of the function at this point.

2.1.4 Logistic Regression

Logistic regression [22] [19] is a classification model based on linear regression [23]. The linear regression model makes predictions based on a sum of the weighted average of input features and a constant (the bias term). This is the general formula for the linear regression:

$$y \;=\; w_0 + w_1 x_1 + \cdots + w_n x_n$$

The value y is the class label, n is the number of features, x_i is the value of the i-th feature. The value w_0 is the bias term and other w_i parameters are the weights of the features, which are calculated during the training of the model.

The logistic regression model uses the logistic function (or sigmoid function) to map output of the linear regression formula to interval between 0 and 1. The general logistic function is defined as:

$$\sigma(x) \;=\; \frac{1}{1 + e^{(-x)}} \;=\; \frac{e^x}{e^x + 1}$$

A part of graph is shown in figure 2.2.

Using sigmoid function on the linear regression formula we get:

$$P(y = 1) \;=\; \sigma(w_0 + \sum_{i=1}^{n} w_i x_i) \;=\; \frac{1}{1 + \exp\left(-(w_0 + \sum_{i=1}^{n} w_i x_i)\right)}$$

$$P(y = 0) \;=\; 1 - P(y = 1) \;=\; \frac{\exp\left(-(w_0 + \sum_{i=1}^{n} w_i x_i)\right)}{1 + \exp\left(-(w_0 + \sum_{i=1}^{n} w_i x_i)\right)}$$

Now we can set a decision boundary which we can use to predict a value of y. For example, if we set the decision boundary to 0.5, then y is predicted as 1 if $P(y = 1) > 0.5$ and 0 otherwise.

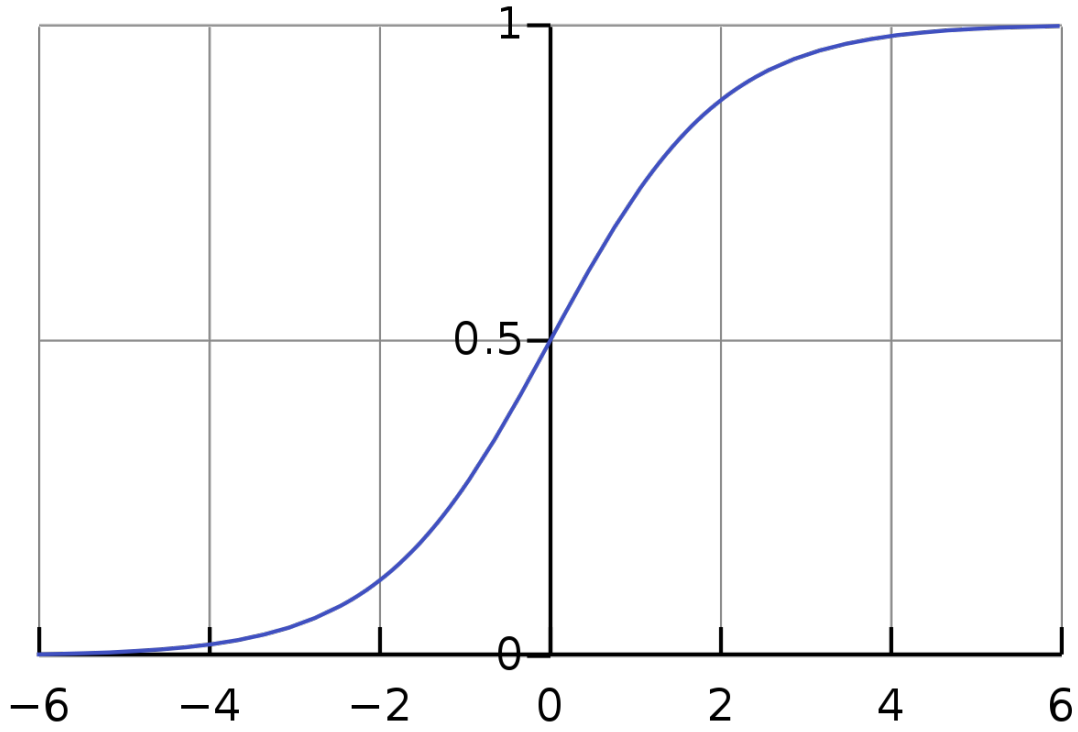

Figure 2.2: The sigmoid function $\sigma(x)$. Qef / Public domain - `https://commons.wikimedia.org/wiki/File:Logistic-curve.svg`

In our work we use logistic regression (LR) with SAGA solver [12].

2.1.5 Support Vector Machine

Support Vector Machine [10] is a supervised learning algorithm which can be used for classification or regression. When using SVM, data is interpreted as vectors in n-dimensional vector space where n is the number of features used. The goal of SVM is to construct a hyperplane in $(n-1)$-dimensional space that separates training data belonging to different classes. Next when making predictions, new data is mapped to the n-dimensional space and it is labeled based on a location relative to the hyperplane.

There could be many possible hyperplanes separating the data of different classes, the goal is to find a hyperplane with maximum margin, i.e. with the largest distance to two nearest data points of different classes. The data points closest to the hyperplane are called support vectors, the hyperplane is dependent only on them. The support vectors determine the position and orientation of the hyperplane.

The support vector machine as described above works only on linearly separable data. To classify non-linearly separated data it is necessary to use a function which maps lower-dimensional space into higher-dimensional space. This function is called kernel and the SVM model can be used with different kernels, for example with linear

kernel, polynomial or RBF (radial basis function) kernel.

In our work we use Support Vector Machine (SVM) with linear kernel.

The support vector machine model on two-dimensional space is illustrated in figure 2.3.

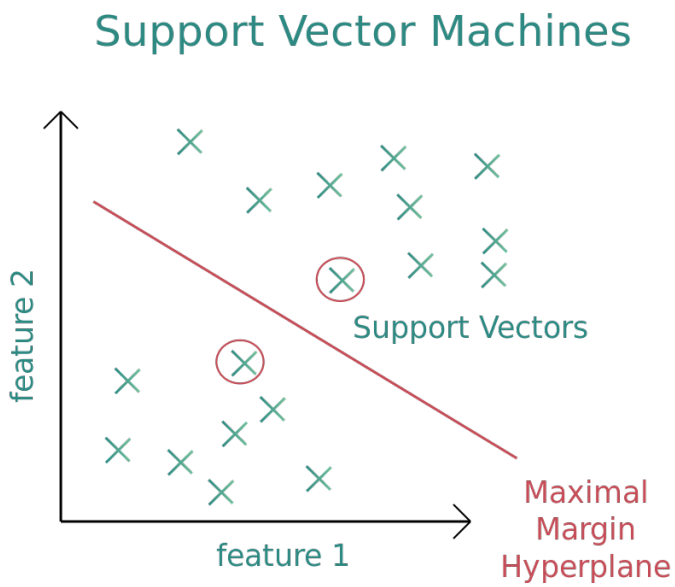

Figure 2.3: Support Vector Machine model. FreeSVG / CC0 - `https://freesvg.org/svm-support-vector-machines-diagram-vector-image`

Chapter 3

Detecting DGA domains

In this chapter we look at various ways of detecting domains generated by DGAs.

By detecting DGA domains we mean classifying input domains into two categories - clean (or benign) domains and domains generated by a DGA in some malware. Machine learning is a very popular method for this task, over the years there have been lots of supervised and unsupervised learning methods tested and used. These methods have been tested either with domain names only, or with some additional information such as DNS traffic data or WHOIS information. Aside from machine learning, graph-based or statistics-based approaches have sometimes been used for detection of DGA domains.

The next section provides an overview of related work.

3.1 Related work

Yadav et al. [35] proposed a methodology to make binary classification of domains based on statistical measures such as Kullback-Leibler divergence, Jaccard index or Levenshtein edit distance. They analyzed the distribution of alphanumeric characters in domains under the assumption that there is a significant difference between the distribution of alphanumeric characters of human-generated and algorithmically generated domains.

The statistical measures are used to measure a distance of the probability distribution of unigrams and bigrams of the domains that map to the same set of IP addresses. However, the downside of this approach is that the results might not be transferable to other malware families, which use different DGAs.

Antonakakis et al. [2] proposed a Pleiades system to detect DGA domains. This system uses a combination of clustering and classification algorithms. The authors assume that DNS response for the DGA domains is in majority of the cases Non-Existent Domain (NXDomain). They cluster similar domains based on lexical and

host-based features. The domains end up in the same cluster if they have similar string patterns and if they are queried by multiple sets of hosts.

Thus, the clusters represent different DGAs and are then classified in order to identify the malware family the DGA belongs to. Their classifier is based on a multi-class version of the Alternating Decision Trees (ADT) learning algorithm.

The authors tested Pleaides system on a real-world DNS traffic provided by large ISPs in North America and they were able to discover six brand new DGAs at the time.

Sivaguru et al. [33] evaluated various tree ensemble models based on human-engineered features and deep learning networks that learn features automatically. They focused on observation time and known seeds of the DGAs and select the training and testing data accordingly. Their goal was to test the robustness of the trained models and see how the models perform on domains generated at a different time or when the seed changes.

They used various kinds of Random Forest models: a binary RF classifier, a multi-class RF classifier and a one-versus-all RF model consisting of 15 binary RF classifiers, where each RF instance is trained on a dataset whose one half consists of domains of only one malware family and the other half is a mix of domains of other families and clean domains. For featureless approach they used various neural networks where each network consists of one or more of the following layers: an embedding layer, a LSTM layer or a CNN layer. All models were trained with no side information and with domain names only.

The authors found that all classifiers are more robust against changes in the seed of the time-dependent DGAs, compared to time-invariant DGAs.

Yu et al. [36] also evaluated deep neural networks - convolutional (CNN) and recurrent (LSTM) neural networks. They used real live traffic domains to train the models and they proposed a novel criteria for building a dataset for training from the real data. They also compared the performance of the neural networks with feature-based classifiers such as SVM or AdaBoost. By setting a threshold on false positive rate to 0.01% they found that the best performing model is the CNN.

The authors also noted that malware families which use wordlist-based DGAs are very hard to detect. An overview of the methods that are more successful in detecting those kinds of DGAs is provided in the next section.

Detecting wordlist-based DGAs

Curtin et al. [11] introduced a score that measures how much a domain is similar to English words, they call it the smashword score. This score is calculated as an average n-gram overlap with words from the English dictionary, for $n = 3, \ldots, 5$.

They provided a machine learning model based on recurrent neural networks. During training, a side information is also considered, if it is available. They used information from WHOIS database such as registrar name, contact email and other contact information, information about when the domain was created, updated or when it expires.

The authors conclude that the combination of DGA domains, their smashword score and the side WHOIS information as the training data provides very good results for their proposed model. Using this model, they were able to detect many malware families, which use wordlist-based DGAs (like suppobox - see section 1.2.3), better than other models they had tested.

Patsakis and Casino [26] proposed a probabilistic approach to detect wordlist-based DGAs. They exploited the fact that these DGAs use wordlists that are limited in size, which results in word repetitions in the generated domains. They collected NXDomains and split them into words, which are then collected in buckets either statistically or based on a specific pattern. Then a threshold derived from the birthday paradox is set on the number of words in buckets. Once some bucket reaches the threshold, an alert is raised, meaning that the bucket may be a part of the DGA wordlist. The authors claim only 3 to 27 NXDomains queries are necessary to detect DGA malware with high confidence.

Pereira et al. [28] used graph-based approach to detect wordlist-based DGAs. They designed a system called WordGraph that can extract dictionaries from the DGAs using only DNS traffic data. The core of their system is a graph, where each vertex is a word and edges connect words that appear together in some domain name. Then a decision tree model is trained with features computed from connected components of the constructed graph.

Their system performed significantly better in detecting wordlist-based DGAs then the RF and CNN models. Testing on real traffic, the authors were able to detect DGAs used by known malware families and also discover new DGAs used by previously unknown malware families.

Chapter 4

Experiments

This chapter provides an overview of experiments we do, the methodology we have chosen and the data we use. Details of the experiments implementation are in appendix A.

4.1 Data

In this section we describe sources of data that we use to build a dataset used in our experiments. We used DGArchive to get malicious domains and TRANCO list to get clean domains.

4.1.1 DGA domains

We use DGArchive [29] [30] as a source of malicious domains. This archive contains AGDs of 86 malware families divided by the seeds used during their generation. The domains can be downloaded based on a date of generation. As of 19th April 2020, there were 86,299,935 unique domains.

We downloaded all available domains from years 2017, 2018 and 2019 and from January 2020. We grouped the domains by malware families, ignoring the seeds in the process, and we removed all duplicates. This way, we obtained 49,745,510 unique domains. In section 4.1.3 we describe what malware families we have chosen for our dataset.

4.1.2 Clean domains

We have chosen a TRANCO [20] as a list of clean domains. This list is composed of four lists of the most popular domains and its authors show how these individual lists can be manipulated or skewed. The authors came up with a way to improve the results

and introduced a new list aggregated by the most popular domains - a TRANCO list. The four sources of domains that are the basis of the TRANCO list are:

- Alexa[1] - this is a list of one million popular domains ranked by Amazon, it is updated daily. The ranks are based on a global traffic data, more specifically, on a proprietary measure of unique visitors and page views.
- Cisco Umbrella[2] - this list also consists of one million domains and it is also updated daily. The domains are taken from Cisco's DNS resolvers - the domains are ranked by the number of unique of IPs issuing DNS queries for them.
- Majestic[3] - this daily updated list consists of one million domains, which are based on backlinks to websites obtained by a crawl of hundreds of billions of websites over a several weeks time frame.
- Quantcast[4] - this is a list of the most visited websites in the USA, it is based on the number of people visiting a website during the previous month of its operation. The list consists of tens of thousands of websites every day. Non-US websites are tracked directly by a tracking script or by data from Internet Service Providers and toolbar providers.

The lists are then combined using the Dowdall rule - the first domain gets 1 point, the second $\frac{1}{2}$, ..., and the last $\frac{1}{N}$ points and unranked domains get 0 points. Also, to improve the stability, the combined list is aggregated from the individual lists of the past 30 days.

For our work we have used the TRANCO list from 1st March 2020, which aggregates the ranks from the lists by Alexa, Umbrella, Majestic and Quantcast from 31st January 2020 to 29th February.

4.1.3 Dataset building

The dataset used for our experiments is built from subsets of AGDs of some chosen malware families and the whole clean domain set. Next, we describe what malware families we have chosen.

We filter the malware families based on the DGAs they use. We use only domains generated by arithmetic-based and hash-based DGAs to build our dataset. There are two main reasons for this. First, a vast majority of known DGAs are arithmetic-based or hash-based, there are only few malware families that use wordlist-based or

[1] `https://s3.amazonaws.com/alexa-static/top-1m.csv.zip`
[2] `https://s3-us-west-1.amazonaws.com/umbrella-static/top-1m.csv.zip`
[3] `http://downloads.majestic.com/majestic_million.csv`
[4] `https://ak.quantcast.com/quantcast-top-sites.zip`

permutation-based DGAs. Second, as we described in section 3.1, there are different, more appropriate ways to detect domains generated by these kinds of DGAs. So we believe our approach and the algorithms we use would result in classifying these types of domains as clean.

Note that we do not distinguish between seeds used in any of the DGAs. We also do not use malware families that generate third-level domains. Again, a vast majority of malware families generates only a second-level domain which is then concatenated with a TLD.

The remaining malware families that we use to build our dataset are listed in a table 4.1, there are names of the families as used by DGArchive, DGA types (A - arithmetic-based or H - hash-based), examples of generated domains and the number of unique domains available to us.

We have used randomly chosen 30,000 domains of each malware family or all domains of a malware family, if there are less than 30,000 unique domains available, resulting in 1,008,828 domains to be included in the dataset.

To sum it up, our dataset contains 2,008,828 domains - one million clean and the rest are DGA domains. This dataset is used in every experiment during the training phase.

4.1.4 Real-world data

We have also tested the models on real-world data provided by cyber security company ESET[5]. We use following collections of data:

- random domains collected on 9$^{\text{th}}$ and 14$^{\text{th}}$ April 2020, total of 1,004,841 domains
- NXDomains (non-existent domains) collected from Whalebone Passive DNS from the first 20 days of April 2020, total of 3,204,821 domains
- Authlist consisting of clean domains generated by ESET and collected on 22$^{\text{nd}}$ April 2020, total of 75,076 domains

For every collection we removed duplicates and kept only the domains of form "sld.tld". We also removed all domains with their SLDs shorter than 5 characters.

[5] `https://www.eset.com/`

Malware family	DGA type	Domain example	Note	Number of unique domains
Bamital	H	873c174ca173b5393e93f9571e8a293b.info		58,552
Bedep	A	yftwlzxtpozg.com		7,458
Blackhole	A	wevydrkvywxqfsul.ru		730
CCleaner	A	ab6d54340c1a.com	DGA found in the backdoored version of CCleaner (5.33)	37
Chinad	A	q60coxnn83zj7i9u.org		288,256
Chir	H	f661e398d876c6f7.cn		100
Conficker	A	kfoqmgax.com		562,962
Corebot	A	c2i032c6o4mhs45vcxgluvo.sg		117,000
Cryptolocker	A	xhlwkqdawjdpi.info		964,982
Diamondfox	A	ddcuhr7.com		474
DirCrypt	A	vlbqryjd.com		1,150
Dmsniff	A	snkrpmnq.net		70
Dyre	H	cdca364b71f0c8506d60eb2939f4b806d9.to		1,126,000
Ebury	A	m9e8t4o6mau3h.biz		2,000
EKforward	A	bd7d817a.eu		1,126
Emotet	A	fvpuplocfrdcuqon.eu		216,168
Feodo	A	xvmzegestulhtvqz.ru		192
Fobber	A	btpnxlsfdqbhzazyx.net		2,000
Gameover	A	99kw7y18sz2ee19xycgb1eckfvd.biz		12,571,000
GozNym	A	toyvsgu.com		332
Gspy	A	9c3b4fe3fba848a3.net		49
Hesperbot	A	iksjxihh.com		178
Infy	H	dce022a0.space		9,660
Locky	A	cbkmotlvy.yt		717,348
MadMax	A	pg0tndvnuq.org		148
Makloader	A	cdpvzekauvhtgrbzhakassjwlummntqseswncnfd.pro		512
Mirai	A	xpknpxmywqsr.online		280
Modpack	A	k1y5u25h.ru		106
Monerominer	A	c0ccdd790a0d2.blackfriday		1,898,700
Murofet	A	vpevhtorzutawui.info		8,512,560
Murofetweekly	A	oui55ixeybyetoyaymun30krlvaxmrp62lt.net		185,000
Mydoom	A	srmseesrwh.biz		2,200
Necurs	A	caxadsjuygrem.ac		6,076,640
Nymaim	A	pnrbassntqm.net		295,084
Oderoor	A	grmcsvspngjj.com		13,389
Omexo	H	35262768764bd6c908c386b532a3dc2f.net		20
Padcrypt	A	mdfeedbdfbfdcabo.info		177,233

Pandabanker	H	28f46950ab54.net		25,364
Proslikefan	A	udahqhqz.ru		146,379
Pushdo	A	cumocuwupjo.com		207,341
Pushdotid	A	ahujctulsb.org		6,000
Pykspa	A	zzfhnetq.info		912,118
Pykspa2	A	kiseccuiwcymao.net		1,442
Pykspa2s	A	tmrvuifox.com		9,960
Qadars	A	v8l6bshunstq.net		324,000
QakBot	A	hluvupofr.net		2,220,000
Ramdo	A	ocqiwseygwqyeuma.org		6,000
Ramnit	A	egopuefrdsefc.com		19,779
Ranbyus	A	gnajdybsuaimhw.me		635,640
Rovnix	A	nn4rzw6r4yv4ezapuu.ru		1,900
Shifu	A	dxnrlqj.eu		2,331
Simda	A	cihunemyror.eu		16,474
Sisron	A	mjuwntiwmtya.com		4,540
Sphinx	A	qeygqpabwinmaoxn.com		134,822
Sutra	A	jpcwcfwiprwifrei.info		738
Szribi	A	tutuitqf.com		5,298
Tempedreve	A	sxilgdils.com		204
TempedreveTDD	A	wrqzuhirg.org		1,126
Tinba	A	dlcbjsrrewrw.me		106,756
TinyNuke	H	ad7a09dc439d2667296b0737abd3c131.xyz		109,184
Tofsee	A	dqhdqhb.ch		3,240
Torpig	A	xgrrunj.net		14,418
UD2	H	5dc52635adcb3d650d90.info	Unknown DGA	380
UD3	A	jyrewq987541.ga	Unknown DGA	60
UD4	A	snfrpmnq.org	Unknown DGA	70
Urlzone	A	e3oa4wglvd21xa.com		32,020
Vawtrak	A	fonizwhgnqp.ru		2,700
Vidro	A	fdmguoewikd.com		32,400
Vidrotid	A	nelazucapqv.com		200
Virut	A	uiaiub.com		10,882,059
WD	H	wd7bdb20e4d622f6569f3e8503138c859d.win		66,344
Xshellghost	A	vqjmpkhmpahuz.com		37
XxHex	A	xxfddc1b01.at		4,400
Total				49,745,510

Table 4.1: Malware families used in building of our dataset

4.2 Types of experiments

In this section we describe the experiments that we perform to classifying algorithms detailed above in detection of AGDs. We use two cross-validation techniques - K-Fold and Leave One Group Out.

4.2.1 K-Fold

In basic usage of machine learning, the data is divided into a training set and a testing set. The training set is used to train the model and the testing set is then used to validate the trained model, i.e. to test how well it performs on unknown data.

This basic approach has a number of drawbacks. Some data is never used to train the model and also, the way the dataset is split into training and testing sets is very important. Different splits can lead to selection bias or an overfitting (the model performs very well on data from the training set, but poorly on unseen data, it doesn't generalize well).

To overcome this, cross-validation techniques can be used. One of them is a k-fold cross-validation. In this technique, the dataset is split into k equal-sized (if possible) subsets (folds). There are also k iterations of training and testing, in each a different fold is taken as a testing set and the remaining $k - 1$ folds are used as a training set. The results are then combined together, usually averaged.

By using k-fold cross-validation, we test how well the models can predict unknown domains of known malware families. In our experiments we use a value of $k = 5$, i.e. our dataset is split into 5 folds and there are 5 iterations of training and testing on one model.

The k-fold cross-validation technique is illustrated in figure 4.1.

4.2.2 Leave One Group Out (LOGO)

Next cross-validation technique we use is Leave One Group Out (LOGO). It is similar to the k-fold cross-validation, but in each iteration we leave out one group of data, this left out group acts as a testing set and other groups are a training set. In our case we leave out domains of one malware family.

As a validation set we take all domains of the left out malware family, not just domains left out in our dataset. By performing the LOGO cross-validation, we test how well the models do in a situation where a new malware family or a new DGA of a known malware appears.

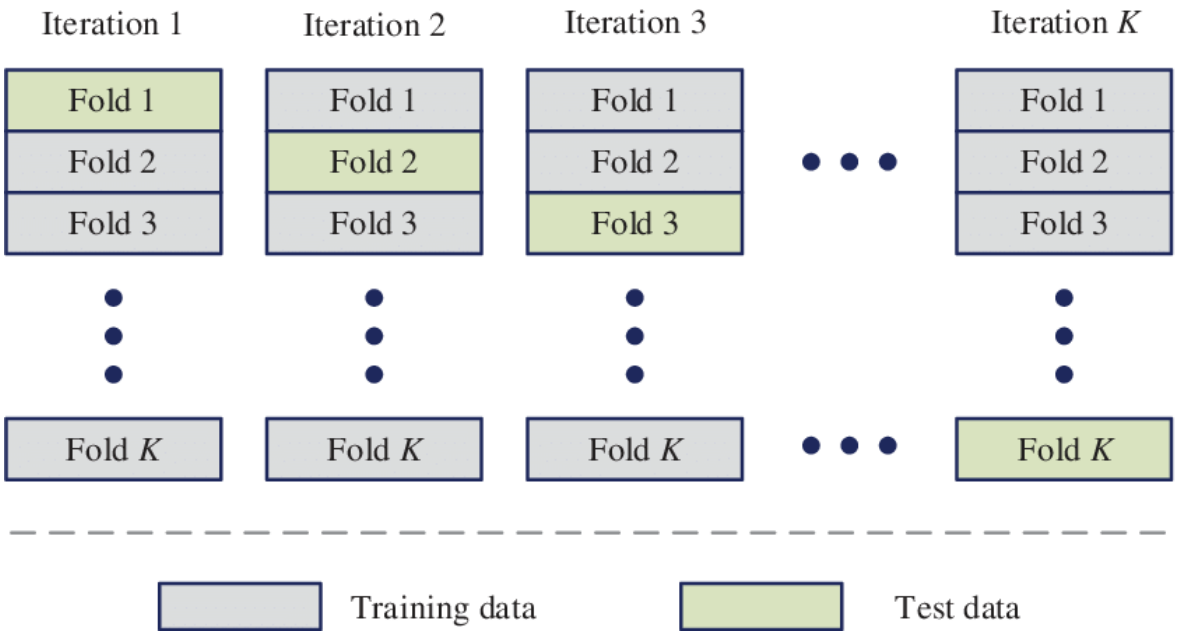

Figure 4.1: K-Fold cross-validation technique. Qiubing Ren, Mingchao Li & Shuai Han (2019) Tectonic discrimination of olivine in basalt using data mining techniques based on major elements: a comparative study from multiple perspectives, Big Earth Data, 3:1, 8-25, DOI: 10.1080/20964471.2019.1572452

4.2.3 Processing results

We use this methodology to compare the results of our experiments. First, we observe number of correctly and incorrectly classified domains:

- True Positive (TP) - number of malicious domains correctly predicted as AGDs
- False Positive (FP) - number of clean domains incorrectly predicted as AGDs
- True Negative (TN) - number of clean domains correctly rejected as AGDs, i.e. correctly identified as clean
- False Negative (FN) - number of malicious domains incorrectly identified as clean

Based on these numbers we compute following metrics:

- Accuracy - ACC = $\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$
- True Positive Rate - TPR = $\frac{\text{TP}}{\text{TP+FN}}$
- True Negative Rate - TNR = $\frac{\text{TN}}{\text{TN+FP}}$
- False Positive Rate - FPR = $\frac{\text{FP}}{\text{FP+TN}} = 1 - \text{TNR}$
- False Negative Rate - FNR = $\frac{\text{FN}}{\text{FN+TP}} = 1 - \text{TPR}$

Next, for these rates we compute mean, median, minimum, maximum and standard deviation.

The important rates are TPR and FPR. The TPR shows us the rate of correctly identified DGA domains and the FPR shows us the rate of clean domains incorrectly identified as DGA domains. If some trained model is used in real-time blocking of DGA domains, high FPR can cause many clean domains to be blocked and the user experience would be very bad. Because of this and because users encounter significantly more clean domains than DGA domains, the FPR should be as low as possible.

4.2.4 Real-world data predictions

For the two best sets of features in terms of DGA domains detection we do predictions on real-world data with all models. Although the ground truth is lacking for this kind of data, we expect similar results as in previous experiments, i.e. the best performing model in previous experiments should also predict the largest number of AGDs among this real-world data.

From the other perspective, we expect that the largest number of AGDs will be detected among the NXDomains and the domains from Authlist should not be detected as AGDs at all.

Chapter 5

Features

In this chapter we describe what features we extract from the domains and list subsets of features used in our experiments.

5.1 Used features

We extract a number of features from domain names (second-level domains) or top-level domains.

We compute the following 14 features that were used in other work regarding DGA detection:

- domain length - length of the domain name (second-level domain) [36]
- TLD length - length of the top-level domain [33]
- TLD hash - CRC32 hash of the top-level domain normalized to a value between 0 and 1 [36]
- is first character digit - Boolean flag, 1 if the first character of the domain name is a digit, 0 otherwise [36]
- number of digits in the domain name [33]
- number of unique characters in the domain name [33]
- vowel ratio - number of vowels in the domain name divided by the domain name length [36]
- consonant ratio - number of consonants in the domain name divided by the domain name length [33]
- hex character ratio - number of hexadecimal characters (0-9 and a-f) in the domain name divided by the domain name length [36]

- digit ratio - number of digits in the domain name divided by the domain name length [33]
- longest consonant sequence in the domain name [33]
- Shannon entropy of the domain name [36]

$$\text{ent} = \frac{-\sum_x D(x) \log D(x)}{\log len(\text{domain})},$$

where $D(x)$ is a distribution of characters.

- Gini index of characters of the domain name [36]

$$\text{gni} = 1 - \sum_x D^2(x)$$

- classification error of characters of the domain name [36]

$$\text{cer} = 1 - \max\{D(x)\}$$

We have come up with another four features:

- longest vowel sequence in the domain name
- longest digit sequence in the domain name
- digit to letter ratio - number of digits in the domain name divided by the number of letters in the domain name
- is MD5 like - Boolean flag, 1 if the domain looks like an MD5 hash, i.e. is 32 characters long and contains only hexadecimal characters, 0 otherwise

We also compute a number of n-gram features [2], for values of $n = 2, ..., 5$. First, we compute the frequency of every n-gram in the TRANCO domain list from the March 1st (same list as in section 4.1.2) and save the frequency values in a look-up dictionary. Next, we look up n-gram frequencies for every n-gram in the domain that we extract the features from. We get a list of frequencies on which we compute the average, median and standard deviation. Finally, we normalize the results by computing the decadic logarithm of them, thus acquiring another 12 features.

Examples of extracted features can be seen in table 5.1. There are extracted features from a clean domain (google.com) and a DGA domain (18ygxbfvc2eov17k.net) of Chinad malware family.

domain	google.com	18ygxbfvc2eov17k.net
domain length	6	16
TLD length	3	3
TLD hash	0.393414	0.948884
is first character digit	0	1
number of digits	0	5
number of unique characters	4	14
vowel ratio	0.5	0.1875
consonant ratio	0.5	0.5
hex character ratio	0.166667	0.5625
digit ratio	0	0.3125
digit to letter ratio	0	0.454545
longest consonant sequence	2	6
longest vowel sequence	2	2
longest digit sequence	0	2
is MD5 like	0	0
Shannon entropy	1.918296	3.75
Gini coefficient of characters	0.722222	0.921875
classification error of characters	0.666667	0.875
2-gram average	4.503576	3.432103
2-gram median	4.319980	2.98945
2-gram standard deviation	4.355612	3.668874
3-gram average	3.219519	1.278754
3-gram median	3.179552	1
3-gram standard deviation	2.976139	1.553702
4-gram average	2.677607	0
4-gram median	2.668386	0
4-gram standard deviation	1.192505	0
5-gram average	2.644439	0
5-gram median	2.644439	0
5-gram standard deviation	0.845098	0
DGA flag	0	1

Table 5.1: Examples of extracted features from a clean domain (google.com) and a DGA domain (18ygxbfvc2eov17k.net)

5.2 Feature subsets

We use different subsets of computed features when performing experiments. By using a smaller set of features we can make the trained models smaller, improve prediction times and, ideally, at the same time maintain the accuracy at the same level. Computing n-gram features is more expensive than computing other features, so we try to train the models with all features except the n-gram features and on the other hand, with n-gram features only and compare the results. Also, 46 of 73 malware families do not use digits in their DGA domains. Therefore, we try to train the models with all features except digit features and see if the overall accuracy changes. All subsets are summarized below.

All features

This subset contains all 30 extracted features described above.

Best features from statistical tests

We did three univariate statistical tests to select the best features:

- chi-squared test
- ANOVA F-test
- mutual information test

For each test and resulting scores, we used Borda count to make rankings of the features, i.e. feature with the highest score was assigned 30 points, with the second highest score 29 points and so on. Then we computed an average of the three rankings and took the first 20 features to make one feature set.

The features are: domain length, number of digits, number of unique characters, vowel ratio, digit ratio, longest consonant sequence, longest digit sequence, Shannon entropy and the average, median and standard deviation of n-grams of the domain, for $n = 2, ..., 5$.

All features except digits features

This subset contains all features except those that involve digits, so these features are left out: is first character digit, number of digits, digit ratio, digit to letter ratio, longest digit sequence.

All features except n-grams features

This subset contains all features except those involving n-grams, the features are: domain length, TLD length, TLD hash, is first character digit, number of digits, number of unique characters, vowel ratio, consonant ratio, hex character ratio, digit ratio, digit to letter ratio, longest consonant sequence, longest vowel sequence, longest digit sequence, is MD5 like, Shannon entropy, Gini index of characters, classification error of characters.

Only n-grams features

This subset contains only features involving n-grams, so the average, median and standard deviation of the domain n-grams, for $n = 2, ..., 5$.

Chapter 6

Results

In this chapter, the results of experiments are discussed. There is also an analysis of domains of malware families which are hard to detect.

6.1 K-Fold

First, we take a look at the results of k-fold experiments.

All features

The best performing model when trained with all features is the Random Forest achieving more than 99% accuracy. It can detect the most DGA domains, the true positive rate is over 98% while the FPR is only 0.15%. The Gradient Boosting Classifier, Logistic Regression and Support Vector Classifier models all perform a bit worse than the RF, achieving accuracy from 96.9 to 98.3 %. The LR and SVC models have worse false positive rates at about 2.5%. On the other hand, the Gaussian Naive Bayes model performs significantly worse with accuracy just above 86% and true positive rate only at 80%, also the false positive rate is at 8%. All results are summarized in table 6.1.

Best features from statistical tests

For this set of features the performance of the trained models perform slightly worse the models trained with all features. The results for the GBC and RF models are about 0.5% worse for accuracy, TPR and FPR. The LR and SVC models also perform worse. The exception is the GNB model, whose accuracy is better by 2% and the TPR by 7%, but the FPR is worse by 3%. The results are summarized in table 6.2.

Gaussian Naive Bayes					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.861829	0.804675	0.195324	0.919487	0.080512
Median	0.861857	0.804379	0.195620	0.919781	0.080218
Min	0.861035	0.802678	0.193478	0.918715	0.080082
Max	0.862760	0.806521	0.197321	0.919917	0.081284
Std	0.000687	0.001433	0.001433	0.000479	0.000479
Gradient Boosting Classifier					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.983142	0.973616	0.026383	0.992752	0.007247
Median	0.983064	0.973817	0.026182	0.992575	0.007424
Min	0.982778	0.972698	0.025310	0.992262	0.006697
Max	0.983527	0.974689	0.027301	0.993302	0.007737
Std	0.000288	0.000691	0.000691	0.000437	0.000437
Logistic Regression					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.969601	0.964417	0.035582	0.974832	0.025167
Median	0.969619	0.964213	0.035786	0.974699	0.025300
Min	0.969330	0.963773	0.035039	0.974312	0.024661
Max	0.969788	0.964960	0.036226	0.975338	0.025687
Std	0.000163	0.000461	0.000461	0.000423	0.000423
Random Forests					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.991807	0.985161	0.014838	0.998510	0.001489
Median	0.991845	0.985312	0.014687	0.998491	0.001508
Min	0.991562	0.984678	0.014567	0.998347	0.001315
Max	0.991943	0.985432	0.015321	0.998684	0.001652
Std	0.000128	0.000272	0.000272	0.000119	0.000119
Support Vector Classifier					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.969891	0.963943	0.036056	0.975893	0.024106
Median	0.969858	0.963929	0.036070	0.975895	0.024104
Min	0.969723	0.963342	0.035449	0.975377	0.023530
Max	0.970189	0.964550	0.036657	0.976469	0.024622
Std	0.000157	0.000469	0.000469	0.000396	0.000396

Table 6.1: Summary of results - all features

Gaussian Naive Bayes					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.883285	0.876898	0.123101	0.889729	0.110270
Median	0.883275	0.876658	0.123341	0.889726	0.110273
Min	0.882660	0.876208	0.122147	0.889154	0.109596
Max	0.883763	0.877852	0.123791	0.890403	0.110845
Std	0.000379	0.000683	0.000683	0.000450	0.000450
Gradient Boosting Classifier					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.978436	0.968841	0.031158	0.988115	0.011884
Median	0.978298	0.968899	0.031100	0.988038	0.011961
Min	0.978151	0.967973	0.030333	0.987771	0.011567
Max	0.978937	0.969666	0.032026	0.988432	0.012228
Std	0.000279	0.000560	0.000560	0.000235	0.000235
Logistic Regression					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.951298	0.951657	0.048342	0.950936	0.049063
Median	0.951531	0.951808	0.048191	0.950915	0.049084
Min	0.950640	0.950945	0.047632	0.950332	0.048161
Max	0.951645	0.952367	0.049054	0.951838	0.049667
Std	0.000379	0.000488	0.000488	0.000555	0.000555
Random Forests					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.987992	0.979911	0.020088	0.996145	0.003854
Median	0.987988	0.979982	0.020017	0.996122	0.003877
Min	0.987866	0.979692	0.019981	0.996053	0.003702
Max	0.988109	0.980018	0.020307	0.996297	0.003946
Std	0.000078	0.000120	0.000120	0.000081	0.000081
Support Vector Classifier					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.951092	0.951794	0.048205	0.950384	0.049615
Median	0.951394	0.951852	0.048147	0.950359	0.049640
Min	0.950344	0.951034	0.047582	0.949647	0.048805
Max	0.951439	0.952417	0.048965	0.951194	0.050352
Std	0.000426	0.000451	0.000451	0.000567	0.000567

Table 6.2: Summary of results - best features from statistical tests

All features except digits features

These results are a bit better compared to results for models trained using all features except for LR and SVC models. The RF model trained using this set of features performs the best across all experiments, achieving 99.2% accuracy, more than a 98.5% TPR and less than 0.15% FPR. The summary of results is in table 6.3.

All features except n-grams features

Models trained with this set of features perform a lot worse, the accuracy and the TPR is much lower than in other experiments. The TPR of the GNB model is even less than 50%, only 45% of DGA domains are detected, on the other hand the FPR is only 4% which is the best for the GNB model across all k-fold experiments. The results summary is in table 6.4.

Only n-grams features

Using this set of features, all models achieve more than 92% TPR, but the GNB, LR and SVC models also have FPR over 10%. Again, the best is the RF model with 98.4% accuracy. The summary is in table 6.5.

Summary

Based on these experiments we can see that the best performing classifier across all feature sets is the Random Forest followed by Gradient Boosting Classifier. Logistic Regression and Support Vector Classifier perform a bit worse and their results are very similar. Finally, the Gaussian Naive Bayes classifier is considerably less effective in detecting DGA domains.

For all experiments, the standard deviation of the results is very low, the mean and median are very close to each other, meaning that in one k-fold experiment, all runs produce very similar models in terms of predictions after training.

Gaussian Naive Bayes					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.895809	0.882240	0.117759	0.909499	0.090500
Median	0.895757	0.881815	0.118184	0.909482	0.090517
Min	0.895262	0.881135	0.116747	0.908855	0.090042
Max	0.896536	0.883252	0.118864	0.909957	0.091144
Std	0.000440	0.000846	0.000846	0.000390	0.000390
Gradient Boosting Classifier					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.983362	0.973969	0.026030	0.992838	0.007161
Median	0.983360	0.974056	0.025943	0.992768	0.007231
Min	0.982977	0.973070	0.025127	0.992738	0.007014
Max	0.983766	0.974872	0.026929	0.992985	0.007261
Std	0.000293	0.000622	0.000622	0.000105	0.000105
Logistic Regression					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.957886	0.956031	0.043968	0.959757	0.040242
Median	0.958072	0.955866	0.044133	0.959678	0.040321
Min	0.957363	0.955204	0.043277	0.959439	0.039565
Max	0.958129	0.956722	0.044795	0.960434	0.040560
Std	0.000292	0.000556	0.000556	0.000350	0.000350
Random Forests					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.992165	0.985813	0.014186	0.998572	0.001427
Median	0.992132	0.985783	0.014216	0.998551	0.001448
Min	0.992042	0.985522	0.013884	0.998437	0.001315
Max	0.992341	0.986115	0.014477	0.998684	0.001562
Std	0.000111	0.000243	0.000243	0.000090	0.000090
Support Vector Classifier					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.957972	0.955549	0.044450	0.960417	0.039582
Median	0.958184	0.955256	0.044743	0.960234	0.039765
Min	0.957360	0.954516	0.043614	0.960105	0.038860
Max	0.958234	0.956385	0.045483	0.961139	0.039894
Std	0.000337	0.000715	0.000715	0.000381	0.000381

Table 6.3: Summary of results - all features except digits features

Gaussian Naive Bayes					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.703719	0.450309	0.549690	0.959366	0.040633
Median	0.703652	0.450014	0.549985	0.959258	0.040741
Min	0.703020	0.449434	0.547871	0.958979	0.039945
Max	0.705101	0.452128	0.550565	0.960054	0.041020
Std	0.000753	0.000973	0.000973	0.000401	0.000401
Gradient Boosting Classifier					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.890158	0.859632	0.140367	0.920953	0.079046
Median	0.890102	0.860276	0.139723	0.921005	0.078994
Min	0.889134	0.857473	0.138739	0.920105	0.078345
Max	0.891130	0.861260	0.142526	0.921654	0.079894
Std	0.000645	0.001321	0.001321	0.000514	0.000514
Logistic Regression					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.838423	0.804833	0.195166	0.872310	0.127689
Median	0.838629	0.804939	0.195060	0.872508	0.127491
Min	0.837577	0.804259	0.194751	0.870654	0.126827
Max	0.838883	0.805248	0.195740	0.873172	0.129345
Std	0.000491	0.000328	0.000328	0.000892	0.000892
Random Forests					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.916170	0.897558	0.102441	0.934948	0.065051
Median	0.916214	0.897369	0.102630	0.935302	0.064697
Min	0.915988	0.897216	0.101543	0.933840	0.064515
Max	0.916286	0.898456	0.102783	0.935484	0.066159
Std	0.000117	0.000454	0.000454	0.000607	0.000607
Support Vector Classifier					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.838507	0.799578	0.200421	0.877781	0.122218
Median	0.838609	0.799579	0.200420	0.877945	0.122054
Min	0.837522	0.799000	0.199894	0.876330	0.121348
Max	0.839180	0.800105	0.200999	0.878651	0.123669
Std	0.000559	0.000367	0.000367	0.000780	0.000780

Table 6.4: Summary of results - all features except n-grams features

Gaussian Naive Bayes					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.894744	0.920795	0.079204	0.868464	0.131535
Median	0.894709	0.921143	0.078856	0.868346	0.131653
Min	0.894216	0.919696	0.078417	0.867780	0.130968
Max	0.895404	0.921582	0.080303	0.869031	0.132219
Std	0.000401	0.000748	0.000748	0.000494	0.000494
Gradient Boosting Classifier					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.962803	0.964849	0.035150	0.960738	0.039261
Median	0.962918	0.964955	0.035044	0.960636	0.039363
Min	0.962147	0.964085	0.034568	0.960189	0.038413
Max	0.963364	0.965431	0.035914	0.961586	0.039810
Std	0.000421	0.000459	0.000459	0.000480	0.000480
Logistic Regression					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.922957	0.947909	0.052090	0.897786	0.102213
Median	0.922922	0.948096	0.051903	0.897863	0.102136
Min	0.922439	0.946769	0.051540	0.897196	0.101569
Max	0.923393	0.948459	0.053230	0.898430	0.102803
Std	0.000340	0.000586	0.000586	0.000480	0.000480
Random Forests					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.984127	0.976396	0.023603	0.991926	0.008073
Median	0.984139	0.976584	0.023415	0.991978	0.008021
Min	0.983998	0.975863	0.023340	0.991464	0.007785
Max	0.984289	0.976659	0.024136	0.992214	0.008535
Std	0.000104	0.000310	0.000310	0.000278	0.000278
Support Vector Classifier					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.920948	0.952934	0.047065	0.888680	0.111319
Median	0.920832	0.953175	0.046824	0.888627	0.111372
Min	0.920299	0.951653	0.046541	0.887962	0.110453
Max	0.921538	0.953458	0.048346	0.889546	0.112037
Std	0.000436	0.000649	0.000649	0.000620	0.000620

Table 6.5: Summary of results - only n-grams features

6.2 Leave One Group Out

Second, we look at the results of LOGO experiments. The results for individual malware families are listed in appendix B.

All features

The two best performing models are the RF and GBC models with the mean of accuracy of almost 98.9% and 98.1% respectively and the median of accuracy of almost 99.8% and 99.3% respectively. The LR and SVM models perform very similarly, about 2 percentage points worse than the RF model. The FPRs of these models range between 0.15% and 2.5%. The worst performing is the GNB model, for some malware families the accuracy drops below 50% and the TPR drops down to only 10%. An analysis of the malware families whose AGDs are hard to detect is in section 6.2.1. The summary of results is in table 6.6.

Best features from statistical tests

Models trained with this set of features perform a bit worse, up to 2 percentage points, than the models trained with all features. There are still some malware families which are hard to detect. All results are in table 6.7.

All features except digits features

The results for this set of features are comparable to the results of models trained with all features, except for the LR and SVC models whose performance is worse by 1.5 p.p. The RF model here performs the best of all LOGO experiments - its mean of accuracy is 98.9% and the median is over 99.8%, the FPR is just below 0.14%. Also, the minimum of the TPR (75.2%) is the best among all LOGO experiments. The results are in table 6.8.

Gaussian Naive Bayes					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.872639	0.800281	0.199718	0.919544	0.080455
Median	0.912126	0.810405	0.189594	0.919611	0.080388
Min	0.356070	0.100233	0.000000	0.912216	0.073050
Max	0.975628	1.000000	0.899766	0.926949	0.087783
Std	0.103409	0.200017	0.200017	0.002661	0.002661
Gradient Boosting Classifier					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.981020	0.962754	0.037245	0.992857	0.007142
Median	0.992631	0.992752	0.007247	0.992857	0.007142
Min	0.802443	0.687670	0.000000	0.990466	0.005538
Max	0.997902	1.000000	0.312329	0.994461	0.009533
Std	0.035211	0.063939	0.063939	0.000738	0.000738
Logistic Regression					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.967199	0.951367	0.048632	0.974853	0.025146
Median	0.976154	0.994900	0.005099	0.974696	0.025303
Min	0.793908	0.197900	0.000000	0.971376	0.020468
Max	0.992645	1.000000	0.802099	0.979531	0.028623
Std	0.039163	0.110240	0.110240	0.001521	0.001521
Random Forests					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.988930	0.973405	0.026594	0.998550	0.001449
Median	0.997904	0.995666	0.004333	0.998562	0.001437
Min	0.827683	0.746069	0.000000	0.997937	0.000963
Max	0.999680	1.000000	0.253930	0.999036	0.002062
Std	0.027291	0.050254	0.050254	0.000252	0.000252
Support Vector Classifier					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.966942	0.949932	0.050067	0.975880	0.024119
Median	0.976678	0.994433	0.005566	0.975689	0.024310
Min	0.793180	0.168570	0.000000	0.973043	0.019504
Max	0.992782	1.000000	0.831429	0.980495	0.026956
Std	0.040457	0.114270	0.114270	0.001436	0.001436

Table 6.6: Summary of LOGO results - all features

Gaussian Naive Bayes					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.881787	0.873563	0.126436	0.889714	0.110285
Median	0.890347	0.898449	0.101550	0.889520	0.110479
Min	0.466331	0.273399	0.000000	0.880903	0.102188
Max	0.965532	1.000000	0.726600	0.897811	0.119096
Std	0.072943	0.143899	0.143899	0.002910	0.002910
Gradient Boosting Classifier					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.978723	0.963493	0.036506	0.988053	0.011946
Median	0.987993	0.991299	0.008700	0.987911	0.012088
Min	0.810573	0.676561	0.000000	0.985112	0.009452
Max	0.996461	1.000000	0.323438	0.990547	0.014887
Std	0.032339	0.060488	0.060488	0.000927	0.000927
Logistic Regression					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.946724	0.940342	0.059657	0.950902	0.049097
Median	0.952341	0.988133	0.011866	0.950725	0.049274
Min	0.716690	0.430071	0.000000	0.946659	0.042682
Max	0.985386	1.000000	0.569928	0.957317	0.053340
Std	0.045835	0.103551	0.103551	0.002302	0.002302
Random Forests					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.987747	0.974031	0.025968	0.996255	0.003744
Median	0.995894	0.994666	0.005333	0.996302	0.003697
Min	0.824835	0.746833	0.000000	0.994279	0.002363
Max	0.999063	1.000000	0.253166	0.997636	0.005720
Std	0.026702	0.049047	0.049047	0.000591	0.000591
Support Vector Classifier					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.946311	0.939798	0.060201	0.950470	0.049529
Median	0.951904	0.988700	0.011299	0.950325	0.049674
Min	0.720757	0.402284	0.000000	0.945841	0.042930
Max	0.985225	1.000000	0.597715	0.957069	0.054158
Std	0.046145	0.105673	0.105673	0.002392	0.002392

Table 6.7: Summary of LOGO results - best features from statistical tests

Gaussian Naive Bayes					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.896108	0.876900	0.123099	0.909487	0.090512
Median	0.909072	0.907333	0.092666	0.909608	0.090391
Min	0.469339	0.269399	0.000000	0.902830	0.082737
Max	0.971357	1.000000	0.730600	0.917262	0.097169
Std	0.072550	0.141822	0.141822	0.002823	0.002823
Gradient Boosting Classifier					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.980782	0.962449	0.037550	0.992786	0.007213
Median	0.992530	0.992133	0.007866	0.992814	0.007185
Min	0.797827	0.685606	0.000000	0.990833	0.005095
Max	0.997945	1.000000	0.314393	0.994904	0.009166
Std	0.035852	0.064398	0.064398	0.000759	0.000759
Logistic Regression					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.953958	0.938441	0.061558	0.959791	0.040208
Median	0.960835	0.988566	0.011433	0.959410	0.040589
Min	0.770693	0.278172	0.000000	0.955467	0.034822
Max	0.987776	1.000000	0.721827	0.965177	0.044532
Std	0.043306	0.111887	0.111887	0.002035	0.002035
Random Forests					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.989277	0.974356	0.025643	0.998611	0.001388
Median	0.998107	0.996415	0.003584	0.998607	0.001392
Min	0.835308	0.752504	0.000000	0.998103	0.000794
Max	0.999610	1.000000	0.247495	0.999205	0.001896
Std	0.026342	0.048662	0.048662	0.000280	0.000280
Support Vector Classifier					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.953855	0.937527	0.062472	0.960325	0.039674
Median	0.961409	0.984334	0.015665	0.959985	0.040014
Min	0.771222	0.244211	0.000000	0.956051	0.034532
Max	0.988052	1.000000	0.755788	0.965467	0.043948
Std	0.044047	0.114630	0.114630	0.002051	0.002051

Table 6.8: Summary of LOGO results - all features except digits features

All features except n-grams features

Similar to the k-fold experiments, models trained with this set of features perform significantly worse than models trained with other sets of features. For all models except the RF model, there are malware families whose domains are not detected at all. All results are in table 6.9.

Only n-grams features

The best model with n-grams features only is again the RF model followed by the GBC model. The LR and SVM models perform very similarly with their mean and median of accuracy being around 90%. The worst is the GNB model, but the difference is not that significant. The last three models have higher FPR ranging from 10% to 13%. All models have at least 50% minimum of the TPR, which is the highest of all feature sets. The results are in table 6.10.

Summary

As in the k-fold experiments, the best performing model is the Random Forest followed by the Gradient Boosting Classifier. The worst performing model is again the Gaussian Naive Bayes classifier and the Logistic Regression and Support Vector Classifier models are in the middle achieving very similar results.

We have seen very high standard deviation values across all LOGO experiments. This means that there are malware families whose domains are very hard to detect. An analysis of their DGAs and generated domains is provided in the next section.

Gaussian Naive Bayes					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.767595	0.429396	0.570603	0.959407	0.040592
Median	0.901027	0.278866	0.721133	0.959319	0.040680
Min	0.297534	0.000000	0.000000	0.954469	0.036397
Max	0.988023	1.000000	1.000000	0.963602	0.045530
Std	0.233735	0.403123	0.403123	0.001747	0.001747
Gradient Boosting Classifier					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.868441	0.784862	0.215137	0.921271	0.078728
Median	0.915596	0.838633	0.161366	0.921501	0.078498
Min	0.287285	0.000000	0.000000	0.912951	0.068970
Max	0.975805	1.000000	1.000000	0.931029	0.087048
Std	0.129336	0.247310	0.247310	0.003468	0.003468
Logistic Regression					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.833367	0.763932	0.236067	0.872598	0.127401
Median	0.867598	0.810433	0.189566	0.872444	0.127555
Min	0.275664	0.000000	0.000000	0.863472	0.113646
Max	0.959836	1.000000	1.000000	0.886353	0.136527
Std	0.127340	0.263642	0.263642	0.004744	0.004744
Random Forests					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.847900	0.713301	0.286698	0.935598	0.064401
Median	0.915931	0.783919	0.216080	0.935281	0.064718
Min	0.309470	0.016913	0.000000	0.931064	0.056882
Max	0.979580	1.000000	0.983086	0.943117	0.068935
Std	0.147418	0.278718	0.278718	0.002432	0.002432
Support Vector Classifier					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.834719	0.760375	0.239624	0.877527	0.122472
Median	0.872766	0.804966	0.195033	0.877102	0.122897
Min	0.276813	0.000000	0.000000	0.866896	0.109954
Max	0.961306	1.000000	1.000000	0.890045	0.133103
Std	0.128337	0.260723	0.260723	0.004559	0.004559

Table 6.9: Summary of LOGO results - all features except n-grams features

Gaussian Naive Bayes					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.885280	0.917354	0.082645	0.868422	0.131577
Median	0.877250	0.933333	0.066666	0.868280	0.131719
Min	0.618414	0.502800	0.000000	0.861176	0.123554
Max	0.958998	1.000000	0.497199	0.876445	0.138823
Std	0.050998	0.096141	0.096141	0.003233	0.003233
Gradient Boosting Classifier					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.960156	0.960977	0.039022	0.960520	0.039479
Median	0.962827	0.989966	0.010033	0.960662	0.039337
Min	0.800423	0.664541	0.000000	0.956702	0.033930
Max	0.987923	1.000000	0.335458	0.966069	0.043297
Std	0.032113	0.063553	0.063553	0.001873	0.001873
Logistic Regression					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.912904	0.941491	0.058508	0.897867	0.102132
Median	0.904527	0.980444	0.019555	0.898012	0.101987
Min	0.724560	0.560358	0.000000	0.891127	0.091493
Max	0.968774	1.000000	0.439641	0.908506	0.108872
Std	0.046121	0.092077	0.092077	0.003233	0.003233
Random Forests					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.984496	0.971221	0.028778	0.992362	0.007637
Median	0.992360	0.991995	0.008004	0.992420	0.007579
Min	0.854624	0.739088	0.000000	0.990272	0.005893
Max	0.997674	1.000000	0.260911	0.994106	0.009727
Std	0.025368	0.048726	0.048726	0.000762	0.000762
Support Vector Classifier					
	Accuracy	**TPR**	**FNR**	**TNR**	**FPR**
Mean	0.909117	0.947647	0.052352	0.888689	0.111310
Median	0.899562	0.984330	0.015669	0.888699	0.111300
Min	0.737193	0.600493	0.000000	0.880870	0.101535
Max	0.965871	1.000000	0.399506	0.898464	0.119129
Std	0.044687	0.084786	0.084786	0.003372	0.003372

Table 6.10: Summary of LOGO results - only n-grams features

6.2.1 Domain analysis of hard-to-detect malware families

As discussed above in the LOGO results section, there are some malware families whose domains are hard to detect compared to others. This is the case for all models and all subsets of features used. In this section, we try to determine a reason behind this and we analyze the domains of said malware families. We try to identify the differences between these domains and the domains of other families and we try to determine how big these differences are.

First, we analyze the mean and median of features of various sets of domains. We compare three sets of domains: clean domains, that is the TRANCO list from section 4.1.2; easy-to-detect domains, these are the domains taken from the malware families that were detected with very high accuracy in vast majority of LOGO tests, the families are bamital, bedep, blackhole, ccleaner, chinad, chir, corebot, cryptolocker, dircrypt, dyre, ebury, emotet, feodo, fobber, gameover, gspy, madmax, makloader, mirai, monerominer, murofet, murofetweekly, oderoor, omexo, pandabanker, qadars, qakbot, ranbyus, rovnix, sisron, sphinx, sutra, tinba, tinynuke, ud2, ud3, urlzone, vidro, vidrotid, wd, xxhex; and hard-to-detect domains, these are taken from families that were detected poorly in LOGO tests, the families are conficker, ekforward, infy, mydoom, nymaim, padcrypt, proslikefan, pushdo, pushdotid, pykspa, pykspa2, pykspa2s, ramdo, shifu, simda, szribi, tempedrevetdd, tofsee, torpig, vawtrak, virut. We take all domains in every domain set and compute mean and median of all features extracted from them. The means and medians of features of individual hard-to-detect malware families are in appendix C.

As can be seen in tables 6.11 and 6.12, there are large differences in some features between the domain sets. The most significant difference that can be observed is in the domain length and the number of unique characters and the number of digits in domains.

The mean of the domain length for easy-to-detect families is more than or almost two times higher than the mean of hard-to-detect families or the mean of clean domains, respectively. The differences in median are also very significant. Note that the lengths of domains of hard-to-detect families tend to be even shorter than the lengths of clean domains.

Another thing that causes differences in many features is that the domains of hard-to-detect families rarely contain a digit. This affects many features, most noticeably number of digits, is first character digit, digit ratio, digit to letter ratio and longest digit sequence features. But it also affects other features like vowel ratio, consonant ratio, hex character ratio, longest vowel sequence or longest consonant sequence. For the domains of hard-to-detect families, the values of all these features are very close to the values for the clean domains.

features	**easy-to-detect families**	**hard-to-detect families**	**clean domains**
domain length	20.882	9.194	10.552
TLD length	3.030	2.933	2.757
TLD hash	0.597	0.567	0.460
is first character digit	0.217	0.058	0.023
number of digits	6.561	0.479	0.182
number of unique characters	13.272	7.118	8.022
vowel ratio	0.181	0.278	0.363
consonant ratio	0.585	0.662	0.595
hex character ratio	0.455	0.347	0.320
digit ratio	0.234	0.060	0.025
digit to letter ratio	0.564	0.230	0.075
longest consonant sequence	4.155	2.722	2.160
longest vowel sequence	1.427	1.221	1.360
longest digit sequence	1.997	0.247	0.076
is md5 like	0.079	0.000	0.000
shannon entropy	3.546	2.706	2.828
gini coefficient	0.902	0.831	0.840
classification error of characters	0.840	0.765	0.776
2-gram avg	3.814	4.042	4.505
2-gram med	3.216	3.730	4.390
2-gram std	4.027	4.045	4.364
3-gram avg	2.052	2.373	3.320
3-gram med	1.302	1.940	3.073
3-gram std	2.319	2.415	3.269
4-gram avg	0.607	0.846	2.293
4-gram med	0.238	0.593	1.956
4-gram std	0.766	0.918	2.304
5-gram avg	0.077	0.158	1.568
5-gram med	0.003	0.067	1.185
5-gram std	0.137	0.191	1.585

Table 6.11: Comparison of mean of features

features	**easy-to-detect families**	**hard-to-detect families**	**clean domains**
domain length	16	8	10
TLD length	3	3	3
TLD hash	0.545	0.495	0.393
is first character digit	0	0	0
number of digits	3	0	0
number of unique characters	12.500	7	8
vowel ratio	0.160	0.250	0.375
consonant ratio	0.583	0.714	0.600
hex character ratio	0.417	0.250	0.308
digit ratio	0.167	0	0
digit to letter ratio	0.250	0	0
longest consonant sequence	3.500	3	2
longest vowel sequence	1	1	1
longest digit sequence	1	0	0
is md5 like	0	0	0
shannon entropy	3.519	2.750	2.918
gini coefficient	0.903	0.844	0.860
classification error of characters	0.843	0.778	0.800
2-gram avg	3.802	4.124	4.612
2-gram med	3.080	3.766	4.534
2-gram std	4.058	4.132	4.476
3-gram avg	1.983	2.368	3.546
3-gram med	1.264	1.848	3.344
3-gram std	2.302	2.454	3.499
4-gram avg	0.527	0.667	2.624
4-gram med	0	0.500	2.211
4-gram std	0.628	0.728	2.653
5-gram avg	0	0	1.733
5-gram med	0	0	1.190
5-gram std	0	0	1.780

Table 6.12: Comparison of median of features

The domains of hard-to-detect families also contain a small number of unique characters, even smaller than the clean domains. In average, they contain only 7 unique characters compared to the domains of easy-to-detect families which contain 13 unique characters. This affects also Shannon entropy, which is very similar for the clean domains and the domains of hard-to-detect families.

The differences in n-gram features are not very large, they are largest for $n = 2$. For other features, the differences are negligible.

Some malware authors design their DGAs to make generated domains look less random. For example, malware Simda (see section 1.2.1) generates domains where consonants and vowels alternate one by one. In this case, the generated domains might even be meaningful words sometimes.

Other families do not use all letters from the alphabet, examples are Vawtrak family [13] and Mydoom family [4]. Vawtrak uses two strings to choose letters from: consonants - "cdfghlmnrstw" and vowels - "aeiou". First, it generates a pseudorandom number and tests if it is even or odd and based on that chooses one of the strings. Then, a letter is chosen from this string randomly. This repeats until the domain is long enough, the length varies from 7 to 11 characters.

Mydoom uses only one string of letters, but very short - "asnhreqwpm". All generated domains are 10 characters long and the letters are chosen by an index computed by pseudorandom number generator seeded with the current date.

All of these approaches above have effect on many features, primarily on the number of unique characters, entropy, character ratios. Also, these families don't include digits in their domains, which affects digit features.

6.3 Real-world data predictions results

The results of real-world data predictions are as we expected in section 4.2.4. The most DGAs were detected among NXDomains and the least among clean domains in the Authlist. Around a quarter of random set of domains were flagged as DGA domains. In the tables below, we list the counts of domains that were labeled as DGA or clean for different models and datasets.

6.3.1 All features

For all datasets (see section 4.1.4) the performance of the models is analogous to the results of k-fold experiments in table 6.1. The Random Forest model detects the most DGA domains for random domains set and for NXDomains and also has the least number of false positives for the Authlist. All results are in tables 6.13, 6.14 and 6.15.

Random domains

Model	DGA	Clean
Gaussian Naive Bayes	229,974	774,867
Gradient Boosting Classifier	263,811	741,030
Logistic Regression	258,325	746,516
Random Forests	273,642	731,199
Support Vector Classifier	252,896	751,945

Table 6.13: Predictions for random domains, trained with all features.

NXDomains

Model	DGA	Clean
Gaussian Naive Bayes	2,823,471	381,350
Gradient Boosting Classifier	2,955,442	249,379
Logistic Regression	2,939,771	265,050
Random Forests	2,976,512	228,309
Support Vector Classifier	2,931,060	273,761

Table 6.14: Predictions for NXDomains, trained with all features.

Authlist

Model	DGA	Clean
Gaussian Naive Bayes	1,331	73,745
Gradient Boosting Classifier	110	74,966
Logistic Regression	1,135	73,941
Random Forests	70	75,006
Support Vector Classifier	1,091	73,985

Table 6.15: Predictions for Authlist, trained with all features.

6.3.2 All features except digit features

As in previous section, the models trained with this set of features are performance-wise analogous to the models in k-fold experiments, see table 6.3. Also, analogously to the k-fold experiments, models trained with these features have a higher detection rate of DGA domains. All results are in tables 6.16, 6.17 and 6.18.

Random domains

Model	DGA	Clean
Gaussian Naive Bayes	240,089	764,752
Gradient Boosting Classifier	264,086	740,755
Logistic Regression	259,102	745,739
Random Forests	276,705	728,136
Support Vector Classifier	257,741	747,100

Table 6.16: Predictions for random domains, trained with all features except digit features.

NXDomains

Model	DGA	Clean
Gaussian Naive Bayes	2,901,030	303,791
Gradient Boosting Classifier	2,954,562	250,259
Logistic Regression	2,950,439	254,382
Random Forests	2,988,122	216,699
Support Vector Classifier	2,950,373	254,448

Table 6.17: Predictions for NXDomains, trained with all features except digit features.

Authlist

Model	DGA	Clean
Gaussian Naive Bayes	1,667	73,409
Gradient Boosting Classifier	99	74,977
Logistic Regression	1,007	74,069
Random Forests	70	75,006
Support Vector Classifier	923	74,153

Table 6.18: Predictions for Authlist, trained with all features except digit features.

6.4 Speed measurements

We have performed our experiments on desktop PC with Intel Core i7-7700 processor @ 3.6 GHz and 16 GB of RAM running Windows 10. The experiments ran on a single thread.

Feature extraction

The extraction of all features mentioned in chapter 5 for about one million domains took approximately 6.5 minutes.

Training

When training models with our dataset and all features, the training times were following: for the GNB model the training took about 15 seconds, for the GBC model it was 64 minutes, for the LR model 24 minutes. For the RF model the training time was 33 minutes and for the SVM model it was 3 minutes.

Testing

Next, making predictions (without feature extraction) on the trained model for the NXDomains set (see section 4.1.4) which contains just over 3.2 million domains took 20 seconds for the GNB model, 16 seconds for the GBC model and the LR model and for the SVM model it was 10 seconds. The predictions took longest for the RF model, 169 seconds.

The fastest training time is achieved by the GNB model, the GBC model is the slowest. For testing, the SVM model is the fastest and the RF model the slowest. All training and testing times are summarized in table 6.19.

Model	**Training**	**Testing**
Gaussian Naive Bayes	0.25 min.	20 s
Gradient Boosting Classifier	64 min.	16 s
Logistic Regression	24 min.	16 s
Random Forest	33 min.	169 s
Support Vector Machine	3 min.	10 s

Table 6.19: Training and testing times.

Conclusion

In this thesis, we compared and evaluated various supervised machine learning algorithms for classification of domains generated by malware domain generation algorithms. We provided an overview of different DGA types and an overview of research related to DGA detection. Our work focused on detecting arithmetic-based and hash-based DGAs and we used the following five classifiers in our experiments: Gaussian Naive Bayes, Random Forest, Gradient Boosting Classifier, Logistic Regression and Support Vector Machine.

We collected data from 73 malware families and built a dataset that was used in two types of experiments, which simulated two real-world situations. First, a situation where new domains from known malware families appear and second, a situation where an entirely new DGA appears. Our results have shown that the best performing models are decision tree-based classifiers, Gradient Boosting Classifier and Random Forest, the latter having the highest DGA detection rate and the lowest false positive rate across all experiments.

We have trained our models with different sets of features to test if some are more important than the others. The results have shown that all models performed significantly worse if we left out n-gram features from the feature set. The Random Forest model did very well when trained with all 30 features, but even slightly better when trained with feature set where we left out all features involving digits.

The results of our experiments have also revealed a number of malware families, whose DGA domains were hard to detect for all models. We did an analysis of these domains and found that they are rather short, only a small portion of them contains digits and that their character distribution is very similar to clean domains. Some DGAs are even designed to generate domains resembling clean domains.

Continuing this line of work, a comparison of various deep learning methods for detecting DGA domains could be made or due to high variety of domain generation algorithms, an evaluation of combination of different models focused on different DGA types.

Appendix A

Implementation

Our experiments were implemented in Python programming language[1]. We used scikit-learn [27][8] framework for model training and testing, with all data stored in pandas [25][21] dataframes.

The attachment contains all datasets and source code:

- folder **datasets** contains our dataset and scripts for downloading and grouping domains from DGArchive and for generating datasets
- folder **ngrams** contains look-up dictionaries of n-gram frequencies and a script used for generating them
- folder **main** contains all source code for performing experiments, settings for them can be edited in the configuration file

[1] `https://www.python.org/`

Appendix B

LOGO - malware families results

B.1 All features

Gaussian Naive Bayes					
Family	**Accuracy**	**TPR**	**FNR**	**TNR**	**FPR**
bamital	0.973777	1.000000	0.000000	0.917003	0.082996
bedep	0.892828	0.848196	0.151803	0.917129	0.082870
blackhole	0.917335	0.854595	0.145404	0.920670	0.079329
ccleaner	0.917924	1.000000	0.000000	0.917708	0.082291
chinad	0.968743	0.993133	0.006866	0.915595	0.084404
chir	0.917574	1.000000	0.000000	0.916981	0.083018
conficker	0.780368	0.716033	0.283966	0.919852	0.080147
corebot	0.965882	0.987466	0.012533	0.918828	0.081171
cryptolocker	0.818593	0.772766	0.227233	0.918520	0.081479
diamondfox	0.923001	0.976744	0.023255	0.921166	0.078833
dircrypt	0.911100	0.806788	0.193211	0.919885	0.080114
dmsniff	0.918985	0.724637	0.275362	0.919959	0.080040
dyre	0.973435	1.000000	0.000000	0.915608	0.084391
ebury	0.927514	0.990495	0.009504	0.918306	0.081693
ekforward	0.929187	0.996444	0.003555	0.923620	0.076379
emotet	0.837198	0.800066	0.199933	0.918088	0.081911
feodo	0.920479	0.916230	0.083769	0.920539	0.079460
fobber	0.905065	0.810405	0.189594	0.918881	0.081118
gameover	0.975093	0.999333	0.000666	0.921313	0.078686
goznym	0.916167	0.640483	0.359516	0.922835	0.077164
gspy	0.922089	1.000000	0.000000	0.921814	0.078185
hesperbot	0.917854	0.762711	0.237288	0.919846	0.080153
infy	0.950761	0.997411	0.002588	0.918057	0.081942
locky	0.806108	0.753800	0.246199	0.923393	0.076606
madmax	0.920416	0.897959	0.102040	0.920659	0.079340
makloader	0.923673	1.000000	0.000000	0.920873	0.079126
mirai	0.918428	0.878136	0.121863	0.919247	0.080752

modpack	0.921487	0.933333	0.066666	0.921397	0.078602
monerominer	0.975628	1.000000	0.000000	0.922574	0.077425
murofet	0.878273	0.856666	0.143333	0.926029	0.073970
murofetweekly	0.974489	0.999966	0.000033	0.918409	0.081590
mydoom	0.860532	0.477489	0.522510	0.921675	0.078324
necurs	0.829413	0.788900	0.211099	0.919463	0.080536
nymaim	0.777394	0.711999	0.288000	0.919423	0.080576
oderoor	0.838013	0.752763	0.247236	0.922575	0.077424
omexo	0.918349	1.000000	0.000000	0.918236	0.081763
padcrypt	0.697248	0.594300	0.405699	0.922208	0.077791
pandabanker	0.971486	0.999921	0.000078	0.919001	0.080998
proslikefan	0.779535	0.711999	0.288000	0.926949	0.073050
pushdo	0.356070	0.100233	0.899766	0.922832	0.077167
pushdotid	0.837728	0.652442	0.347557	0.918737	0.081262
pykspa	0.766982	0.697999	0.302000	0.919610	0.080389
pykspa2	0.885395	0.551006	0.448993	0.920980	0.079019
pykspa2s	0.758861	0.532784	0.467215	0.923060	0.076939
qadars	0.948273	0.960533	0.039466	0.921485	0.078514
qakbot	0.870215	0.848899	0.151100	0.916775	0.083224
ramdo	0.794059	0.507417	0.492582	0.919611	0.080388
ramnit	0.823940	0.756092	0.243907	0.921731	0.078268
ranbyus	0.853981	0.823666	0.176333	0.920397	0.079602
rovnix	0.927422	0.992627	0.007372	0.918346	0.081653
shifu	0.887135	0.710729	0.289270	0.917141	0.082858
simda	0.524025	0.190918	0.809081	0.922635	0.077364
sisron	0.933040	0.969596	0.030403	0.920925	0.079074
sphinx	0.827156	0.783966	0.216033	0.920716	0.079283
sutra	0.912126	0.862957	0.137042	0.914784	0.085215
szribi	0.812473	0.544836	0.455163	0.916660	0.083339
tempedreve	0.911781	0.694581	0.305418	0.915012	0.084987
tempedrevetdd	0.901891	0.683555	0.316444	0.919641	0.080358
tinba	0.832581	0.793166	0.206833	0.920547	0.079452
tinynuke	0.973380	1.000000	0.000000	0.915525	0.084474
tofsee	0.858163	0.587218	0.412781	0.921619	0.078380
torpig	0.786812	0.661718	0.338281	0.918012	0.081987
ud2	0.923587	1.000000	0.000000	0.921486	0.078513
ud3	0.923115	1.000000	0.000000	0.922786	0.077213
ud4	0.916380	0.724637	0.275362	0.917352	0.082647
urlzone	0.944868	0.957466	0.042533	0.917563	0.082436
vawtrak	0.823738	0.342719	0.657280	0.919235	0.080764
vidro	0.798353	0.744366	0.255633	0.916399	0.083600
vidrotid	0.913850	0.778894	0.221105	0.915797	0.084202
virut	0.582674	0.430966	0.569033	0.920201	0.079798
wd	0.974279	1.000000	0.000000	0.917476	0.082523

xshellghost	0.911348	0.583333	0.416666	0.912216	0.087783
xxhex	0.935517	1.000000	0.000000	0.915133	0.084866
Gradient Boosting Classifier					
Family	**Accuracy**	**TPR**	**FNR**	**TNR**	**FPR**
bamital	0.997902	1.000000	0.000000	0.993360	0.006639
bedep	0.993428	0.995306	0.004693	0.992406	0.007593
blackhole	0.993630	0.995884	0.004115	0.993510	0.006489
ccleaner	0.992783	1.000000	0.000000	0.992764	0.007235
chinad	0.997121	0.999800	0.000199	0.991283	0.008716
chir	0.992854	1.000000	0.000000	0.992803	0.007196
conficker	0.889773	0.841999	0.158000	0.993351	0.006648
corebot	0.996869	0.998733	0.001266	0.992805	0.007194
cryptolocker	0.992344	0.991966	0.008033	0.993167	0.006832
diamondfox	0.992460	0.978858	0.021141	0.992925	0.007074
dircrypt	0.992631	0.986074	0.013925	0.993183	0.006816
dmsniff	0.992556	0.956521	0.043478	0.992736	0.007263
dyre	0.997875	1.000000	0.000000	0.993251	0.006748
ebury	0.993555	0.998999	0.001000	0.992759	0.007240
ekforward	0.993611	0.999111	0.000888	0.993156	0.006843
emotet	0.996618	0.998600	0.001399	0.992302	0.007697
feodo	0.993714	1.000000	0.000000	0.993625	0.006374
fobber	0.991398	0.989994	0.010005	0.991603	0.008396
gameover	0.997840	1.000000	0.000000	0.993048	0.006951
goznym	0.990938	0.864048	0.135951	0.994008	0.005991
gspy	0.992003	1.000000	0.000000	0.991975	0.008024
hesperbot	0.993053	0.977401	0.022598	0.993254	0.006745
infy	0.993642	0.992752	0.007247	0.994266	0.005733
locky	0.977685	0.970333	0.029666	0.994170	0.005829
madmax	0.991845	0.993197	0.006802	0.991830	0.008169
makloader	0.991619	1.000000	0.000000	0.991311	0.008688
mirai	0.993500	1.000000	0.000000	0.993367	0.006632
modpack	0.993063	0.866666	0.133333	0.994020	0.005979
monerominer	0.997738	0.999900	0.000099	0.993033	0.006966
murofet	0.996098	0.997666	0.002333	0.992632	0.007367
murofetweekly	0.997730	1.000000	0.000000	0.992736	0.007263
mydoom	0.983474	0.926784	0.073215	0.992523	0.007476
necurs	0.985217	0.981999	0.018000	0.992368	0.007631
nymaim	0.911738	0.873800	0.126199	0.994135	0.005864
oderoor	0.984340	0.975948	0.024051	0.992665	0.007334
omexo	0.992563	1.000000	0.000000	0.992553	0.007446
padcrypt	0.978755	0.972300	0.027699	0.992861	0.007138
pandabanker	0.997545	0.999763	0.000236	0.993450	0.006549
proslikefan	0.973299	0.964066	0.035933	0.993451	0.006548
pushdo	0.802443	0.716466	0.283533	0.992910	0.007089

pushdotid	0.977738	0.943990	0.056009	0.992493	0.007506
pykspa	0.979751	0.973933	0.026066	0.992624	0.007375
pykspa2	0.984981	0.895905	0.104094	0.994461	0.005538
pykspa2s	0.940222	0.868460	0.131539	0.992342	0.007657
qadars	0.995540	0.996533	0.003466	0.993372	0.006627
qakbot	0.993780	0.994266	0.005733	0.992718	0.007281
ramdo	0.993602	0.995165	0.004834	0.992917	0.007082
ramnit	0.987850	0.984275	0.015724	0.993003	0.006996
ranbyus	0.994209	0.995233	0.004766	0.991966	0.008033
rovnix	0.994595	1.000000	0.000000	0.993842	0.006157
shifu	0.990703	0.983261	0.016738	0.991969	0.008030
simda	0.826912	0.687670	0.312329	0.993534	0.006465
sisron	0.993748	1.000000	0.000000	0.991676	0.008323
sphinx	0.995347	0.996233	0.003766	0.993429	0.006570
sutra	0.990885	0.998643	0.001356	0.990466	0.009533
szribi	0.978205	0.940909	0.059090	0.992724	0.007275
tempedreve	0.992636	0.945812	0.054187	0.993332	0.006667
tempedrevetdd	0.988905	0.944888	0.055111	0.992484	0.007515
tinba	0.992150	0.991999	0.008000	0.992486	0.007513
tinynuke	0.997899	1.000000	0.000000	0.993334	0.006665
tofsee	0.976858	0.907996	0.092003	0.992986	0.007013
torpig	0.970280	0.947839	0.052160	0.993816	0.006183
ud2	0.993644	1.000000	0.000000	0.993469	0.006530
ud3	0.992990	1.000000	0.000000	0.992960	0.007039
ud4	0.991082	0.956521	0.043478	0.991257	0.008742
urlzone	0.996008	0.997500	0.002499	0.992775	0.007224
vawtrak	0.959371	0.791033	0.208966	0.992791	0.007208
vidro	0.981381	0.976133	0.023866	0.992857	0.007142
vidrotid	0.992642	0.989949	0.010050	0.992681	0.007318
virut	0.879334	0.828433	0.171566	0.992583	0.007416
wd	0.997843	1.000000	0.000000	0.993080	0.006919
xshellghost	0.992307	0.944444	0.055555	0.992433	0.007566
xxhex	0.993830	0.999090	0.000909	0.992167	0.007832
Logistic Regression					
Family	**Accuracy**	**TPR**	**FNR**	**TNR**	**FPR**
bamital	0.991563	1.000000	0.000000	0.973296	0.026703
bedep	0.981421	0.996647	0.003352	0.973130	0.026869
blackhole	0.974868	0.995884	0.004115	0.973751	0.026248
ccleaner	0.976820	1.000000	0.000000	0.976759	0.023240
chinad	0.991637	0.999900	0.000099	0.973632	0.026367
chir	0.973222	1.000000	0.000000	0.973029	0.026970
conficker	0.931587	0.911033	0.088966	0.976150	0.023849
corebot	0.991750	0.999333	0.000666	0.975219	0.024780
cryptolocker	0.988459	0.995633	0.004366	0.972815	0.027184

diamondfox	0.972286	0.928118	0.071881	0.973794	0.026205
dircrypt	0.976068	0.989556	0.010443	0.974932	0.025067
dmsniff	0.973187	0.956521	0.043478	0.973271	0.026728
dyre	0.992119	1.000000	0.000000	0.974965	0.025034
ebury	0.977029	0.999499	0.000500	0.973743	0.026256
ekforward	0.972137	0.917333	0.082666	0.976674	0.023325
emotet	0.989559	0.996399	0.003600	0.974656	0.025343
feodo	0.976757	1.000000	0.000000	0.976428	0.023571
fobber	0.979929	0.991995	0.008004	0.978168	0.021831
gameover	0.991866	0.999833	0.000166	0.974190	0.025809
goznym	0.971960	0.924471	0.075528	0.973109	0.026890
gspy	0.975937	1.000000	0.000000	0.975852	0.024147
hesperbot	0.973644	0.977401	0.022598	0.973596	0.026403
infy	0.919656	0.840149	0.159850	0.975395	0.024604
locky	0.973098	0.972566	0.027433	0.974289	0.025710
madmax	0.976263	1.000000	0.000000	0.976006	0.023993
makloader	0.975758	1.000000	0.000000	0.974868	0.025131
mirai	0.974285	1.000000	0.000000	0.973762	0.026237
modpack	0.972756	0.923809	0.076190	0.973126	0.026873
monerominer	0.992645	1.000000	0.000000	0.976634	0.023365
murofet	0.991003	0.998433	0.001566	0.974581	0.025418
murofetweekly	0.989594	0.996500	0.003499	0.974392	0.025607
mydoom	0.967824	0.921782	0.078217	0.975174	0.024825
necurs	0.982435	0.986066	0.013933	0.974364	0.025635
nymaim	0.925615	0.902000	0.097999	0.976905	0.023094
oderoor	0.976306	0.978040	0.021959	0.974586	0.025413
omexo	0.974192	1.000000	0.000000	0.974156	0.025843
padcrypt	0.988908	0.994933	0.005066	0.975744	0.024255
pandabanker	0.991714	1.000000	0.000000	0.976420	0.023579
proslikefan	0.953387	0.942733	0.057266	0.976644	0.023355
pushdo	0.819852	0.749333	0.250666	0.976074	0.023925
pushdotid	0.972718	0.964827	0.035172	0.976167	0.023832
pykspa	0.962556	0.956833	0.043166	0.975219	0.024780
pykspa2	0.972500	0.931991	0.068008	0.976811	0.023188
pykspa2s	0.956909	0.931318	0.068681	0.975495	0.024504
qadars	0.991013	0.997166	0.002833	0.977567	0.022432
qakbot	0.988178	0.994399	0.005600	0.974588	0.025411
ramdo	0.982178	0.998999	0.001000	0.974810	0.025189
ramnit	0.980925	0.986045	0.013954	0.973546	0.026453
ranbyus	0.990250	0.997199	0.002800	0.975023	0.024976
rovnix	0.976708	1.000000	0.000000	0.973466	0.026533
shifu	0.971175	0.936480	0.063519	0.977076	0.022923
simda	0.793908	0.640259	0.359740	0.977771	0.022228
sisron	0.979435	1.000000	0.000000	0.972619	0.027380

sphinx	0.990284	0.997700	0.002299	0.974221	0.025778
sutra	0.976344	1.000000	0.000000	0.975066	0.024933
szribi	0.957680	0.914479	0.085520	0.974498	0.025501
tempedreve	0.974371	0.931034	0.068965	0.975016	0.024983
tempedrevetdd	0.972131	0.934222	0.065777	0.975213	0.024786
tinba	0.988582	0.994900	0.005099	0.974482	0.025517
tinynuke	0.991758	1.000000	0.000000	0.973846	0.026153
tofsee	0.826879	0.197900	0.802099	0.974186	0.025813
torpig	0.962290	0.949365	0.050634	0.975847	0.024152
ud2	0.973022	1.000000	0.000000	0.972280	0.027719
ud3	0.976154	1.000000	0.000000	0.976052	0.023947
ud4	0.976536	0.956521	0.043478	0.976638	0.023361
urlzone	0.990579	0.998600	0.001399	0.973195	0.026804
vawtrak	0.944580	0.792886	0.207113	0.974696	0.025303
vidro	0.976692	0.977600	0.022399	0.974708	0.025291
vidrotid	0.971569	0.984924	0.015075	0.971376	0.028623
virut	0.824924	0.755433	0.244566	0.979531	0.020468
wd	0.991533	1.000000	0.000000	0.972835	0.027164
xshellghost	0.972671	0.944444	0.055555	0.972746	0.027253
xxhex	0.979415	0.998408	0.001591	0.973411	0.026588
Random Forests					
Family	**Accuracy**	**TPR**	**FNR**	**TNR**	**FPR**
bamital	0.999452	1.000000	0.000000	0.998267	0.001732
bedep	0.997825	0.996379	0.003620	0.998612	0.001387
blackhole	0.998753	0.998628	0.001371	0.998760	0.001239
ccleaner	0.998469	1.000000	0.000000	0.998465	0.001534
chinad	0.999474	0.999800	0.000199	0.998765	0.001234
chir	0.998556	1.000000	0.000000	0.998546	0.001453
conficker	0.939389	0.911900	0.088099	0.998988	0.001011
corebot	0.999017	0.999266	0.000733	0.998473	0.001526
cryptolocker	0.995337	0.993866	0.006133	0.998546	0.001453
diamondfox	0.998254	0.983086	0.016913	0.998772	0.001227
dircrypt	0.997904	0.989556	0.010443	0.998607	0.001392
dmsniff	0.998554	1.000000	0.000000	0.998547	0.001452
dyre	0.999520	1.000000	0.000000	0.998476	0.001523
ebury	0.998404	0.998999	0.001000	0.998317	0.001682
ekforward	0.997825	0.985777	0.014222	0.998822	0.001177
emotet	0.997601	0.997366	0.002633	0.998111	0.001888
feodo	0.999049	1.000000	0.000000	0.999036	0.000963
fobber	0.997132	0.989994	0.010005	0.998174	0.001825
gameover	0.999609	1.000000	0.000000	0.998742	0.001257
goznym	0.997360	0.942598	0.057401	0.998684	0.001315
gspy	0.998606	1.000000	0.000000	0.998601	0.001398
hesperbot	0.998352	0.983050	0.016949	0.998549	0.001450

infy	0.995733	0.991614	0.008385	0.998620	0.001379
locky	0.984900	0.978899	0.021100	0.998355	0.001644
madmax	0.997961	0.993197	0.006802	0.998012	0.001987
makloader	0.998199	1.000000	0.000000	0.998133	0.001866
mirai	0.998857	0.996415	0.003584	0.998906	0.001093
modpack	0.997425	0.847619	0.152380	0.998559	0.001440
monerominer	0.999451	0.999900	0.000099	0.998476	0.001523
murofet	0.998141	0.998233	0.001766	0.997937	0.002062
murofetweekly	0.999541	1.000000	0.000000	0.998532	0.001467
mydoom	0.990422	0.939517	0.060482	0.998548	0.001451
necurs	0.987585	0.982633	0.017366	0.998592	0.001407
nymaim	0.951795	0.930499	0.069500	0.998045	0.001954
oderoor	0.997173	0.995966	0.004033	0.998370	0.001629
omexo	0.998614	1.000000	0.000000	0.998612	0.001387
padcrypt	0.984792	0.978366	0.021633	0.998834	0.001165
pandabanker	0.999207	0.999526	0.000473	0.998617	0.001382
proslikefan	0.978831	0.969866	0.030133	0.998399	0.001600
pushdo	0.827683	0.750533	0.249466	0.998596	0.001403
pushdotid	0.985192	0.955825	0.044174	0.998032	0.001967
pykspa	0.984182	0.977600	0.022399	0.998746	0.001253
pykspa2	0.996796	0.977793	0.022206	0.998818	0.001181
pykspa2s	0.978792	0.951802	0.048197	0.998395	0.001604
qadars	0.996615	0.995666	0.004333	0.998689	0.001310
qakbot	0.996730	0.995733	0.004266	0.998907	0.001092
ramdo	0.997918	0.996666	0.003333	0.998466	0.001533
ramnit	0.992328	0.988219	0.011780	0.998250	0.001749
ranbyus	0.997688	0.997166	0.002833	0.998831	0.001168
rovnix	0.998777	1.000000	0.000000	0.998607	0.001392
shifu	0.995882	0.980686	0.019313	0.998466	0.001533
simda	0.860841	0.746069	0.253930	0.998183	0.001816
sisron	0.999177	1.000000	0.000000	0.998904	0.001095
sphinx	0.997651	0.997233	0.002766	0.998555	0.001444
sutra	0.998678	0.998643	0.001356	0.998679	0.001320
szribi	0.985399	0.951481	0.048518	0.998603	0.001396
tempedreve	0.997906	0.940886	0.059113	0.998754	0.001245
tempedrevetdd	0.995455	0.955555	0.044444	0.998699	0.001300
tinba	0.994682	0.992999	0.007000	0.998437	0.001562
tinynuke	0.999680	1.000000	0.000000	0.998985	0.001014
tofsee	0.983010	0.918184	0.081815	0.998192	0.001807
torpig	0.979192	0.960671	0.039328	0.998617	0.001382
ud2	0.998940	1.000000	0.000000	0.998911	0.001088
ud3	0.998554	1.000000	0.000000	0.998548	0.001451
ud4	0.998538	1.000000	0.000000	0.998530	0.001469
urlzone	0.998220	0.997900	0.002099	0.998916	0.001083

vawtrak	0.970295	0.827713	0.172286	0.998602	0.001397
vidro	0.991925	0.989033	0.010966	0.998250	0.001749
vidrotid	0.998642	0.989949	0.010050	0.998768	0.001231
virut	0.936850	0.909233	0.090766	0.998294	0.001705
wd	0.999449	1.000000	0.000000	0.998233	0.001766
xshellghost	0.998607	0.944444	0.055555	0.998751	0.001248
xxhex	0.998525	0.998408	0.001591	0.998562	0.001437
Support Vector Classifier					
Family	**Accuracy**	**TPR**	**FNR**	**TNR**	**FPR**
bamital	0.991791	1.000000	0.000000	0.974018	0.025981
bedep	0.982130	0.997049	0.002950	0.974007	0.025992
blackhole	0.975491	0.997256	0.002743	0.974334	0.025665
ccleaner	0.978278	1.000000	0.000000	0.978221	0.021778
chinad	0.992185	0.999900	0.000099	0.975375	0.024624
chir	0.973944	1.000000	0.000000	0.973756	0.026243
conficker	0.929192	0.906766	0.093233	0.977813	0.022186
corebot	0.992161	0.999466	0.000533	0.976237	0.023762
cryptolocker	0.989076	0.995800	0.004199	0.974414	0.025585
diamondfox	0.973752	0.947145	0.052854	0.974660	0.025339
dircrypt	0.976338	0.987815	0.012184	0.975371	0.024628
dmsniff	0.974344	0.971014	0.028985	0.974360	0.025639
dyre	0.992599	1.000000	0.000000	0.976489	0.023510
ebury	0.977794	0.999499	0.000500	0.974621	0.025378
ekforward	0.974040	0.928888	0.071111	0.977777	0.022222
emotet	0.990061	0.996800	0.003199	0.975383	0.024616
feodo	0.977488	1.000000	0.000000	0.977169	0.022830
fobber	0.980057	0.991495	0.008504	0.978387	0.021612
gameover	0.992141	0.999800	0.000199	0.975151	0.024848
goznym	0.972888	0.924471	0.075528	0.974059	0.025940
gspy	0.976377	1.000000	0.000000	0.976293	0.023706
hesperbot	0.974432	0.977401	0.022598	0.974394	0.025605
infy	0.897853	0.785485	0.214514	0.976629	0.023370
locky	0.973536	0.972633	0.027366	0.975560	0.024439
madmax	0.977137	1.000000	0.000000	0.976889	0.023110
makloader	0.976728	1.000000	0.000000	0.975874	0.024125
mirai	0.975500	1.000000	0.000000	0.975001	0.024998
modpack	0.974472	0.923809	0.076190	0.974855	0.025144
murofet	0.992782	1.000000	0.000000	0.977069	0.022930
monerominer	0.991623	0.998500	0.001499	0.976423	0.023576
murofetweekly	0.971303	0.969333	0.030666	0.975640	0.024359
mydoom	0.968388	0.920873	0.079126	0.975972	0.024027
necurs	0.983102	0.986066	0.013933	0.976513	0.023486
nymaim	0.924930	0.900566	0.099433	0.977846	0.022153
oderoor	0.976678	0.977741	0.022258	0.975624	0.024375

omexo	0.975723	1.000000	0.000000	0.975689	0.024310
padcrypt	0.988840	0.994433	0.005566	0.976618	0.023381
pandabanker	0.992149	1.000000	0.000000	0.977658	0.022341
proslikefan	0.952747	0.941400	0.058599	0.977517	0.022482
pushdo	0.821712	0.752033	0.247966	0.976074	0.023925
pushdotid	0.973377	0.964827	0.035172	0.977115	0.022884
pykspa	0.961592	0.954999	0.045000	0.976178	0.023821
pykspa2	0.973101	0.929909	0.070090	0.977697	0.022302
pykspa2s	0.955937	0.927703	0.072296	0.976443	0.023556
qadars	0.991493	0.997533	0.002466	0.978295	0.021704
qakbot	0.988521	0.994500	0.005499	0.975462	0.024537
ramdo	0.982584	0.999166	0.000833	0.975321	0.024678
ramnit	0.981611	0.986095	0.013904	0.975149	0.024850
ranbyus	0.991096	0.997366	0.002633	0.977360	0.022639
rovnix	0.977158	1.000000	0.000000	0.973979	0.026020
shifu	0.971237	0.935622	0.064377	0.977295	0.022704
simda	0.793180	0.638256	0.361743	0.978570	0.021429
sisron	0.980531	1.000000	0.000000	0.974080	0.025919
sphinx	0.990558	0.997900	0.002099	0.974655	0.025344
sutra	0.977109	1.000000	0.000000	0.975872	0.024127
szribi	0.957099	0.909760	0.090239	0.975527	0.024472
tempedreve	0.974588	0.931034	0.068965	0.975236	0.024763
tempedrevetdd	0.972599	0.933333	0.066666	0.975791	0.024208
tinba	0.989134	0.994966	0.005033	0.976119	0.023880
tinynuke	0.991964	0.999933	0.000066	0.974643	0.025356
tofsee	0.821957	0.168570	0.831429	0.974981	0.025018
torpig	0.961971	0.948047	0.051952	0.976575	0.023424
ud2	0.974435	1.000000	0.000000	0.973731	0.026268
ud3	0.977455	1.000000	0.000000	0.977358	0.022641
ud4	0.977633	0.956521	0.043478	0.977740	0.022259
urlzone	0.991058	0.998766	0.001233	0.974351	0.025648
vawtrak	0.945010	0.791404	0.208595	0.975505	0.024494
vidro	0.976944	0.977366	0.022633	0.976020	0.023979
vidrotid	0.973212	0.984924	0.015075	0.973043	0.026956
virut	0.814276	0.739566	0.260433	0.980495	0.019504
wd	0.991923	1.000000	0.000000	0.974087	0.025912
xshellghost	0.974210	0.944444	0.055555	0.974289	0.025710
xxhex	0.980453	0.999090	0.000909	0.974561	0.025438

Table B.1: LOGO results for individual malware families - all features

B.2 Best features from statistical tests

Gaussian Naive Bayes					
Family	**Accuracy**	**TPR**	**FNR**	**TNR**	**FPR**
bamital	0.964999	1.000000	0.000000	0.889217	0.110782
bedep	0.901290	0.924902	0.075097	0.888434	0.111565
blackhole	0.893381	0.932784	0.067215	0.891286	0.108713
ccleaner	0.888913	1.000000	0.000000	0.888620	0.111379
chinad	0.961592	0.996366	0.003633	0.885813	0.114186
chir	0.886972	1.000000	0.000000	0.886158	0.113841
conficker	0.856422	0.842233	0.157766	0.887186	0.112813
corebot	0.960329	0.992133	0.007866	0.890996	0.109003
cryptolocker	0.874354	0.867366	0.132633	0.889591	0.110408
diamondfox	0.893752	0.970401	0.029598	0.891134	0.108865
dircrypt	0.890481	0.903394	0.096605	0.889393	0.110606
dmsniff	0.891956	0.768115	0.231884	0.892576	0.107423
dyre	0.964185	1.000000	0.000000	0.886220	0.113779
ebury	0.903330	0.992496	0.007503	0.890294	0.109705
ekforward	0.900033	0.997333	0.002666	0.891979	0.108020
emotet	0.893376	0.895366	0.104633	0.889042	0.110957
feodo	0.891974	0.958115	0.041884	0.891038	0.108961
fobber	0.890347	0.898449	0.101550	0.889164	0.110835
gameover	0.965327	0.999566	0.000433	0.889365	0.110634
goznym	0.893478	0.833836	0.166163	0.894921	0.105078
gspy	0.891644	1.000000	0.000000	0.891261	0.108738
hesperbot	0.887703	0.841807	0.158192	0.888292	0.111707
infy	0.932670	0.997204	0.002795	0.887429	0.112570
locky	0.864130	0.849733	0.150266	0.896412	0.103587
madmax	0.890927	0.959183	0.040816	0.890189	0.109810
makloader	0.895414	1.000000	0.000000	0.891577	0.108422
mirai	0.891857	0.870967	0.129032	0.892281	0.107718
modpack	0.889953	0.942857	0.057142	0.889553	0.110446
monerominer	0.965532	0.999966	0.000033	0.890573	0.109426
murofet	0.912583	0.919266	0.080733	0.897811	0.102188
murofetweekly	0.965206	0.999966	0.000033	0.888693	0.111306
mydoom	0.857527	0.653024	0.346975	0.890171	0.109828
necurs	0.883693	0.881866	0.118133	0.887752	0.112247
nymaim	0.856640	0.842133	0.157866	0.888148	0.111851
oderoor	0.874837	0.855841	0.144158	0.893680	0.106319
omexo	0.889698	1.000000	0.000000	0.889545	0.110454
padcrypt	0.739440	0.669266	0.330733	0.892781	0.107218
pandabanker	0.961103	0.999921	0.000078	0.889454	0.110545
proslikefan	0.857603	0.841533	0.158466	0.892680	0.107319
pushdo	0.466331	0.273399	0.726600	0.893738	0.106261
pushdotid	0.855425	0.776962	0.223037	0.889731	0.110268

pykspa	0.840078	0.818233	0.181766	0.888413	0.111586
pykspa2	0.869443	0.680777	0.319222	0.889520	0.110479
pykspa2s	0.797051	0.664223	0.335776	0.893523	0.106476
qadars	0.947724	0.973799	0.026200	0.890750	0.109249
qakbot	0.908766	0.918399	0.081600	0.887723	0.112276
ramdo	0.846864	0.746791	0.253208	0.890698	0.109301
ramnit	0.880447	0.871523	0.128476	0.893310	0.106689
ranbyus	0.899274	0.902299	0.097700	0.892645	0.107354
rovnix	0.902007	0.997367	0.002632	0.888734	0.111265
shifu	0.876216	0.824892	0.175107	0.884946	0.115053
simda	0.611825	0.374673	0.625326	0.895612	0.104387
sisron	0.918782	1.000000	0.000000	0.891866	0.108133
sphinx	0.891536	0.891766	0.108233	0.891039	0.108960
sutra	0.883462	0.930800	0.069199	0.880903	0.119096
szribi	0.856908	0.778742	0.221257	0.887337	0.112662
tempedreve	0.884421	0.778325	0.221674	0.885998	0.114001
tempedrevetdd	0.882911	0.788444	0.211555	0.890591	0.109408
tinba	0.886054	0.884900	0.115099	0.888632	0.111367
tinynuke	0.964500	1.000000	0.000000	0.887343	0.112656
tofsee	0.874860	0.815375	0.184624	0.888792	0.111207
torpig	0.836487	0.789068	0.210931	0.886221	0.113778
ud2	0.895692	1.000000	0.000000	0.892823	0.107176
ud3	0.893344	1.000000	0.000000	0.892888	0.107111
ud4	0.886558	0.768115	0.231884	0.887158	0.112841
urlzone	0.947583	0.975266	0.024733	0.887580	0.112419
vawtrak	0.819933	0.462763	0.537236	0.890842	0.109157
vidro	0.864089	0.854400	0.145599	0.885276	0.114723
vidrotid	0.885348	0.854271	0.145728	0.885797	0.114202
virut	0.758048	0.699400	0.300599	0.888534	0.111465
wd	0.964849	1.000000	0.000000	0.887220	0.112779
xshellghost	0.884533	0.777777	0.222222	0.884815	0.115184
xxhex	0.914441	1.000000	0.000000	0.887395	0.112604
Gradient Boosting Classifier					
Family	**Accuracy**	**TPR**	**FNR**	**TNR**	**FPR**
bamital	0.996032	1.000000	0.000000	0.987442	0.012557
bedep	0.990686	0.995976	0.004023	0.987806	0.012193
blackhole	0.988507	0.997256	0.002743	0.988042	0.011957
ccleaner	0.987025	1.000000	0.000000	0.986991	0.013008
chinad	0.995864	0.999833	0.000166	0.987215	0.012784
chir	0.988451	1.000000	0.000000	0.988368	0.011631
conficker	0.907931	0.870666	0.129333	0.988725	0.011274
corebot	0.996183	0.999600	0.000399	0.988736	0.011263
cryptolocker	0.990516	0.991133	0.008866	0.989169	0.010830
diamondfox	0.987993	0.974630	0.025369	0.988449	0.011550

dircrypt	0.987493	0.985204	0.014795	0.987685	0.012314
dmsniff	0.987714	0.956521	0.043478	0.987870	0.012129
dyre	0.995797	1.000000	0.000000	0.986648	0.013351
ebury	0.988259	0.996998	0.003001	0.986981	0.013018
ekforward	0.989534	0.995555	0.004444	0.989036	0.010963
emotet	0.994448	0.996633	0.003366	0.989688	0.010311
feodo	0.988086	1.000000	0.000000	0.987917	0.012082
fobber	0.987766	0.990995	0.009004	0.987295	0.012704
gameover	0.996461	1.000000	0.000000	0.988611	0.011388
goznym	0.985801	0.903323	0.096676	0.987796	0.012203
gspy	0.988262	1.000000	0.000000	0.988220	0.011779
hesperbot	0.987538	0.977401	0.022598	0.987668	0.012331
infy	0.989845	0.991303	0.008696	0.988822	0.011177
locky	0.978769	0.973533	0.026466	0.990508	0.009491
madmax	0.988131	0.965986	0.034013	0.988371	0.011628
makloader	0.986771	1.000000	0.000000	0.986285	0.013714
mirai	0.988857	0.996415	0.003584	0.988703	0.011296
modpack	0.988559	0.876190	0.123809	0.989409	0.010590
monerominer	0.996185	0.999900	0.000099	0.988099	0.011900
murofet	0.994492	0.997033	0.002966	0.988874	0.011125
murofetweekly	0.995943	1.000000	0.000000	0.987012	0.012987
mydoom	0.978841	0.919054	0.080945	0.988385	0.011614
necurs	0.986872	0.986766	0.013233	0.987108	0.012891
nymaim	0.931002	0.904166	0.095833	0.989285	0.010714
oderoor	0.980955	0.974902	0.025097	0.986960	0.013039
omexo	0.987096	1.000000	0.000000	0.987078	0.012921
padcrypt	0.977795	0.972400	0.027599	0.989584	0.010415
pandabanker	0.995575	0.999802	0.000197	0.987773	0.012226
proslikefan	0.969161	0.959866	0.040133	0.989449	0.010550
pushdo	0.810573	0.729333	0.270666	0.990547	0.009452
pushdotid	0.974543	0.943990	0.056009	0.987901	0.012098
pykspa	0.972244	0.964500	0.035499	0.989379	0.010620
pykspa2	0.981110	0.908396	0.091603	0.988848	0.011151
pykspa2s	0.943982	0.883723	0.116276	0.987747	0.012252
qadars	0.992476	0.994366	0.005633	0.988346	0.011653
qakbot	0.990670	0.992033	0.007966	0.987694	0.012305
ramdo	0.990759	0.996166	0.003833	0.988390	0.011609
ramnit	0.985164	0.983517	0.016482	0.987538	0.012461
ranbyus	0.994095	0.996900	0.003099	0.987950	0.012049
rovnix	0.990027	1.000000	0.000000	0.988638	0.011361
shifu	0.982218	0.955364	0.044635	0.986786	0.013213
simda	0.818281	0.676561	0.323438	0.987868	0.012131
sisron	0.990841	1.000000	0.000000	0.987806	0.012193
sphinx	0.993660	0.996833	0.003166	0.986786	0.013213

sutra	0.985737	0.997286	0.002713	0.985112	0.014887
szribi	0.973180	0.934868	0.065131	0.988094	0.011905
tempedreve	0.986933	0.921182	0.078817	0.987911	0.012088
tempedrevetdd	0.983759	0.933333	0.066666	0.987859	0.012140
tinba	0.990009	0.991299	0.008700	0.987129	0.012870
tinynuke	0.996096	1.000000	0.000000	0.987611	0.012388
tofsee	0.972699	0.907996	0.092003	0.987852	0.012147
torpig	0.964101	0.940903	0.059096	0.988432	0.011567
ud2	0.989618	1.000000	0.000000	0.989333	0.010666
ud3	0.988510	1.000000	0.000000	0.988461	0.011538
ud4	0.987939	0.971014	0.028985	0.988025	0.011974
urlzone	0.993248	0.995766	0.004233	0.987789	0.012210
vawtrak	0.955505	0.790292	0.209707	0.988304	0.011695
vidro	0.978408	0.974066	0.025933	0.987900	0.012099
vidrotid	0.987141	0.989949	0.010050	0.987101	0.012898
virut	0.907805	0.871866	0.128133	0.987763	0.012236
wd	0.996099	1.000000	0.000000	0.987485	0.012514
xshellghost	0.987032	0.944444	0.055555	0.987144	0.012855
xxhex	0.991100	1.000000	0.000000	0.988286	0.011713
Logistic Regression					
Family	**Accuracy**	**TPR**	**FNR**	**TNR**	**FPR**
bamital	0.984198	1.000000	0.000000	0.949985	0.050014
bedep	0.963740	0.993026	0.006973	0.947794	0.052205
blackhole	0.952644	0.991769	0.008230	0.950565	0.049434
ccleaner	0.951745	1.000000	0.000000	0.951618	0.048381
chinad	0.983572	0.999166	0.000833	0.949589	0.050410
chir	0.951497	1.000000	0.000000	0.951148	0.048851
conficker	0.890229	0.861466	0.138533	0.952590	0.047409
corebot	0.982541	0.997533	0.002466	0.949858	0.050141
cryptolocker	0.975821	0.988033	0.011966	0.949193	0.050806
diamondfox	0.950017	0.875264	0.124735	0.952570	0.047429
dircrypt	0.951798	0.983463	0.016536	0.949131	0.050868
dmsniff	0.949266	0.956521	0.043478	0.949230	0.050769
dyre	0.983668	1.000000	0.000000	0.948116	0.051883
ebury	0.955079	0.996998	0.003001	0.948950	0.051049
ekforward	0.947400	0.927111	0.072888	0.949080	0.050919
emotet	0.980146	0.994466	0.005533	0.948950	0.051049
feodo	0.953442	1.000000	0.000000	0.952783	0.047216
fobber	0.955845	0.986493	0.013506	0.951372	0.048627
gameover	0.985386	0.999966	0.000033	0.953039	0.046960
goznym	0.946204	0.891238	0.108761	0.947533	0.052466
gspy	0.954075	1.000000	0.000000	0.953912	0.046087
hesperbot	0.951801	0.966101	0.033898	0.951617	0.048382
infy	0.875709	0.764882	0.235117	0.953403	0.046596

locky	0.957561	0.959666	0.040333	0.952840	0.047159
madmax	0.951652	0.993197	0.006802	0.951203	0.048796
makloader	0.949577	1.000000	0.000000	0.947727	0.052272
mirai	0.952357	0.996415	0.003584	0.951461	0.048538
modpack	0.950804	0.923809	0.076190	0.951008	0.048991
monerominer	0.984993	1.000000	0.000000	0.952325	0.047674
murofet	0.981869	0.994833	0.005166	0.953215	0.046784
murofetweekly	0.984161	1.000000	0.000000	0.949299	0.050700
mydoom	0.938716	0.874033	0.125966	0.949041	0.050958
necurs	0.969905	0.979566	0.020433	0.948432	0.051567
nymaim	0.905712	0.882033	0.117966	0.957141	0.042858
oderoor	0.952017	0.956080	0.043919	0.947988	0.052011
omexo	0.948312	1.000000	0.000000	0.948240	0.051759
padcrypt	0.978572	0.990933	0.009066	0.951562	0.048437
pandabanker	0.983812	0.999921	0.000078	0.954079	0.045920
proslikefan	0.926458	0.914300	0.085699	0.952997	0.047002
pushdo	0.789398	0.713600	0.286399	0.957317	0.042682
pushdotid	0.947870	0.939156	0.060843	0.951679	0.048320
pykspa	0.931036	0.919566	0.080433	0.956412	0.043587
pykspa2	0.947670	0.893823	0.106176	0.953400	0.046599
pykspa2s	0.926619	0.894467	0.105532	0.949970	0.050029
qadars	0.977086	0.988133	0.011866	0.952949	0.047050
qakbot	0.976951	0.988766	0.011233	0.951143	0.048856
ramdo	0.966895	0.998499	0.001500	0.953051	0.046948
ramnit	0.967343	0.978208	0.021791	0.951683	0.048316
ranbyus	0.981003	0.994866	0.005133	0.950631	0.049368
rovnix	0.955797	0.999473	0.000526	0.949717	0.050282
shifu	0.939730	0.882403	0.117596	0.949481	0.050518
simda	0.716690	0.517817	0.482182	0.954670	0.045329
sisron	0.960625	1.000000	0.000000	0.947575	0.052424
sphinx	0.981002	0.995166	0.004833	0.950321	0.049678
sutra	0.952341	1.000000	0.000000	0.949765	0.050234
szribi	0.928745	0.874834	0.125165	0.949731	0.050268
tempedreve	0.947877	0.921182	0.078817	0.948274	0.051725
tempedrevetdd	0.945532	0.903111	0.096888	0.948981	0.051018
tinba	0.975875	0.988966	0.011033	0.946659	0.053340
tinynuke	0.984201	1.000000	0.000000	0.949865	0.050134
tofsee	0.852188	0.430071	0.569928	0.951048	0.048951
torpig	0.930760	0.911840	0.088159	0.950603	0.049396
ud2	0.955225	1.000000	0.000000	0.953994	0.046005
ud3	0.950935	1.000000	0.000000	0.950725	0.049274
ud4	0.951100	0.956521	0.043478	0.951072	0.048927
urlzone	0.980155	0.994166	0.005833	0.949786	0.050213
vawtrak	0.917945	0.746943	0.253056	0.951894	0.048105

vidro	0.955123	0.956666	0.043333	0.951749	0.048250
vidrotid	0.948710	0.969849	0.030150	0.948405	0.051594
virut	0.782632	0.705066	0.294933	0.955206	0.044793
wd	0.984696	1.000000	0.000000	0.950898	0.049101
xshellghost	0.948347	0.944444	0.055555	0.948358	0.051641
xxhex	0.960469	0.999090	0.000909	0.948260	0.051739
Random Forests					
Family	**Accuracy**	**TPR**	**FNR**	**TNR**	**FPR**
bamital	0.998882	1.000000	0.000000	0.996463	0.003536
bedep	0.996076	0.995842	0.004157	0.996203	0.003796
blackhole	0.996122	0.998628	0.001371	0.995989	0.004010
ccleaner	0.995772	1.000000	0.000000	0.995761	0.004238
chinad	0.998811	0.999766	0.000233	0.996731	0.003268
chir	0.995885	1.000000	0.000000	0.995856	0.004143
conficker	0.943791	0.919633	0.080366	0.996169	0.003830
corebot	0.998400	0.999566	0.000433	0.995857	0.004142
cryptolocker	0.993258	0.991900	0.008099	0.996220	0.003779
diamondfox	0.996020	0.976744	0.023255	0.996679	0.003320
dircrypt	0.996214	0.988685	0.011314	0.996848	0.003151
dmsniff	0.996964	1.000000	0.000000	0.996949	0.003050
dyre	0.999063	1.000000	0.000000	0.997024	0.002975
ebury	0.996618	0.998999	0.001000	0.996270	0.003729
ekforward	0.996126	0.994666	0.005333	0.996247	0.003752
emotet	0.997395	0.997766	0.002233	0.996587	0.003412
feodo	0.997149	1.000000	0.000000	0.997109	0.002890
fobber	0.996049	0.991495	0.008504	0.996714	0.003285
gameover	0.998851	1.000000	0.000000	0.996302	0.003697
goznym	0.995077	0.939577	0.060422	0.996419	0.003580
gspy	0.996625	1.000000	0.000000	0.996613	0.003386
hesperbot	0.995774	0.988700	0.011299	0.995865	0.004134
infy	0.994581	0.991303	0.008696	0.996879	0.003120
locky	0.984946	0.979899	0.020100	0.996263	0.003736
madmax	0.995922	0.979591	0.020408	0.996099	0.003900
makloader	0.995359	1.000000	0.000000	0.995189	0.004810
mirai	0.996714	0.996415	0.003584	0.996720	0.003279
modpack	0.995495	0.866666	0.133333	0.996469	0.003530
monerominer	0.998401	0.999900	0.000099	0.995138	0.004861
murofet	0.997016	0.997666	0.002333	0.995579	0.004420
murofetweekly	0.998533	1.000000	0.000000	0.995304	0.004695
mydoom	0.987543	0.936334	0.063665	0.995717	0.004282
necurs	0.991677	0.989900	0.010099	0.995628	0.004371
nymaim	0.955310	0.936699	0.063300	0.995728	0.004271
oderoor	0.996503	0.997236	0.002763	0.995776	0.004223
omexo	0.996792	1.000000	0.000000	0.996787	0.003212

padcrypt	0.984815	0.979333	0.020666	0.996795	0.003204
pandabanker	0.998516	0.999448	0.000551	0.996797	0.003202
proslikefan	0.978945	0.970766	0.029233	0.996798	0.003201
pushdo	0.824835	0.746833	0.253166	0.997636	0.002363
pushdotid	0.983367	0.954492	0.045507	0.995991	0.004008
pykspa	0.981473	0.974733	0.025266	0.996386	0.003613
pykspa2	0.995260	0.976405	0.023594	0.997267	0.002732
pykspa2s	0.977947	0.953408	0.046591	0.995770	0.004229
qadars	0.995403	0.994966	0.005033	0.996358	0.003641
qakbot	0.994695	0.993700	0.006299	0.996869	0.003130
ramdo	0.996953	0.997999	0.002000	0.996495	0.003504
ramnit	0.991074	0.987561	0.012438	0.996137	0.003862
ranbyus	0.997001	0.997099	0.002900	0.996786	0.003213
rovnix	0.996911	1.000000	0.000000	0.996481	0.003518
shifu	0.992700	0.971244	0.028755	0.996349	0.003650
simda	0.861635	0.748861	0.251138	0.996585	0.003414
sisron	0.995941	1.000000	0.000000	0.994596	0.005403
sphinx	0.996830	0.997233	0.002766	0.995956	0.004043
sutra	0.994434	0.997286	0.002713	0.994279	0.005720
szribi	0.981115	0.942986	0.057013	0.995957	0.004042
tempedreve	0.995163	0.916256	0.083743	0.996336	0.003663
tempedrevetdd	0.993116	0.952888	0.047111	0.996386	0.003613
tinba	0.993439	0.992233	0.007766	0.996131	0.003868
tinynuke	0.998653	1.000000	0.000000	0.995725	0.004274
tofsee	0.985412	0.937635	0.062364	0.996601	0.003398
torpig	0.974150	0.953180	0.046819	0.996144	0.003855
ud2	0.996892	1.000000	0.000000	0.996807	0.003192
ud3	0.996025	1.000000	0.000000	0.996008	0.003991
ud4	0.996418	1.000000	0.000000	0.996400	0.003599
urlzone	0.995894	0.995566	0.004433	0.996604	0.003395
vawtrak	0.969682	0.835865	0.164134	0.996248	0.003751
vidro	0.994922	0.994233	0.005766	0.996428	0.003571
vidrotid	0.995713	0.989949	0.010050	0.995797	0.004202
virut	0.947980	0.926300	0.073699	0.996217	0.003782
wd	0.998416	1.000000	0.000000	0.994920	0.005079
xshellghost	0.996190	0.972222	0.027777	0.996253	0.003746
xxhex	0.997870	1.000000	0.000000	0.997197	0.002802
Support Vector Classifier					
Family	**Accuracy**	**TPR**	**FNR**	**TNR**	**FPR**
bamital	0.984015	1.000000	0.000000	0.949408	0.050591
bedep	0.963598	0.993160	0.006839	0.947502	0.052497
blackhole	0.952298	0.991769	0.008230	0.950200	0.049799
ccleaner	0.951016	1.000000	0.000000	0.950887	0.049112
chinad	0.983389	0.999033	0.000966	0.949299	0.050700

chir	0.951281	1.000000	0.000000	0.950930	0.049069
conficker	0.890777	0.862199	0.137800	0.952735	0.047264
corebot	0.982061	0.996600	0.003399	0.950366	0.049633
cryptolocker	0.976004	0.988633	0.011366	0.948466	0.051533
diamondfox	0.950157	0.892177	0.107822	0.952136	0.047863
dircrypt	0.951798	0.986074	0.013925	0.948911	0.051088
dmsniff	0.949338	0.956521	0.043478	0.949302	0.050697
dyre	0.983805	1.000000	0.000000	0.948552	0.051447
ebury	0.954313	0.996998	0.003001	0.948072	0.051927
ekforward	0.947060	0.917333	0.082666	0.949521	0.050478
emotet	0.980420	0.994933	0.005066	0.948805	0.051194
feodo	0.952565	1.000000	0.000000	0.951893	0.048106
fobber	0.955272	0.987493	0.012506	0.950569	0.049430
gameover	0.985225	0.999866	0.000133	0.952743	0.047256
goznym	0.946347	0.888217	0.111782	0.947753	0.052246
gspy	0.954075	1.000000	0.000000	0.953912	0.046087
hesperbot	0.951156	0.960451	0.039548	0.951037	0.048962
infy	0.865085	0.740345	0.259654	0.952533	0.047466
locky	0.957630	0.960366	0.039633	0.951494	0.048505
madmax	0.950560	0.993197	0.006802	0.950099	0.049900
makloader	0.948053	1.000000	0.000000	0.946147	0.053852
mirai	0.951214	0.992831	0.007168	0.950368	0.049631
modpack	0.950017	0.923809	0.076190	0.950216	0.049783
murofet	0.984765	1.000000	0.000000	0.951600	0.048399
monerominer	0.981823	0.995266	0.004733	0.952110	0.047889
murofetweekly	0.983818	0.999966	0.000033	0.948272	0.051727
mydoom	0.938967	0.877216	0.122783	0.948824	0.051175
necurs	0.970549	0.980533	0.019466	0.948358	0.051641
nymaim	0.905895	0.882333	0.117666	0.957069	0.042930
oderoor	0.952798	0.957947	0.042052	0.947692	0.052307
omexo	0.947510	1.000000	0.000000	0.947437	0.052562
padcrypt	0.978160	0.990366	0.009633	0.951489	0.048510
pandabanker	0.983505	0.999724	0.000275	0.953569	0.046430
proslikefan	0.927075	0.915066	0.084933	0.953288	0.046711
pushdo	0.792636	0.718433	0.281566	0.957022	0.042977
pushdotid	0.947667	0.940490	0.059509	0.950805	0.049194
pykspa	0.931104	0.919833	0.080166	0.956043	0.043956
pykspa2	0.947069	0.887578	0.112421	0.953400	0.046599
pykspa2s	0.923281	0.886032	0.113967	0.950335	0.049664
qadars	0.977612	0.988700	0.011299	0.953386	0.046613
qakbot	0.977180	0.989500	0.010499	0.950269	0.049730
ramdo	0.966438	0.998666	0.001333	0.952321	0.047678
ramnit	0.967731	0.978915	0.021084	0.951610	0.048389
ranbyus	0.981026	0.994833	0.005166	0.950777	0.049222

rovnix	0.955346	0.999473	0.000526	0.949204	0.050795
shifu	0.939293	0.884549	0.115450	0.948605	0.051394
simda	0.720757	0.525101	0.474898	0.954888	0.045111
sisron	0.960789	1.000000	0.000000	0.947794	0.052205
sphinx	0.980752	0.995299	0.004700	0.949238	0.050761
sutra	0.951088	1.000000	0.000000	0.948445	0.051554
szribi	0.928110	0.874457	0.125542	0.948996	0.051003
tempedreve	0.947227	0.921182	0.078817	0.947615	0.052384
tempedrevetdd	0.946000	0.907555	0.092444	0.949125	0.050874
tinba	0.975875	0.989333	0.010666	0.945841	0.054158
tinynuke	0.983791	1.000000	0.000000	0.948561	0.051438
tofsee	0.846329	0.402284	0.597715	0.950325	0.049674
torpig	0.931931	0.914961	0.085038	0.949730	0.050269
ud2	0.955155	1.000000	0.000000	0.953922	0.046077
ud3	0.950791	1.000000	0.000000	0.950580	0.049419
ud4	0.951904	0.956521	0.043478	0.951880	0.048119
urlzone	0.980246	0.994766	0.005233	0.948775	0.051224
vawtrak	0.918436	0.751018	0.248981	0.951673	0.048326
vidro	0.955763	0.958133	0.041866	0.950583	0.049416
vidrotid	0.948567	0.969849	0.030150	0.948260	0.051739
virut	0.775480	0.694766	0.305233	0.955057	0.044942
wd	0.984535	1.000000	0.000000	0.950382	0.049617
xshellghost	0.947468	0.944444	0.055555	0.947476	0.052523
xxhex	0.959923	0.998181	0.001818	0.947829	0.052170

Table B.2: LOGO results for individual malware families - best features from statistical tests

B.3 All features except digits features

Gaussian Naive Bayes					
Family	**Accuracy**	**TPR**	**FNR**	**TNR**	**FPR**
bamital	0.970859	1.000000	0.000000	0.907765	0.092234
bedep	0.912683	0.921282	0.078717	0.908002	0.091997
blackhole	0.914012	0.939643	0.060356	0.912650	0.087349
ccleaner	0.908958	1.000000	0.000000	0.908718	0.091281
chinad	0.963648	0.991533	0.008466	0.902883	0.097116
chir	0.906676	1.000000	0.000000	0.906004	0.093995
conficker	0.863129	0.842566	0.157433	0.907711	0.092288
corebot	0.958890	0.981133	0.018866	0.910398	0.089601
cryptolocker	0.894236	0.888066	0.111933	0.907690	0.092309
diamondfox	0.912530	0.957716	0.042283	0.910987	0.089012
dircrypt	0.909072	0.906875	0.093124	0.909257	0.090742
dmsniff	0.911830	0.811594	0.188405	0.912332	0.087667

dyre	0.970420	1.000000	0.000000	0.906030	0.093969
ebury	0.918772	0.992496	0.007503	0.907993	0.092006
ekforward	0.920557	0.996444	0.003555	0.914275	0.085724
emotet	0.906810	0.907333	0.092666	0.905671	0.094328
feodo	0.912366	0.963350	0.036649	0.911644	0.088355
fobber	0.909397	0.907953	0.092046	0.909608	0.090391
gameover	0.971347	0.998600	0.001399	0.910885	0.089114
goznym	0.913598	0.815709	0.184290	0.915966	0.084033
gspy	0.911158	1.000000	0.000000	0.910844	0.089155
hesperbot	0.907469	0.836158	0.163841	0.908385	0.091614
infy	0.944916	0.995962	0.004037	0.909130	0.090869
locky	0.882872	0.869266	0.130733	0.913378	0.086621
madmax	0.911096	0.952380	0.047619	0.910649	0.089350
makloader	0.913492	1.000000	0.000000	0.910318	0.089681
mirai	0.912571	0.953405	0.046594	0.911741	0.088258
modpack	0.909760	0.895238	0.104761	0.909870	0.090129
monerominer	0.971357	0.999933	0.000066	0.909150	0.090849
murofet	0.928510	0.933599	0.066400	0.917262	0.082737
murofetweekly	0.970753	0.999566	0.000433	0.907329	0.092670
mydoom	0.873114	0.644383	0.355616	0.909625	0.090374
necurs	0.899096	0.895499	0.104500	0.907090	0.092909
nymaim	0.863990	0.842700	0.157299	0.910229	0.089770
oderoor	0.890050	0.867941	0.132058	0.911980	0.088019
omexo	0.909382	1.000000	0.000000	0.909256	0.090743
padcrypt	0.804363	0.754633	0.245366	0.913030	0.086969
pandabanker	0.967778	0.999645	0.000354	0.908958	0.091041
proslikefan	0.865535	0.843566	0.156433	0.913489	0.086510
pushdo	0.469339	0.269399	0.730600	0.912272	0.087727
pushdotid	0.873529	0.792132	0.207867	0.909117	0.090882
pykspa	0.855988	0.831866	0.168133	0.909359	0.090640
pykspa2	0.890602	0.703678	0.296321	0.910494	0.089505
pykspa2s	0.816526	0.682699	0.317300	0.913725	0.086274
qadars	0.942259	0.956799	0.043200	0.910487	0.089512
qakbot	0.923423	0.930466	0.069533	0.908038	0.091961
ramdo	0.871439	0.780796	0.219203	0.911141	0.088858
ramnit	0.890567	0.876428	0.123571	0.910945	0.089054
ranbyus	0.916599	0.919300	0.080699	0.910684	0.089315
rovnix	0.920216	0.994207	0.005792	0.909917	0.090082
shifu	0.895245	0.834763	0.165236	0.905533	0.094466
simda	0.628526	0.389910	0.610089	0.914063	0.085936
sisron	0.932547	0.999559	0.000440	0.910338	0.089661
sphinx	0.901958	0.899166	0.100833	0.908007	0.091992
sutra	0.904682	0.938941	0.061058	0.902830	0.097169
szribi	0.859394	0.735321	0.264678	0.907694	0.092305

tempedreve	0.903479	0.793103	0.206896	0.905121	0.094878
tempedrevetdd	0.902158	0.802666	0.197333	0.910247	0.089752
tinba	0.901063	0.897900	0.102099	0.908123	0.091876
tinynuke	0.970892	1.000000	0.000000	0.907628	0.092371
tofsee	0.873044	0.716579	0.283420	0.909689	0.090310
torpig	0.853389	0.801345	0.198654	0.907973	0.092026
ud2	0.914971	1.000000	0.000000	0.912633	0.087366
ud3	0.913649	0.932203	0.067796	0.913570	0.086429
ud4	0.906512	0.811594	0.188405	0.906993	0.093006
urlzone	0.949362	0.968700	0.031299	0.907448	0.092551
vawtrak	0.836013	0.458688	0.541311	0.910923	0.089076
vidro	0.877630	0.865199	0.134800	0.904810	0.095189
vidrotid	0.906350	0.879396	0.120603	0.906739	0.093260
virut	0.723185	0.639233	0.360766	0.909967	0.090032
wd	0.970860	0.999966	0.000033	0.906581	0.093418
xshellghost	0.904168	0.805555	0.194444	0.904429	0.095570
xxhex	0.929238	1.000000	0.000000	0.906869	0.093130
Gradient Boosting Classifier					
Family	**Accuracy**	**TPR**	**FNR**	**TNR**	**FPR**
bamital	0.997833	1.000000	0.000000	0.993143	0.006856
bedep	0.993097	0.995440	0.004559	0.991822	0.008177
blackhole	0.993907	0.995884	0.004115	0.993802	0.006197
ccleaner	0.992419	1.000000	0.000000	0.992399	0.007600
chinad	0.997440	0.999800	0.000199	0.992300	0.007699
chir	0.992638	1.000000	0.000000	0.992585	0.007414
conficker	0.889362	0.841600	0.158399	0.992917	0.007082
corebot	0.996960	0.998399	0.001600	0.993823	0.006176
cryptolocker	0.992595	0.992433	0.007566	0.992949	0.007050
diamondfox	0.992530	0.976744	0.023255	0.993069	0.006930
dircrypt	0.991617	0.986074	0.013925	0.992083	0.007916
dmsniff	0.992122	0.956521	0.043478	0.992300	0.007699
dyre	0.997647	1.000000	0.000000	0.992525	0.007474
ebury	0.993172	0.998999	0.001000	0.992320	0.007679
ekforward	0.993340	0.999111	0.000888	0.992862	0.007137
emotet	0.996961	0.998900	0.001099	0.992738	0.007261
feodo	0.993129	1.000000	0.000000	0.993032	0.006967
fobber	0.992163	0.989994	0.010005	0.992479	0.007520
gameover	0.997610	1.000000	0.000000	0.992308	0.007691
goznym	0.990011	0.888217	0.111782	0.992473	0.007526
gspy	0.991196	1.000000	0.000000	0.991165	0.008834
hesperbot	0.992981	0.971751	0.028248	0.993254	0.006745
infy	0.993045	0.990682	0.009317	0.994701	0.005298
locky	0.977547	0.970333	0.029666	0.993721	0.006278
madmax	0.992718	0.986394	0.013605	0.992787	0.007212

makloader	0.991619	1.000000	0.000000	0.991311	0.008688
mirai	0.993428	1.000000	0.000000	0.993294	0.006705
modpack	0.992134	0.866666	0.133333	0.993083	0.006916
monerominer	0.997670	0.999900	0.000099	0.992816	0.007183
murofet	0.996213	0.997833	0.002166	0.992632	0.007367
murofetweekly	0.997616	1.000000	0.000000	0.992369	0.007630
mydoom	0.983787	0.930422	0.069577	0.992305	0.007694
necurs	0.984688	0.981633	0.018366	0.991479	0.008520
nymaim	0.912423	0.874900	0.125099	0.993918	0.006081
oderoor	0.983596	0.975873	0.024126	0.991257	0.008742
omexo	0.992272	1.000000	0.000000	0.992261	0.007738
padcrypt	0.979327	0.972899	0.027100	0.993371	0.006628
pandabanker	0.997493	0.999802	0.000197	0.993231	0.006768
proslikefan	0.973002	0.963633	0.036366	0.993451	0.006548
pushdo	0.797827	0.709766	0.290233	0.992910	0.007089
pushdotid	0.978448	0.945157	0.054842	0.993003	0.006996
pykspa	0.979430	0.973300	0.026699	0.992993	0.007006
pykspa2	0.985582	0.897987	0.102012	0.994904	0.005095
pykspa2s	0.935744	0.857817	0.142182	0.992342	0.007657
qadars	0.995586	0.996933	0.003066	0.992643	0.007356
qakbot	0.993689	0.994099	0.005900	0.992791	0.007208
ramdo	0.993500	0.994665	0.005334	0.992990	0.007009
ramnit	0.988179	0.984983	0.015016	0.992785	0.007214
ranbyus	0.994209	0.995099	0.004900	0.992258	0.007741
rovnix	0.994916	1.000000	0.000000	0.994209	0.005790
shifu	0.990391	0.984549	0.015450	0.991385	0.008614
simda	0.825556	0.685606	0.314393	0.993026	0.006973
sisron	0.993857	1.000000	0.000000	0.991822	0.008177
sphinx	0.995074	0.996099	0.003900	0.992851	0.007148
sutra	0.991163	0.997286	0.002713	0.990833	0.009166
szribi	0.977359	0.937511	0.062488	0.992871	0.007128
tempedreve	0.991986	0.940886	0.059113	0.992746	0.007253
tempedrevetdd	0.989039	0.943999	0.056000	0.992701	0.007298
tinba	0.992104	0.992133	0.007866	0.992039	0.007960
tinynuke	0.997945	1.000000	0.000000	0.993479	0.006520
tofsee	0.976917	0.908305	0.091694	0.992986	0.007013
torpig	0.970635	0.948186	0.051813	0.994180	0.005819
ud2	0.993926	1.000000	0.000000	0.993759	0.006240
ud3	0.993207	1.000000	0.000000	0.993178	0.006821
ud4	0.992617	0.956521	0.043478	0.992800	0.007199
urlzone	0.996145	0.997666	0.002333	0.992847	0.007152
vawtrak	0.959064	0.787699	0.212300	0.993085	0.006914
vidro	0.982159	0.976866	0.023133	0.993731	0.006268
vidrotid	0.991927	0.989949	0.010050	0.991956	0.008043

virut	0.874850	0.821999	0.178000	0.992435	0.007564
wd	0.997705	1.000000	0.000000	0.992638	0.007361
xshellghost	0.992893	0.944444	0.055555	0.993021	0.006978
xxhex	0.994157	0.998408	0.001591	0.992814	0.007185
Logistic Regression					
Family	**Accuracy**	**TPR**	**FNR**	**TNR**	**FPR**
bamital	0.987162	1.000000	0.000000	0.959367	0.040632
bedep	0.970500	0.994635	0.005364	0.957359	0.042640
blackhole	0.963445	0.991769	0.008230	0.961939	0.038060
ccleaner	0.961075	1.000000	0.000000	0.960973	0.039026
chinad	0.985856	0.998633	0.001366	0.958015	0.041984
chir	0.959725	1.000000	0.000000	0.959435	0.040564
conficker	0.911672	0.888766	0.111233	0.961335	0.038664
corebot	0.981033	0.990266	0.009733	0.960904	0.039095
cryptolocker	0.982997	0.993500	0.006499	0.960095	0.039904
diamondfox	0.958394	0.915433	0.084566	0.959861	0.040138
dircrypt	0.960451	0.983463	0.016536	0.958513	0.041486
dmsniff	0.958372	0.927536	0.072463	0.958527	0.041472
dyre	0.986889	1.000000	0.000000	0.958348	0.041651
ebury	0.963884	0.998499	0.001500	0.958823	0.041176
ekforward	0.940944	0.694222	0.305777	0.961368	0.038631
emotet	0.982819	0.993766	0.006233	0.958971	0.041028
feodo	0.963236	1.000000	0.000000	0.962715	0.037284
fobber	0.964383	0.990495	0.009504	0.960572	0.039427
gameover	0.987776	0.999199	0.000800	0.962431	0.037568
goznym	0.959474	0.918429	0.081570	0.960467	0.039532
gspy	0.961264	1.000000	0.000000	0.961127	0.038872
hesperbot	0.958461	0.966101	0.033898	0.958363	0.041636
infy	0.901309	0.813024	0.186975	0.963202	0.036797
locky	0.959958	0.959833	0.040166	0.960239	0.039760
madmax	0.960608	0.993197	0.006802	0.960256	0.039743
makloader	0.957819	1.000000	0.000000	0.956271	0.043728
mirai	0.960571	1.000000	0.000000	0.959769	0.040230
modpack	0.958670	0.885714	0.114285	0.959221	0.040778
monerominer	0.987688	0.999633	0.000366	0.961686	0.038313
murofet	0.985380	0.997399	0.002600	0.958815	0.041184
murofetweekly	0.981869	0.992299	0.007700	0.958911	0.041088
mydoom	0.948920	0.874488	0.125511	0.960801	0.039198
necurs	0.970733	0.976866	0.023133	0.957101	0.042898
nymaim	0.901604	0.872333	0.127666	0.965177	0.034822
oderoor	0.962730	0.968778	0.031221	0.956731	0.043268
omexo	0.955529	1.000000	0.000000	0.955467	0.044532
padcrypt	0.984335	0.994866	0.005133	0.961322	0.038677
pandabanker	0.979516	0.988566	0.011433	0.962812	0.037187

proslikefan	0.930253	0.915733	0.084266	0.961947	0.038052
pushdo	0.774792	0.689033	0.310966	0.964776	0.035223
pushdotid	0.956237	0.947157	0.052842	0.960206	0.039793
pykspa	0.954016	0.950600	0.049399	0.961575	0.038424
pykspa2	0.960619	0.930603	0.069396	0.963813	0.036186
pykspa2s	0.946897	0.930515	0.069484	0.958795	0.041204
qadars	0.980951	0.988833	0.011166	0.963729	0.036270
qakbot	0.982233	0.993099	0.006900	0.958497	0.041502
ramdo	0.972226	0.998499	0.001500	0.960718	0.039281
ramnit	0.972537	0.982101	0.017898	0.958752	0.041247
ranbyus	0.984871	0.996133	0.003866	0.960198	0.039801
rovnix	0.963904	0.998946	0.001053	0.959026	0.040973
shifu	0.952208	0.909871	0.090128	0.959410	0.040589
simda	0.770693	0.610999	0.389000	0.961789	0.038210
sisron	0.968247	1.000000	0.000000	0.957724	0.042275
sphinx	0.985085	0.997266	0.002733	0.958697	0.041302
sutra	0.959229	1.000000	0.000000	0.957025	0.042974
szribi	0.939007	0.895223	0.104776	0.956052	0.043947
tempedreve	0.958273	0.921182	0.078817	0.958824	0.041175
tempedrevetdd	0.955423	0.924444	0.075555	0.957941	0.042058
tinba	0.981446	0.991633	0.008366	0.958711	0.041288
tinynuke	0.987010	1.000000	0.000000	0.958777	0.041222
tofsee	0.829222	0.278172	0.721827	0.958279	0.041720
torpig	0.947093	0.935492	0.064507	0.959260	0.040739
ud2	0.961723	1.000000	0.000000	0.960670	0.039329
ud3	0.960835	0.898305	0.101694	0.961103	0.038896
ud4	0.961113	0.927536	0.072463	0.961284	0.038715
urlzone	0.983987	0.996199	0.003800	0.957517	0.042482
vawtrak	0.928439	0.768803	0.231196	0.960132	0.039867
vidro	0.966971	0.970966	0.029033	0.958236	0.041763
vidrotid	0.956425	0.969849	0.030150	0.956231	0.043768
virut	0.795304	0.720333	0.279666	0.962103	0.037896
wd	0.987288	1.000000	0.000000	0.959216	0.040783
xshellghost	0.957652	0.944444	0.055555	0.957687	0.042312
xxhex	0.959650	0.962491	0.037508	0.958752	0.041247
Random Forests					
Family	**Accuracy**	**TPR**	**FNR**	**TNR**	**FPR**
bamital	0.999475	1.000000	0.000000	0.998340	0.001659
bedep	0.998203	0.996647	0.003352	0.999050	0.000949
blackhole	0.998476	0.998628	0.001371	0.998468	0.001531
ccleaner	0.998542	1.000000	0.000000	0.998538	0.001461
chinad	0.999497	0.999833	0.000166	0.998765	0.001234
chir	0.998412	1.000000	0.000000	0.998400	0.001599
conficker	0.940187	0.912966	0.087033	0.999205	0.000794

corebot	0.998903	0.999166	0.000833	0.998328	0.001671
cryptolocker	0.995292	0.993800	0.006199	0.998546	0.001453
diamondfox	0.998394	0.983086	0.016913	0.998917	0.001082
dircrypt	0.998107	0.989556	0.010443	0.998827	0.001172
dmsniff	0.998843	1.000000	0.000000	0.998837	0.001162
dyre	0.999588	1.000000	0.000000	0.998693	0.001306
ebury	0.998404	0.998999	0.001000	0.998317	0.001682
ekforward	0.998029	0.991111	0.008888	0.998601	0.001398
emotet	0.997806	0.997433	0.002566	0.998620	0.001379
feodo	0.998976	1.000000	0.000000	0.998962	0.001037
fobber	0.997642	0.990495	0.009504	0.998685	0.001314
gameover	0.999425	1.000000	0.000000	0.998151	0.001848
goznym	0.997003	0.945619	0.054380	0.998246	0.001753
gspy	0.998459	1.000000	0.000000	0.998453	0.001546
hesperbot	0.998567	0.983050	0.016949	0.998766	0.001233
infy	0.997909	0.996479	0.003520	0.998911	0.001088
locky	0.984993	0.978899	0.021100	0.998654	0.001345
madmax	0.998325	1.000000	0.000000	0.998307	0.001692
makloader	0.998268	1.000000	0.000000	0.998204	0.001795
mirai	0.999142	0.996415	0.003584	0.999198	0.000801
modpack	0.997211	0.847619	0.152380	0.998342	0.001657
monerominer	0.999406	0.999900	0.000099	0.998331	0.001668
murofet	0.998347	0.998299	0.001700	0.998452	0.001547
murofetweekly	0.999610	1.000000	0.000000	0.998752	0.001247
mydoom	0.990923	0.942246	0.057753	0.998693	0.001306
necurs	0.984895	0.978766	0.021233	0.998518	0.001481
nymaim	0.952593	0.931566	0.068433	0.998262	0.001737
oderoor	0.997247	0.996041	0.003958	0.998444	0.001555
omexo	0.998469	1.000000	0.000000	0.998466	0.001533
padcrypt	0.985089	0.978899	0.021100	0.998616	0.001383
pandabanker	0.999309	0.999605	0.000394	0.998762	0.001237
proslikefan	0.979882	0.971366	0.028633	0.998472	0.001527
pushdo	0.835308	0.761566	0.238433	0.998670	0.001329
pushdotid	0.985192	0.954992	0.045007	0.998396	0.001603
pykspa	0.984848	0.978400	0.021599	0.999114	0.000885
pykspa2	0.996595	0.975017	0.024982	0.998892	0.001107
pykspa2s	0.979003	0.952706	0.047293	0.998103	0.001896
qadars	0.997324	0.996633	0.003366	0.998834	0.001165
qakbot	0.996913	0.995966	0.004033	0.998980	0.001019
ramdo	0.998121	0.996499	0.003500	0.998831	0.001168
ramnit	0.992626	0.988674	0.011325	0.998323	0.001676
ranbyus	0.997665	0.997066	0.002933	0.998977	0.001022
rovnix	0.998777	1.000000	0.000000	0.998607	0.001392
shifu	0.995819	0.979399	0.020600	0.998612	0.001387

simda	0.864545	0.752504	0.247495	0.998619	0.001380
sisron	0.999396	1.000000	0.000000	0.999196	0.000803
sphinx	0.997719	0.997399	0.002600	0.998411	0.001588
sutra	0.998608	0.998643	0.001356	0.998606	0.001393
szribi	0.986193	0.955068	0.044931	0.998309	0.001690
tempedreve	0.998050	0.940886	0.059113	0.998901	0.001098
tempedrevetdd	0.995789	0.959111	0.040888	0.998771	0.001228
tinba	0.994751	0.993166	0.006833	0.998288	0.001711
tinynuke	0.999543	1.000000	0.000000	0.998551	0.001448
tofsee	0.985060	0.927755	0.072244	0.998481	0.001518
torpig	0.979441	0.960879	0.039120	0.998908	0.001091
ud2	0.998587	1.000000	0.000000	0.998548	0.001451
ud3	0.998193	1.000000	0.000000	0.998185	0.001814
ud4	0.998391	1.000000	0.000000	0.998383	0.001616
urlzone	0.998517	0.998299	0.001700	0.998988	0.001011
vawtrak	0.971646	0.835865	0.164134	0.998602	0.001397
vidro	0.991834	0.988966	0.011033	0.998104	0.001895
vidrotid	0.998499	0.989949	0.010050	0.998623	0.001376
virut	0.939104	0.912333	0.087666	0.998665	0.001334
wd	0.999449	1.000000	0.000000	0.998233	0.001766
xshellghost	0.999047	0.944444	0.055555	0.999191	0.000808
xxhex	0.998798	0.999318	0.000681	0.998634	0.001365
Support Vector Classifier					
Family	**Accuracy**	**TPR**	**FNR**	**TNR**	**FPR**
bamital	0.987230	1.000000	0.000000	0.959584	0.040415
bedep	0.971257	0.995172	0.004827	0.958235	0.041764
blackhole	0.963652	0.993141	0.006858	0.962085	0.037914
ccleaner	0.961148	1.000000	0.000000	0.961046	0.038953
chinad	0.986016	0.998766	0.001233	0.958233	0.041766
chir	0.960014	1.000000	0.000000	0.959726	0.040273
conficker	0.909984	0.885966	0.114033	0.962058	0.037941
corebot	0.981536	0.990366	0.009633	0.962284	0.037715
cryptolocker	0.982997	0.993466	0.006533	0.960168	0.039831
diamondfox	0.958813	0.913319	0.086680	0.960366	0.039633
dircrypt	0.961533	0.984334	0.015665	0.959612	0.040387
dmsniff	0.958300	0.927536	0.072463	0.958454	0.041545
dyre	0.987117	1.000000	0.000000	0.959074	0.040925
ebury	0.964203	0.998499	0.001500	0.959189	0.040810
ekforward	0.942779	0.709333	0.290666	0.962104	0.037895
emotet	0.983025	0.993866	0.006133	0.959407	0.040592
feodo	0.963090	1.000000	0.000000	0.962567	0.037432
fobber	0.964829	0.989494	0.010505	0.961229	0.038770
gameover	0.988052	0.999166	0.000833	0.963392	0.036607
goznym	0.959902	0.912386	0.087613	0.961052	0.038947

gspy	0.962511	1.000000	0.000000	0.962379	0.037620
hesperbot	0.958390	0.966101	0.033898	0.958291	0.041708
infy	0.905149	0.822341	0.177658	0.963202	0.036797
locky	0.960165	0.959666	0.040333	0.961285	0.038714
madmax	0.961409	0.993197	0.006802	0.961065	0.038934
makloader	0.958304	1.000000	0.000000	0.956774	0.043225
mirai	0.960428	1.000000	0.000000	0.959623	0.040376
modpack	0.958455	0.885714	0.114285	0.959005	0.040994
murofet	0.987780	0.999633	0.000366	0.961976	0.038023
monerominer	0.985839	0.997500	0.002499	0.960067	0.039932
murofetweekly	0.973160	0.979433	0.020566	0.959351	0.040648
mydoom	0.948732	0.869940	0.130059	0.961309	0.038690
necurs	0.970480	0.976633	0.023366	0.956805	0.043194
nymaim	0.900143	0.870066	0.129933	0.965467	0.034532
oderoor	0.963622	0.969002	0.030997	0.958287	0.041712
omexo	0.956112	1.000000	0.000000	0.956051	0.043948
padcrypt	0.984541	0.994833	0.005166	0.962051	0.037948
pandabanker	0.978263	0.986476	0.013523	0.963103	0.036896
proslikefan	0.929064	0.913633	0.086366	0.962747	0.037252
pushdo	0.775274	0.689433	0.310566	0.965440	0.034559
pushdotid	0.957302	0.948824	0.051175	0.961008	0.038991
pykspa	0.953557	0.949766	0.050233	0.961944	0.038055
pykspa2	0.961353	0.929909	0.070090	0.964699	0.035300
pykspa2s	0.945798	0.926699	0.073300	0.959670	0.040329
qadars	0.981522	0.989466	0.010533	0.964166	0.035833
qakbot	0.982462	0.993099	0.006900	0.959225	0.040774
ramdo	0.972835	0.998499	0.001500	0.961594	0.038405
ramnit	0.973044	0.982404	0.017595	0.959554	0.040445
ranbyus	0.985009	0.996166	0.003833	0.960563	0.039436
rovnix	0.964225	0.998946	0.001053	0.959393	0.040606
shifu	0.952582	0.906866	0.093133	0.960359	0.039640
simda	0.771222	0.611546	0.388453	0.962298	0.037701
sisron	0.968247	1.000000	0.000000	0.957724	0.042275
sphinx	0.985313	0.997266	0.002733	0.959419	0.040580
sutra	0.959298	1.000000	0.000000	0.957098	0.042901
szribi	0.938796	0.892014	0.107985	0.957007	0.042992
tempedreve	0.958489	0.921182	0.078817	0.959044	0.040955
tempedrevetdd	0.956359	0.922666	0.077333	0.959098	0.040901
tinba	0.981699	0.991766	0.008233	0.959232	0.040767
tinynuke	0.987032	1.000000	0.000000	0.958849	0.041150
tofsee	0.823129	0.244211	0.755788	0.958712	0.041287
torpig	0.947271	0.935562	0.064437	0.959551	0.040448
ud2	0.962288	1.000000	0.000000	0.961250	0.038749
ud3	0.961341	0.898305	0.101694	0.961611	0.038388

ud4	0.961844	0.927536	0.072463	0.962018	0.037981
urlzone	0.984352	0.996199	0.003800	0.958673	0.041326
vawtrak	0.927948	0.766580	0.233419	0.959985	0.040014
vidro	0.967200	0.970899	0.029100	0.959110	0.040889
vidrotid	0.957139	0.974874	0.025125	0.956884	0.043115
virut	0.784495	0.704466	0.295533	0.962548	0.037451
wd	0.987724	1.000000	0.000000	0.960615	0.039384
xshellghost	0.958165	0.944444	0.055555	0.958201	0.041798
xxhex	0.959049	0.960900	0.039099	0.958465	0.041534

Table B.3: LOGO results for individual malware families - all features except digits features

B.4 All features except ngrams features

Gaussian Naive Bayes					
Family	**Accuracy**	**TPR**	**FNR**	**TNR**	**FPR**
bamital	0.987185	1.000000	0.000000	0.959439	0.040560
bedep	0.740462	0.341558	0.658441	0.957651	0.042348
blackhole	0.927305	0.334705	0.665294	0.958804	0.041195
ccleaner	0.958743	1.000000	0.000000	0.958634	0.041365
chinad	0.936299	0.924466	0.075533	0.962083	0.037916
chir	0.959870	1.000000	0.000000	0.959581	0.040418
conficker	0.323334	0.032233	0.967766	0.954469	0.045530
corebot	0.947761	0.942866	0.057133	0.958433	0.041566
cryptolocker	0.439622	0.201300	0.798699	0.959296	0.040703
diamondfox	0.950366	0.735729	0.264270	0.957695	0.042304
dircrypt	0.901027	0.208877	0.791122	0.959319	0.040680
dmsniff	0.954036	0.014492	0.985507	0.958744	0.041255
dyre	0.987665	1.000000	0.000000	0.960815	0.039184
ebury	0.956355	0.931465	0.068534	0.959994	0.040005
ekforward	0.959836	0.966222	0.033777	0.959308	0.040691
emotet	0.505380	0.295899	0.704099	0.961731	0.038268
feodo	0.950665	0.256544	0.743455	0.960492	0.039507
fobber	0.870659	0.254627	0.745372	0.960572	0.039427
gameover	0.986305	0.997600	0.002399	0.961248	0.038751
goznym	0.937642	0.018126	0.981873	0.959883	0.040116
gspy	0.958036	1.000000	0.000000	0.957888	0.042111
hesperbot	0.948864	0.203389	0.796610	0.958436	0.041563
infy	0.968297	0.983849	0.016150	0.957395	0.042604
locky	0.406270	0.157700	0.842300	0.963602	0.036397
madmax	0.953473	0.380952	0.619047	0.959667	0.040332
makloader	0.960590	1.000000	0.000000	0.959144	0.040855
mirai	0.952642	0.663082	0.336917	0.958530	0.041469

modpack	0.956167	0.638095	0.361904	0.958573	0.041426
monerominer	0.987574	0.999900	0.000099	0.960743	0.039256
murofet	0.528446	0.332399	0.667599	0.961762	0.038237
murofetweekly	0.987118	1.000000	0.000000	0.958764	0.041235
mydoom	0.833802	0.059572	0.940427	0.957389	0.042610
necurs	0.489780	0.278866	0.721133	0.958583	0.041416
nymaim	0.329879	0.040599	0.959400	0.958155	0.041844
oderoor	0.512404	0.064983	0.935016	0.956212	0.043787
omexo	0.961216	1.000000	0.000000	0.961162	0.038837
padcrypt	0.535457	0.341299	0.658700	0.959720	0.040279
pandabanker	0.985167	0.999566	0.000433	0.958591	0.041408
proslikefan	0.317346	0.022166	0.977833	0.961655	0.038344
pushdo	0.298447	0.000000	1.000000	0.959607	0.040392
pushdotid	0.683316	0.055509	0.944490	0.957801	0.042198
pykspa	0.337450	0.057566	0.942433	0.956707	0.043292
pykspa2	0.872179	0.041637	0.958362	0.960564	0.039435
pykspa2s	0.574753	0.043277	0.956722	0.960764	0.039235
qadars	0.803018	0.731533	0.268466	0.959213	0.040786
qakbot	0.565189	0.385666	0.614333	0.957332	0.042667
ramdo	0.676263	0.027337	0.972662	0.960499	0.039500
ramnit	0.510417	0.197593	0.802406	0.961303	0.038696
ranbyus	0.532854	0.339366	0.660633	0.956766	0.043233
rovnix	0.958756	0.942601	0.057398	0.961005	0.038994
shifu	0.822123	0.000000	1.000000	0.961965	0.038034
simda	0.438010	0.000000	1.000000	0.962153	0.037846
sisron	0.719111	0.000000	1.000000	0.957432	0.042567
sphinx	0.493511	0.277566	0.722433	0.961296	0.038703
sutra	0.930216	0.362279	0.637720	0.960912	0.039087
szribi	0.688584	0.000000	1.000000	0.956639	0.043360
tempedreve	0.943329	0.039408	0.960591	0.956773	0.043226
tempedrevetdd	0.890195	0.027555	0.972444	0.960326	0.039673
tinba	0.407485	0.159799	0.840200	0.960273	0.039726
tinynuke	0.986781	1.000000	0.000000	0.958052	0.041947
tofsee	0.779190	0.000000	1.000000	0.961677	0.038322
torpig	0.488371	0.039467	0.960532	0.959188	0.040811
ud2	0.961370	1.000000	0.000000	0.960307	0.039692
ud3	0.962280	1.000000	0.000000	0.962119	0.037880
ud4	0.954681	0.014492	0.985507	0.959447	0.040552
urlzone	0.704477	0.585999	0.414000	0.961274	0.038725
vawtrak	0.802197	0.014449	0.985550	0.958587	0.041412
vidro	0.344533	0.063566	0.936433	0.958892	0.041107
vidrotid	0.944496	0.100502	0.899497	0.956666	0.043333
virut	0.297534	0.000000	1.000000	0.959507	0.040492
wd	0.988023	1.000000	0.000000	0.961572	0.038427

xshellghost	0.955088	0.222222	0.777777	0.957026	0.042973
xxhex	0.967185	0.993407	0.006592	0.958896	0.041103
Gradient Boosting Classifier					
Family	**Accuracy**	**TPR**	**FNR**	**TNR**	**FPR**
bamital	0.975670	1.000000	0.000000	0.922993	0.077006
bedep	0.890511	0.836261	0.163738	0.920049	0.079950
blackhole	0.917474	0.943758	0.056241	0.916077	0.083922
ccleaner	0.919381	1.000000	0.000000	0.919169	0.080830
chinad	0.969269	0.992033	0.007966	0.919662	0.080337
chir	0.917069	1.000000	0.000000	0.916472	0.083527
conficker	0.737048	0.649900	0.350099	0.925995	0.074004
corebot	0.950595	0.964333	0.035666	0.920645	0.079354
cryptolocker	0.884866	0.869633	0.130366	0.918084	0.081915
diamondfox	0.894171	0.194503	0.805496	0.918062	0.081937
dircrypt	0.911235	0.781549	0.218450	0.922157	0.077842
dmsniff	0.919346	0.840579	0.159420	0.919741	0.080258
dyre	0.975080	1.000000	0.000000	0.920833	0.079166
ebury	0.930768	1.000000	0.000000	0.920646	0.079353
ekforward	0.920897	0.980444	0.019555	0.915967	0.084032
emotet	0.958991	0.976199	0.023800	0.921501	0.078498
feodo	0.916678	0.952879	0.047120	0.916166	0.083833
fobber	0.915387	0.828914	0.171085	0.928008	0.071991
gameover	0.975805	0.999266	0.000733	0.923753	0.076246
goznym	0.915596	0.522658	0.477341	0.925100	0.074899
gspy	0.916587	1.000000	0.000000	0.916292	0.083707
hesperbot	0.913986	0.644067	0.355932	0.917452	0.082547
infy	0.672142	0.316285	0.683714	0.921614	0.078385
locky	0.822360	0.778366	0.221633	0.921001	0.078998
madmax	0.921945	0.863945	0.136054	0.922573	0.077426
makloader	0.922842	0.998043	0.001956	0.920083	0.079916
mirai	0.916785	0.860215	0.139784	0.917936	0.082063
modpack	0.920986	0.780952	0.219047	0.922046	0.077953
monerominer	0.964048	0.985333	0.014666	0.917712	0.082287
murofet	0.944897	0.953699	0.046300	0.925440	0.074559
murofetweekly	0.974695	0.999933	0.000066	0.919143	0.080856
mydoom	0.905539	0.833560	0.166439	0.917029	0.082970
necurs	0.811688	0.762299	0.237700	0.921464	0.078535
nymaim	0.695364	0.590999	0.409000	0.922029	0.077970
oderoor	0.825032	0.726172	0.273827	0.923094	0.076905
omexo	0.916672	1.000000	0.000000	0.916557	0.083442
padcrypt	0.864071	0.838633	0.161366	0.919659	0.080340
pandabanker	0.972381	0.999921	0.000078	0.921548	0.078451
proslikefan	0.768013	0.697100	0.302899	0.922802	0.077197
pushdo	0.287285	0.000000	1.000000	0.923718	0.076281

pushdotid	0.837728	0.651275	0.348724	0.919247	0.080752
pykspa	0.822447	0.777066	0.222933	0.922855	0.077144
pykspa2	0.897543	0.665510	0.334489	0.922236	0.077763
pykspa2s	0.818005	0.673561	0.326438	0.922914	0.077085
qadars	0.914269	0.910000	0.089999	0.923597	0.076402
qakbot	0.906457	0.900066	0.099933	0.920416	0.079583
ramdo	0.829398	0.616269	0.383730	0.922751	0.077248
ramnit	0.816328	0.743199	0.256800	0.921731	0.078268
ranbyus	0.859336	0.832500	0.167499	0.918133	0.081866
rovnix	0.930510	0.983149	0.016850	0.923184	0.076815
shifu	0.926628	0.941201	0.058798	0.924149	0.075850
simda	0.427031	0.010926	0.989073	0.924960	0.075039
sisron	0.927008	0.956378	0.043621	0.917275	0.082724
sphinx	0.862254	0.834133	0.165866	0.923171	0.076828
sutra	0.916370	0.979647	0.020352	0.912951	0.087048
szribi	0.750581	0.312818	0.687181	0.920996	0.079003
tempedreve	0.918134	0.709359	0.290640	0.921239	0.078760
tempedrevetdd	0.905834	0.700444	0.299555	0.922532	0.077467
tinba	0.897955	0.884766	0.115233	0.927391	0.072608
tinynuke	0.975412	1.000000	0.000000	0.921973	0.078026
tofsee	0.828695	0.438715	0.561284	0.920028	0.079971
torpig	0.792032	0.660331	0.339668	0.930161	0.069838
ud2	0.920409	1.000000	0.000000	0.918220	0.081779
ud3	0.918707	0.932203	0.067796	0.918650	0.081349
ud4	0.924128	0.826086	0.173913	0.924625	0.075374
urlzone	0.897037	0.885666	0.114333	0.921681	0.078318
vawtrak	0.817356	0.284920	0.715079	0.923059	0.076940
vidro	0.777012	0.709066	0.290933	0.925583	0.074416
vidrotid	0.918494	0.793969	0.206030	0.920289	0.079710
virut	0.288703	0.000000	1.000000	0.931029	0.068970
wd	0.973155	1.000000	0.000000	0.913869	0.086130
xshellghost	0.924756	0.722222	0.277777	0.925292	0.074707
xxhex	0.943325	0.997044	0.002955	0.926343	0.073656
Logistic Regression					
Family	**Accuracy**	**TPR**	**FNR**	**TNR**	**FPR**
bamital	0.959663	0.999933	0.000066	0.872474	0.127525
bedep	0.883704	0.906664	0.093335	0.871203	0.128796
blackhole	0.870672	0.890260	0.109739	0.869631	0.130368
ccleaner	0.872804	1.000000	0.000000	0.872469	0.127530
chinad	0.958004	0.998766	0.001233	0.869179	0.130820
chir	0.867845	1.000000	0.000000	0.866894	0.133105
conficker	0.760293	0.705233	0.294766	0.879670	0.120329
corebot	0.955462	0.995966	0.004033	0.867160	0.132839
cryptolocker	0.887220	0.894100	0.105899	0.872219	0.127780

diamondfox	0.847260	0.298097	0.701902	0.866012	0.133987
dircrypt	0.865467	0.795474	0.204525	0.871362	0.128637
dmsniff	0.873599	0.797101	0.202898	0.873983	0.126016
dyre	0.957789	1.000000	0.000000	0.865902	0.134097
ebury	0.889739	0.999499	0.000500	0.873692	0.126307
ekforward	0.834386	0.392000	0.607999	0.871008	0.128991
emotet	0.844623	0.829899	0.170100	0.876697	0.123302
feodo	0.870559	0.947643	0.052356	0.869468	0.130531
fobber	0.868875	0.855427	0.144572	0.870838	0.129161
gameover	0.959836	0.998999	0.001000	0.872947	0.127052
goznym	0.861586	0.522658	0.477341	0.869784	0.130215
gspy	0.876384	1.000000	0.000000	0.875947	0.124052
hesperbot	0.865716	0.751412	0.248587	0.867184	0.132815
infy	0.559969	0.117921	0.882078	0.869865	0.130134
locky	0.809751	0.779900	0.220099	0.876681	0.123318
madmax	0.874326	0.931972	0.068027	0.873702	0.126297
makloader	0.867086	0.853228	0.146771	0.867595	0.132404
mirai	0.874071	0.903225	0.096774	0.873478	0.126521
modpack	0.862710	0.761904	0.238095	0.863472	0.136527
monerominer	0.957515	0.999466	0.000533	0.866192	0.133807
murofet	0.919353	0.938633	0.061366	0.876740	0.123259
murofetweekly	0.958101	0.999900	0.000099	0.866094	0.133905
mydoom	0.854773	0.763528	0.236471	0.869337	0.130662
necurs	0.817458	0.790266	0.209733	0.877898	0.122101
nymaim	0.712802	0.637199	0.362800	0.876999	0.123000
oderoor	0.834554	0.800418	0.199581	0.868415	0.131584
omexo	0.874097	1.000000	0.000000	0.873923	0.126076
padcrypt	0.605753	0.482833	0.517166	0.874353	0.125646
pandabanker	0.953636	0.999881	0.000118	0.868277	0.131722
proslikefan	0.715183	0.639433	0.360566	0.880529	0.119470
pushdo	0.275664	0.000000	1.000000	0.886353	0.113646
pushdotid	0.830679	0.731455	0.268544	0.874061	0.125938
pykspa	0.798595	0.762833	0.237166	0.877719	0.122280
pykspa2	0.840008	0.563497	0.436502	0.869433	0.130566
pykspa2s	0.746947	0.569836	0.430163	0.875583	0.124416
qadars	0.909741	0.928833	0.071166	0.868026	0.131973
qakbot	0.884323	0.889100	0.110899	0.873889	0.126110
ramdo	0.721350	0.376396	0.623603	0.872444	0.127555
ramnit	0.810149	0.766407	0.233592	0.873196	0.126803
ranbyus	0.886869	0.892700	0.107299	0.874096	0.125903
rovnix	0.891970	0.999473	0.000526	0.877006	0.122993
shifu	0.847641	0.685836	0.314163	0.875164	0.124835
simda	0.398855	0.000000	1.000000	0.876144	0.123855
sisron	0.885988	0.940515	0.059484	0.867917	0.132082

sphinx	0.830851	0.810433	0.189566	0.875081	0.124918
sutra	0.867598	0.940298	0.059701	0.863669	0.136330
szribi	0.722757	0.327921	0.672078	0.876460	0.123539
tempedreve	0.866806	0.699507	0.300492	0.869294	0.130705
tempedrevetdd	0.859854	0.710222	0.289777	0.872019	0.127980
tinba	0.868721	0.862133	0.137866	0.883425	0.116574
tinynuke	0.959660	0.999900	0.000099	0.872201	0.127798
tofsee	0.729333	0.097252	0.902747	0.877368	0.122631
torpig	0.836132	0.795241	0.204758	0.879019	0.120980
ud2	0.873375	1.000000	0.000000	0.869893	0.130106
ud3	0.868198	1.000000	0.000000	0.867634	0.132365
ud4	0.873254	0.797101	0.202898	0.873640	0.126359
urlzone	0.937638	0.969400	0.030599	0.868795	0.131204
vawtrak	0.777218	0.282697	0.717302	0.875395	0.124604
vidro	0.810590	0.778766	0.221233	0.880174	0.119825
vidrotid	0.867419	0.773869	0.226130	0.868768	0.131231
virut	0.345368	0.102733	0.897266	0.885197	0.114802
wd	0.958425	1.000000	0.000000	0.866607	0.133392
xshellghost	0.872078	0.750000	0.250000	0.872401	0.127598
xxhex	0.899099	0.983859	0.016140	0.872305	0.127694
Random Forests					
Family	**Accuracy**	**TPR**	**FNR**	**TNR**	**FPR**
bamital	0.979341	1.000000	0.000000	0.934613	0.065386
bedep	0.884791	0.798578	0.201421	0.931731	0.068268
blackhole	0.935128	0.965706	0.034293	0.933503	0.066496
ccleaner	0.933304	1.000000	0.000000	0.933128	0.066871
chinad	0.936938	0.937066	0.062933	0.936660	0.063339
chir	0.934536	1.000000	0.000000	0.934065	0.065934
conficker	0.750347	0.663866	0.336133	0.937847	0.062152
corebot	0.913256	0.904933	0.095066	0.931400	0.068599
cryptolocker	0.875679	0.849400	0.150599	0.932984	0.067015
diamondfox	0.901570	0.016913	0.983086	0.931778	0.068221
dircrypt	0.923066	0.769364	0.230635	0.936011	0.063988
dmsniff	0.933945	0.898550	0.101449	0.934122	0.065877
dyre	0.979580	1.000000	0.000000	0.935128	0.064871
ebury	0.941105	0.995497	0.004502	0.933152	0.066847
ekforward	0.925790	0.839999	0.160000	0.932891	0.067108
emotet	0.940485	0.941833	0.058166	0.937549	0.062450
feodo	0.935462	0.973821	0.026178	0.934919	0.065080
fobber	0.917871	0.794897	0.205102	0.935820	0.064179
gameover	0.762373	0.683566	0.316433	0.937213	0.062786
goznym	0.923373	0.492447	0.507552	0.933796	0.066203
gspy	0.934634	1.000000	0.000000	0.934403	0.065596
hesperbot	0.934326	0.666666	0.333333	0.937762	0.062237

infy	0.585015	0.088311	0.911688	0.933226	0.066773
locky	0.835039	0.790433	0.209566	0.935052	0.064947
madmax	0.932867	0.802721	0.197278	0.934275	0.065724
makloader	0.934824	0.988258	0.011741	0.932864	0.067135
mirai	0.932642	0.802867	0.197132	0.935281	0.064718
modpack	0.929138	0.095238	0.904761	0.935446	0.064553
monerominer	0.570430	0.404633	0.595366	0.931354	0.068645
murofet	0.940789	0.942566	0.057433	0.936859	0.063140
murofetweekly	0.309679	0.026266	0.973733	0.933524	0.066475
mydoom	0.915931	0.804001	0.195998	0.933797	0.066202
necurs	0.823803	0.771900	0.228099	0.939171	0.060828
nymaim	0.752105	0.665633	0.334366	0.939911	0.060088
oderoor	0.848168	0.756797	0.243202	0.938801	0.061198
omexo	0.932711	1.000000	0.000000	0.932617	0.067382
padcrypt	0.855725	0.819266	0.180733	0.935392	0.064607
pandabanker	0.972125	0.991720	0.008279	0.935958	0.064041
proslikefan	0.754663	0.671100	0.328899	0.937063	0.062936
pushdo	0.316682	0.036400	0.963600	0.937601	0.062398
pushdotid	0.848123	0.649608	0.350391	0.934917	0.065082
pykspa	0.823297	0.772633	0.227366	0.935393	0.064606
pykspa2	0.922039	0.811936	0.188063	0.933756	0.066243
pykspa2s	0.856406	0.744753	0.255246	0.937500	0.062500
qadars	0.924719	0.918300	0.081699	0.938747	0.061252
qakbot	0.883042	0.859033	0.140966	0.935488	0.064511
ramdo	0.832394	0.598599	0.401400	0.934798	0.065201
ramnit	0.799970	0.707301	0.292698	0.933537	0.066462
ranbyus	0.903119	0.888499	0.111500	0.935149	0.064850
rovnix	0.937974	0.971563	0.028436	0.933299	0.066700
shifu	0.937671	0.923175	0.076824	0.940137	0.059862
simda	0.627137	0.369513	0.630486	0.935420	0.064579
sisron	0.892130	0.763384	0.236615	0.934798	0.065201
sphinx	0.826609	0.774666	0.225333	0.939129	0.060870
sutra	0.933555	0.968792	0.031207	0.931651	0.068348
szribi	0.755660	0.292618	0.707381	0.935915	0.064084
tempedreve	0.933439	0.699507	0.300492	0.936918	0.063081
tempedrevetdd	0.921205	0.716444	0.283555	0.937852	0.062147
tinba	0.894641	0.875633	0.124366	0.937062	0.062937
tinynuke	0.493048	0.288433	0.711566	0.937767	0.062232
tofsee	0.892846	0.695893	0.304106	0.938973	0.061026
torpig	0.761744	0.595269	0.404730	0.936345	0.063654
ud2	0.932909	1.000000	0.000000	0.931064	0.068935
ud3	0.933304	0.508474	0.491525	0.935123	0.064876
ud4	0.938454	0.898550	0.101449	0.938657	0.061342
urlzone	0.749070	0.663166	0.336833	0.935264	0.064735

vawtrak	0.825211	0.270841	0.729158	0.935270	0.064729
vidro	0.782982	0.711266	0.288733	0.939795	0.060204
vidrotid	0.933995	0.783919	0.216080	0.936159	0.063840
virut	0.309470	0.024666	0.975333	0.943117	0.056882
wd	0.979533	1.000000	0.000000	0.934334	0.065665
xshellghost	0.936405	0.583333	0.416666	0.937339	0.062660
xxhex	0.735353	0.090020	0.909979	0.939350	0.060649
Support Vector Classifier					
Family	**Accuracy**	**TPR**	**FNR**	**TNR**	**FPR**
bamital	0.960940	0.999766	0.000233	0.876876	0.123123
bedep	0.884271	0.897545	0.102454	0.877044	0.122955
blackhole	0.877457	0.891632	0.108367	0.876704	0.123295
ccleaner	0.876448	0.944444	0.055555	0.876269	0.123730
chinad	0.959055	0.997933	0.002066	0.874337	0.125662
chir	0.872825	1.000000	0.000000	0.871910	0.128089
conficker	0.755298	0.695866	0.304133	0.884151	0.115848
corebot	0.955256	0.993866	0.006133	0.871084	0.128915
cryptolocker	0.886512	0.889966	0.110033	0.878979	0.121020
diamondfox	0.858638	0.437632	0.562367	0.873014	0.126985
dircrypt	0.869929	0.789382	0.210617	0.876713	0.123286
dmsniff	0.878008	0.768115	0.231884	0.878558	0.121441
dyre	0.959822	1.000000	0.000000	0.872360	0.127639
ebury	0.893568	0.998499	0.001500	0.878227	0.121772
ekforward	0.840231	0.413333	0.586666	0.875570	0.124429
emotet	0.845742	0.829366	0.170633	0.881417	0.118582
feodo	0.874945	0.947643	0.052356	0.873915	0.126084
fobber	0.872953	0.850425	0.149574	0.876241	0.123758
gameover	0.961306	0.999099	0.000900	0.877458	0.122541
goznym	0.865867	0.495468	0.504531	0.874826	0.125173
gspy	0.882620	1.000000	0.000000	0.882205	0.117794
hesperbot	0.869082	0.740112	0.259887	0.870738	0.129261
infy	0.576182	0.150636	0.849363	0.874510	0.125489
locky	0.805578	0.771566	0.228433	0.881838	0.118161
madmax	0.877675	0.931972	0.068027	0.877088	0.122911
makloader	0.872766	0.866927	0.133072	0.872980	0.127019
mirai	0.878642	0.903225	0.096774	0.878142	0.121857
modpack	0.866857	0.771428	0.228571	0.867579	0.132420
murofet	0.958429	0.998399	0.001600	0.871417	0.128582
monerominer	0.919238	0.936133	0.063866	0.881897	0.118102
murofetweekly	0.959591	0.999900	0.000099	0.870863	0.129136
mydoom	0.856964	0.747157	0.252842	0.874491	0.125508
necurs	0.815642	0.786333	0.213666	0.880788	0.119211
nymaim	0.704425	0.622566	0.377433	0.882212	0.117787
oderoor	0.830351	0.785554	0.214445	0.874786	0.125213

omexo	0.879128	1.000000	0.000000	0.878960	0.121039
padcrypt	0.594182	0.464299	0.535700	0.877995	0.122004
pandabanker	0.954889	0.999053	0.000946	0.873371	0.126628
proslikefan	0.702222	0.618666	0.381333	0.884604	0.115395
pushdo	0.276813	0.000000	1.000000	0.890045	0.109954
pushdotid	0.829716	0.717119	0.282880	0.878944	0.121055
pykspa	0.796391	0.757533	0.242466	0.882365	0.117634
pykspa2	0.844813	0.562803	0.437196	0.874824	0.125175
pykspa2s	0.749566	0.567828	0.432171	0.881563	0.118436
qadars	0.895883	0.906299	0.093700	0.873124	0.126875
qakbot	0.885603	0.888199	0.111800	0.879933	0.120066
ramdo	0.718659	0.357892	0.642107	0.876679	0.123320
ramnit	0.805014	0.754424	0.245575	0.877933	0.122066
ranbyus	0.889547	0.894366	0.105633	0.878989	0.121010
rovnix	0.895058	0.999473	0.000526	0.880524	0.119475
shifu	0.851759	0.685407	0.314592	0.880055	0.119944
simda	0.401302	0.000000	1.000000	0.881519	0.118480
sisron	0.889388	0.933024	0.066975	0.874926	0.125073
sphinx	0.828707	0.804966	0.195033	0.880135	0.119864
sutra	0.870729	0.941655	0.058344	0.866896	0.133103
szribi	0.721223	0.308665	0.691334	0.881825	0.118174
tempedreve	0.872076	0.689655	0.310344	0.874789	0.125210
tempedrevetdd	0.863530	0.701333	0.298666	0.876716	0.123283
tinba	0.864163	0.853133	0.146866	0.888781	0.111218
tinynuke	0.960664	0.998900	0.001099	0.877562	0.122437
tofsee	0.732731	0.097252	0.902747	0.881561	0.118438
torpig	0.837517	0.794131	0.205868	0.883020	0.116979
ud2	0.878248	1.000000	0.000000	0.874900	0.125099
ud3	0.874485	1.000000	0.000000	0.873947	0.126052
ud4	0.877421	0.768115	0.231884	0.877975	0.122024
urlzone	0.936611	0.964999	0.035000	0.875081	0.124918
vawtrak	0.778630	0.267876	0.732123	0.880029	0.119970
vidro	0.805169	0.769333	0.230666	0.883527	0.116472
vidrotid	0.872848	0.768844	0.231155	0.874347	0.125652
virut	0.338147	0.090333	0.909666	0.889498	0.110501
wd	0.959641	0.999900	0.000099	0.870730	0.129269
xshellghost	0.876767	0.750000	0.250000	0.877102	0.122897
xxhex	0.902102	0.979995	0.020004	0.877479	0.122520

Table B.4: LOGO results for individual malware families - all features except ngrams features

B.5 Only ngrams features

Gaussian Naive Bayes					
Family	**Accuracy**	**TPR**	**FNR**	**TNR**	**FPR**
bamital	0.957520	1.000000	0.000000	0.865545	0.134454
bedep	0.890606	0.937508	0.062491	0.865070	0.134929
blackhole	0.873234	0.945130	0.054869	0.869413	0.130586
ccleaner	0.868357	1.000000	0.000000	0.868011	0.131988
chinad	0.952201	0.993233	0.006766	0.862787	0.137212
chir	0.864814	1.000000	0.000000	0.863841	0.136158
conficker	0.900974	0.917533	0.082466	0.865071	0.134928
corebot	0.945316	0.980199	0.019800	0.869268	0.130731
cryptolocker	0.898761	0.914599	0.085400	0.864224	0.135775
diamondfox	0.875532	0.989429	0.010570	0.871643	0.128356
dircrypt	0.873242	0.940818	0.059181	0.867551	0.132448
dmsniff	0.873672	0.855072	0.144927	0.873765	0.126234
dyre	0.957926	1.000000	0.000000	0.866337	0.133662
ebury	0.884252	0.993496	0.006503	0.868280	0.131719
ekforward	0.881073	0.998222	0.001777	0.871376	0.128623
emotet	0.903360	0.920699	0.079300	0.865587	0.134412
feodo	0.871071	0.947643	0.052356	0.869987	0.130012
fobber	0.875756	0.935967	0.064032	0.866968	0.133031
gameover	0.957033	0.996766	0.003233	0.868880	0.131119
goznym	0.875428	0.924471	0.075528	0.874241	0.125758
gspy	0.871029	1.000000	0.000000	0.870573	0.129426
hesperbot	0.868008	0.898305	0.101694	0.867619	0.132380
infy	0.921619	0.997204	0.002795	0.868631	0.131368
locky	0.899008	0.910533	0.089466	0.873168	0.126831
madmax	0.869957	0.965986	0.034013	0.868918	0.131081
makloader	0.871242	0.933463	0.066536	0.868959	0.131040
mirai	0.872714	0.921146	0.078853	0.871729	0.128270
modpack	0.867500	0.933333	0.066666	0.867002	0.132997
monerominer	0.958749	0.999800	0.000199	0.869385	0.130614
murofet	0.919789	0.939400	0.060599	0.876445	0.123554
murofetweekly	0.955855	0.996800	0.003199	0.865727	0.134272
mydoom	0.852707	0.760800	0.239199	0.867378	0.132621
necurs	0.900728	0.915766	0.084233	0.867303	0.132696
nymaim	0.901102	0.917433	0.082566	0.865633	0.134366
oderoor	0.897191	0.923140	0.076859	0.871452	0.128547
omexo	0.865130	1.000000	0.000000	0.864943	0.135056
padcrypt	0.793386	0.756533	0.243466	0.873916	0.126083
pandabanker	0.954301	0.999645	0.000354	0.870606	0.129393
proslikefan	0.903392	0.918166	0.081833	0.871143	0.128856
pushdo	0.618414	0.502800	0.497199	0.874538	0.125461
pushdotid	0.865415	0.856142	0.143857	0.869470	0.130529

pykspa	0.896783	0.910066	0.089933	0.867394	0.132605
pykspa2	0.860899	0.782095	0.217904	0.869285	0.130714
pykspa2s	0.827045	0.764735	0.235264	0.872301	0.127698
qadars	0.942373	0.975433	0.024566	0.870138	0.129861
qakbot	0.914048	0.935433	0.064566	0.867336	0.132663
ramdo	0.883523	0.916652	0.083347	0.869012	0.130987
ramnit	0.897014	0.916220	0.083779	0.869333	0.130666
ranbyus	0.903966	0.918433	0.081566	0.872270	0.127729
rovnix	0.883927	0.992627	0.007372	0.868797	0.131202
shifu	0.868355	0.907296	0.092703	0.861731	0.138268
simda	0.714573	0.580465	0.419534	0.875054	0.124945
sisron	0.902495	1.000000	0.000000	0.870181	0.129818
sphinx	0.901890	0.917566	0.082433	0.867932	0.132067
sutra	0.865859	0.952510	0.047489	0.861176	0.138823
szribi	0.873677	0.889371	0.110628	0.867568	0.132431
tempedreve	0.867167	0.857142	0.142857	0.867316	0.132683
tempedrevetdd	0.869143	0.876444	0.123555	0.868550	0.131449
tinba	0.913954	0.935166	0.064833	0.866612	0.133387
tinynuke	0.958998	1.000000	0.000000	0.869883	0.130116
tofsee	0.870701	0.882062	0.117937	0.868040	0.131959
torpig	0.877250	0.884719	0.115280	0.869416	0.130583
ud2	0.876694	1.000000	0.000000	0.873303	0.126696
ud3	0.871377	1.000000	0.000000	0.870827	0.129172
ud4	0.868138	0.855072	0.144927	0.868204	0.131795
urlzone	0.943614	0.981199	0.018800	0.862148	0.137851
vawtrak	0.824413	0.608002	0.391997	0.867377	0.132622
vidro	0.904162	0.922766	0.077233	0.863483	0.136516
vidrotid	0.866061	0.919597	0.080402	0.865289	0.134710
virut	0.850174	0.843066	0.156933	0.865989	0.134010
wd	0.958654	0.999966	0.000033	0.867417	0.132582
xshellghost	0.863213	0.805555	0.194444	0.863365	0.136634
xxhex	0.897952	1.000000	0.000000	0.865694	0.134305
Gradient Boosting Classifier					
Family	**Accuracy**	**TPR**	**FNR**	**TNR**	**FPR**
bamital	0.986660	1.000000	0.000000	0.957780	0.042219
bedep	0.973242	0.994099	0.005900	0.961886	0.038113
blackhole	0.963860	0.997256	0.002743	0.962085	0.037914
ccleaner	0.959545	1.000000	0.000000	0.959438	0.040561
chinad	0.987227	0.999500	0.000499	0.960485	0.039514
chir	0.959581	1.000000	0.000000	0.959290	0.040709
conficker	0.933777	0.921066	0.078933	0.961335	0.038664
corebot	0.986929	0.999199	0.000800	0.960177	0.039822
cryptolocker	0.979432	0.988600	0.011399	0.959441	0.040558
diamondfox	0.966073	0.966173	0.033826	0.966069	0.033930

dircrypt	0.959369	0.986074	0.013925	0.957120	0.042879
dmsniff	0.959384	0.971014	0.028985	0.959325	0.040674
dyre	0.986729	1.000000	0.000000	0.957840	0.042159
ebury	0.963055	0.992496	0.007503	0.958750	0.041249
ekforward	0.962827	0.994666	0.005333	0.960191	0.039808
emotet	0.983550	0.994500	0.005499	0.959697	0.040302
feodo	0.963017	1.000000	0.000000	0.962493	0.037506
fobber	0.963937	0.989494	0.010505	0.960207	0.039792
gameover	0.987707	1.000000	0.000000	0.960434	0.039565
goznym	0.961401	0.927492	0.072507	0.962221	0.037778
gspy	0.959283	1.000000	0.000000	0.959140	0.040859
hesperbot	0.960825	0.977401	0.022598	0.960612	0.039387
infy	0.974911	0.992235	0.007764	0.962766	0.037233
locky	0.970170	0.972433	0.027566	0.965097	0.034902
madmax	0.959662	0.979591	0.020408	0.959446	0.040553
makloader	0.958235	1.000000	0.000000	0.956702	0.043297
mirai	0.961642	0.992831	0.007168	0.961008	0.038991
modpack	0.960100	0.885714	0.114285	0.960662	0.039337
monerominer	0.987483	0.999766	0.000233	0.960743	0.039256
murofet	0.985289	0.995433	0.004566	0.962867	0.037132
murofetweekly	0.986912	1.000000	0.000000	0.958104	0.041895
mydoom	0.954241	0.921327	0.078672	0.959494	0.040505
necurs	0.976550	0.984033	0.015966	0.959917	0.040082
nymaim	0.942323	0.933833	0.066166	0.960761	0.039238
oderoor	0.965408	0.970047	0.029952	0.960806	0.039193
omexo	0.959320	1.000000	0.000000	0.959264	0.040735
padcrypt	0.964737	0.965999	0.034000	0.961978	0.038021
pandabanker	0.986267	0.999487	0.000512	0.961865	0.038134
proslikefan	0.952404	0.947633	0.052366	0.962820	0.037179
pushdo	0.812640	0.744533	0.255466	0.963520	0.036479
pushdotid	0.955983	0.945157	0.054842	0.960717	0.039282
pykspa	0.954245	0.951133	0.048866	0.961132	0.038867
pykspa2	0.949005	0.843858	0.156141	0.960194	0.039805
pykspa2s	0.910523	0.839843	0.160156	0.961858	0.038141
qadars	0.983535	0.992600	0.007399	0.963729	0.036270
qakbot	0.981684	0.991199	0.008800	0.960899	0.039100
ramdo	0.971667	0.993832	0.006167	0.961959	0.038040
ramnit	0.974029	0.982101	0.017898	0.962396	0.037603
ranbyus	0.983681	0.993966	0.006033	0.961148	0.038851
rovnix	0.966349	1.000000	0.000000	0.961665	0.038334
shifu	0.954267	0.936480	0.063519	0.957293	0.042706
simda	0.800423	0.664541	0.335458	0.963024	0.036975
sisron	0.970660	1.000000	0.000000	0.960937	0.039062
sphinx	0.983534	0.993866	0.006133	0.961152	0.038847

sutra	0.959159	0.994572	0.005427	0.957245	0.042754
szribi	0.954136	0.940343	0.059656	0.959506	0.040493
tempedreve	0.957984	0.926108	0.073891	0.958458	0.041541
tempedrevetdd	0.959099	0.936000	0.063999	0.960977	0.039022
tinba	0.979904	0.989966	0.010033	0.957446	0.042553
tinynuke	0.987923	1.000000	0.000000	0.961674	0.038325
tofsee	0.941590	0.869095	0.130904	0.958568	0.041431
torpig	0.943578	0.925712	0.074287	0.962316	0.037683
ud2	0.962429	1.000000	0.000000	0.961396	0.038603
ud3	0.962135	1.000000	0.000000	0.961973	0.038026
ud4	0.959798	0.971014	0.028985	0.959741	0.040258
urlzone	0.984215	0.996199	0.003800	0.958240	0.041759
vawtrak	0.930281	0.775842	0.224157	0.960941	0.039058
vidro	0.966903	0.969033	0.030966	0.962244	0.037755
vidrotid	0.957139	0.979899	0.020100	0.956811	0.043188
virut	0.863995	0.820633	0.179366	0.960471	0.039528
wd	0.987243	1.000000	0.000000	0.959069	0.040930
xshellghost	0.959264	0.944444	0.055555	0.959303	0.040696
xxhex	0.969314	1.000000	0.000000	0.959614	0.040385
Logistic Regression					
Family	**Accuracy**	**TPR**	**FNR**	**TNR**	**FPR**
bamital	0.966914	1.000000	0.000000	0.895280	0.104719
bedep	0.927386	0.991015	0.008984	0.892742	0.107257
blackhole	0.904527	0.991769	0.008230	0.899890	0.100109
ccleaner	0.897878	1.000000	0.000000	0.897610	0.102389
chinad	0.967098	0.999399	0.000600	0.896709	0.103290
chir	0.898231	1.000000	0.000000	0.897499	0.102500
conficker	0.886009	0.879733	0.120266	0.899616	0.100383
corebot	0.966774	0.997999	0.002000	0.898699	0.101300
cryptolocker	0.955116	0.982833	0.017166	0.894679	0.105320
diamondfox	0.903525	0.959830	0.040169	0.901602	0.098397
dircrypt	0.899810	0.976501	0.023498	0.893351	0.106648
dmsniff	0.897737	0.927536	0.072463	0.897588	0.102411
dyre	0.967040	1.000000	0.000000	0.895290	0.104709
ebury	0.908818	0.993996	0.006003	0.896365	0.103634
ekforward	0.904383	0.980444	0.019555	0.898086	0.101913
emotet	0.960933	0.990600	0.009399	0.896303	0.103696
feodo	0.903595	1.000000	0.000000	0.902231	0.097768
fobber	0.906148	0.981990	0.018009	0.895078	0.104921
gameover	0.968774	0.999866	0.000133	0.899792	0.100207
goznym	0.899828	0.924471	0.075528	0.899232	0.100767
gspy	0.899200	1.000000	0.000000	0.898844	0.101155
hesperbot	0.898589	0.943502	0.056497	0.898012	0.101987
infy	0.928403	0.969562	0.030437	0.899550	0.100449

locky	0.938404	0.955833	0.044166	0.899327	0.100672
madmax	0.895077	0.979591	0.020408	0.894163	0.105836
makloader	0.899224	1.000000	0.000000	0.895526	0.104473
mirai	0.899428	0.985663	0.014336	0.897675	0.102324
modpack	0.895745	0.904761	0.095238	0.895677	0.104322
monerominer	0.966126	0.996333	0.003666	0.900370	0.099629
murofet	0.964106	0.992299	0.007700	0.901790	0.098209
murofetweekly	0.968530	1.000000	0.000000	0.899258	0.100741
mydoom	0.895899	0.894042	0.105957	0.896196	0.103803
necurs	0.950525	0.974800	0.025200	0.896569	0.103430
nymaim	0.894825	0.892800	0.107199	0.899225	0.100774
oderoor	0.918653	0.944054	0.055945	0.893457	0.106542
omexo	0.895603	1.000000	0.000000	0.895459	0.104540
padcrypt	0.930000	0.943633	0.056366	0.900211	0.099788
pandabanker	0.960822	0.993573	0.006426	0.900371	0.099628
proslikefan	0.900946	0.900599	0.099400	0.901702	0.098297
pushdo	0.736070	0.658233	0.341766	0.908506	0.091493
pushdotid	0.907657	0.930821	0.069178	0.897529	0.102470
pykspa	0.910374	0.913333	0.086666	0.903827	0.096172
pykspa2	0.893538	0.828591	0.171408	0.900450	0.099549
pykspa2s	0.872333	0.832613	0.167386	0.901181	0.098818
qadars	0.963800	0.991999	0.008000	0.902184	0.097815
qakbot	0.958842	0.985099	0.014900	0.901485	0.098514
ramdo	0.923178	0.979829	0.020170	0.898364	0.101635
ramnit	0.942835	0.971989	0.028010	0.900816	0.099183
ranbyus	0.963563	0.992033	0.007966	0.901190	0.098809
rovnix	0.908184	1.000000	0.000000	0.895404	0.104595
shifu	0.889193	0.874248	0.125751	0.891736	0.108263
simda	0.724560	0.574212	0.425787	0.904474	0.095525
sisron	0.923553	1.000000	0.000000	0.898218	0.101781
sphinx	0.962028	0.991166	0.008833	0.898909	0.101090
sutra	0.898142	0.994572	0.005427	0.892930	0.107069
szribi	0.899121	0.909382	0.090617	0.895127	0.104872
tempedreve	0.891207	0.896551	0.103448	0.891127	0.108872
tempedrevetdd	0.898349	0.889777	0.110222	0.899046	0.100953
tinba	0.954882	0.981233	0.018766	0.896072	0.103927
tinynuke	0.967924	1.000000	0.000000	0.898210	0.101789
tofsee	0.830101	0.560358	0.439641	0.893275	0.106724
torpig	0.908781	0.920926	0.079073	0.896042	0.103957
ud2	0.903742	1.000000	0.000000	0.901095	0.098904
ud3	0.899486	1.000000	0.000000	0.899056	0.100943
ud4	0.896937	0.927536	0.072463	0.896782	0.103217
urlzone	0.963162	0.995333	0.004666	0.893432	0.106567
vawtrak	0.864244	0.712856	0.287143	0.894299	0.105700

vidro	0.930375	0.944599	0.055400	0.899271	0.100728
vidrotid	0.893849	0.969849	0.030150	0.892753	0.107246
virut	0.787600	0.736033	0.263966	0.902328	0.097671
wd	0.968038	1.000000	0.000000	0.897452	0.102547
xshellghost	0.894644	0.916666	0.083333	0.894586	0.105413
xxhex	0.921048	1.000000	0.000000	0.896090	0.103909
Random Forests					
Family	**Accuracy**	**TPR**	**FNR**	**TNR**	**FPR**
bamital	0.997674	1.000000	0.000000	0.992638	0.007361
bedep	0.993003	0.994501	0.005498	0.992187	0.007812
blackhole	0.991276	0.998628	0.001371	0.990885	0.009114
ccleaner	0.991763	1.000000	0.000000	0.991741	0.008258
chinad	0.997052	0.999566	0.000433	0.991574	0.008425
chir	0.992854	1.000000	0.000000	0.992803	0.007196
conficker	0.939046	0.914200	0.085799	0.992917	0.007082
corebot	0.997052	0.999199	0.000800	0.992369	0.007630
cryptolocker	0.991818	0.991566	0.008433	0.992368	0.007631
diamondfox	0.992321	0.978858	0.021141	0.992780	0.007219
dircrypt	0.992969	0.990426	0.009573	0.993183	0.006816
dmsniff	0.992411	1.000000	0.000000	0.992373	0.007626
dyre	0.997601	1.000000	0.000000	0.992380	0.007619
ebury	0.991641	0.990995	0.009004	0.991735	0.008264
ekforward	0.992388	0.990222	0.009777	0.992568	0.007431
emotet	0.994859	0.996299	0.003700	0.991721	0.008278
feodo	0.993568	1.000000	0.000000	0.993477	0.006522
fobber	0.991908	0.991995	0.008004	0.991895	0.008104
gameover	0.997403	1.000000	0.000000	0.991643	0.008356
goznym	0.992080	0.927492	0.072507	0.993642	0.006357
gspy	0.993250	1.000000	0.000000	0.993226	0.006773
hesperbot	0.991119	0.988700	0.011299	0.991150	0.008849
infy	0.989247	0.983849	0.016150	0.993032	0.006967
locky	0.981996	0.977033	0.022966	0.993124	0.006875
madmax	0.992791	0.979591	0.020408	0.992934	0.007065
makloader	0.992242	1.000000	0.000000	0.991958	0.008041
mirai	0.992571	0.989247	0.010752	0.992639	0.007360
modpack	0.990990	0.866666	0.133333	0.991930	0.008069
monerominer	0.996962	0.999700	0.000299	0.991002	0.008997
murofet	0.994836	0.996399	0.003600	0.991379	0.008620
murofetweekly	0.997341	0.999933	0.000066	0.991635	0.008364
mydoom	0.983161	0.938608	0.061391	0.990272	0.009727
necurs	0.989355	0.988333	0.011666	0.991627	0.008372
nymaim	0.950311	0.930833	0.069166	0.992615	0.007384
oderoor	0.994457	0.996638	0.003361	0.992294	0.007705
omexo	0.993147	1.000000	0.000000	0.993137	0.006862

padcrypt	0.981385	0.976133	0.023866	0.992861	0.007138
pandabanker	0.996854	0.999172	0.000827	0.992576	0.007423
proslikefan	0.972521	0.962633	0.037366	0.994106	0.005893
pushdo	0.857953	0.796833	0.203166	0.993354	0.006645
pushdotid	0.981997	0.955992	0.044007	0.993367	0.006632
pykspa	0.976216	0.969233	0.030766	0.991666	0.008333
pykspa2	0.990989	0.963913	0.036086	0.993870	0.006129
pykspa2s	0.971737	0.943568	0.056431	0.992196	0.007803
qadars	0.992522	0.992399	0.007600	0.992789	0.007210
qakbot	0.993277	0.993366	0.006633	0.993082	0.006917
ramdo	0.993703	0.995332	0.004667	0.992990	0.007009
ramnit	0.988985	0.986601	0.013398	0.992420	0.007579
ranbyus	0.995399	0.996600	0.003399	0.992770	0.007229
rovnix	0.993308	1.000000	0.000000	0.992377	0.007622
shifu	0.986274	0.957510	0.042489	0.991166	0.008833
simda	0.854624	0.739088	0.260911	0.992881	0.007118
sisron	0.993857	1.000000	0.000000	0.991822	0.008177
sphinx	0.994298	0.995533	0.004466	0.991623	0.008376
sutra	0.991094	0.997286	0.002713	0.990759	0.009240
szribi	0.980057	0.950349	0.049650	0.991621	0.008378
tempedreve	0.991264	0.911330	0.088669	0.992453	0.007546
tempedrevetdd	0.989774	0.953777	0.046222	0.992701	0.007298
tinba	0.991413	0.991199	0.008800	0.991891	0.008108
tinynuke	0.997671	1.000000	0.000000	0.992610	0.007389
tofsee	0.972640	0.888237	0.111762	0.992407	0.007592
torpig	0.973937	0.956093	0.043906	0.992652	0.007347
ud2	0.993008	1.000000	0.000000	0.992816	0.007183
ud3	0.992701	1.000000	0.000000	0.992670	0.007329
ud4	0.992398	1.000000	0.000000	0.992359	0.007640
urlzone	0.995324	0.996999	0.003000	0.991691	0.008308
vawtrak	0.968270	0.840681	0.159318	0.993600	0.006399
vidro	0.992360	0.992099	0.007900	0.992930	0.007069
vidrotid	0.990999	0.979899	0.020100	0.991159	0.008840
virut	0.911714	0.875166	0.124833	0.993028	0.006971
wd	0.997522	1.000000	0.000000	0.992049	0.007950
xshellghost	0.992380	0.944444	0.055555	0.992507	0.007492
xxhex	0.993338	0.998181	0.001818	0.991807	0.008192
Support Vector Classifier					
Family	**Accuracy**	**TPR**	**FNR**	**TNR**	**FPR**
bamital	0.963767	1.000000	0.000000	0.885320	0.114679
bedep	0.921902	0.992087	0.007912	0.883688	0.116311
blackhole	0.894904	0.994513	0.005486	0.889609	0.110390
ccleaner	0.889569	1.000000	0.000000	0.889278	0.110721
chinad	0.963808	0.999433	0.000566	0.886177	0.113822

chir	0.888415	1.000000	0.000000	0.887612	0.112387
conficker	0.891529	0.892533	0.107466	0.889354	0.110645
corebot	0.964397	0.998299	0.001700	0.890487	0.109512
cryptolocker	0.953745	0.985366	0.014633	0.884794	0.115205
diamondfox	0.893961	0.961945	0.038054	0.891640	0.108359
dircrypt	0.891224	0.978241	0.021758	0.883896	0.116103
dmsniff	0.891956	0.956521	0.043478	0.891632	0.108367
dyre	0.963911	1.000000	0.000000	0.885349	0.114650
ebury	0.901288	0.993996	0.006003	0.887734	0.112265
ekforward	0.895820	0.979555	0.020444	0.888888	0.111111
emotet	0.958876	0.992166	0.007833	0.886355	0.113644
feodo	0.895629	1.000000	0.000000	0.894151	0.105848
fobber	0.898630	0.984492	0.015507	0.886098	0.113901
gameover	0.965327	0.999866	0.000133	0.888699	0.111300
goznym	0.890054	0.942598	0.057401	0.888783	0.111216
gspy	0.891570	1.000000	0.000000	0.891187	0.108812
hesperbot	0.889278	0.949152	0.050847	0.888510	0.111489
infy	0.923411	0.968009	0.031990	0.892146	0.107853
locky	0.939096	0.961266	0.038733	0.889387	0.110612
madmax	0.886267	0.993197	0.006802	0.885110	0.114889
makloader	0.890912	1.000000	0.000000	0.886910	0.113089
mirai	0.891071	0.989247	0.010752	0.889075	0.110924
modpack	0.887236	0.914285	0.085714	0.887031	0.112968
murofet	0.963774	0.996199	0.003800	0.893186	0.106813
monerominer	0.962362	0.993666	0.006333	0.893170	0.106829
murofetweekly	0.965871	1.000000	0.000000	0.890747	0.109252
mydoom	0.888888	0.901773	0.098226	0.886832	0.113167
necurs	0.949720	0.977899	0.022100	0.887086	0.112913
nymaim	0.900372	0.905433	0.094566	0.889379	0.110620
oderoor	0.919062	0.952718	0.047281	0.885678	0.114321
omexo	0.885324	1.000000	0.000000	0.885165	0.114834
padcrypt	0.931944	0.950366	0.049633	0.891689	0.108310
pandabanker	0.957932	0.993336	0.006663	0.892584	0.107415
proslikefan	0.906044	0.912266	0.087733	0.892462	0.107537
pushdo	0.750470	0.683666	0.316333	0.898464	0.101535
pushdotid	0.902789	0.935489	0.064510	0.888492	0.111507
pykspa	0.914323	0.923366	0.076633	0.894313	0.105686
pykspa2	0.886530	0.838306	0.161693	0.891662	0.108337
pykspa2s	0.872713	0.845265	0.154734	0.892648	0.107351
qadars	0.961033	0.992900	0.007099	0.891405	0.108594
qakbot	0.957401	0.986999	0.013000	0.892747	0.107252
ramdo	0.919167	0.984330	0.015669	0.890625	0.109375
ramnit	0.941283	0.976084	0.023915	0.891123	0.108876
ranbyus	0.961618	0.993433	0.006566	0.891915	0.108084

rovnix	0.899562	1.000000	0.000000	0.885582	0.114417
shifu	0.883391	0.891845	0.108154	0.881953	0.118046
simda	0.737193	0.605536	0.394463	0.894740	0.105259
sisron	0.916753	1.000000	0.000000	0.889164	0.110835
sphinx	0.960363	0.992433	0.007566	0.890894	0.109105
sutra	0.889236	0.997286	0.002713	0.883396	0.116603
szribi	0.894519	0.918821	0.081178	0.885059	0.114940
tempedreve	0.881100	0.896551	0.103448	0.880870	0.119129
tempedrevetdd	0.891064	0.905777	0.094222	0.889868	0.110131
tinba	0.953938	0.984166	0.015833	0.886475	0.113524
tinynuke	0.964796	1.000000	0.000000	0.888285	0.111714
tofsee	0.831800	0.600493	0.399506	0.885972	0.114027
torpig	0.909881	0.931955	0.068044	0.886730	0.113269
ud2	0.895268	1.000000	0.000000	0.892388	0.107611
ud3	0.890165	1.000000	0.000000	0.889695	0.110304
ud4	0.887435	0.956521	0.043478	0.887084	0.112915
urlzone	0.960835	0.996366	0.003633	0.883823	0.116176
vawtrak	0.860869	0.733975	0.266024	0.886061	0.113938
vidro	0.932296	0.952300	0.047699	0.888556	0.111443
vidrotid	0.885063	0.969849	0.030150	0.883840	0.116159
virut	0.799581	0.757399	0.242600	0.893429	0.106570
wd	0.965262	1.000000	0.000000	0.888545	0.111454
xshellghost	0.885266	0.916666	0.083333	0.885183	0.114816
xxhex	0.913731	1.000000	0.000000	0.886461	0.113538

Table B.5: LOGO results for individual malware families - only ngrams features

Appendix C

Feature values of hard-to-detect malware families

C.1 Mean

features	conficker	ekforward	infy	mydoom	nymaim	padcrypt	proslikefan
domain length	8.000506	7.935169	8	10	8.494656	16	8.014012
TLD length	2.974945	2	3.666667	2.623636	2.623314	3.237354	2.615389
TLD hash	0.682881	0.473572	0.410431	0.623243	0.592182	0.545552	0.584529
is first character digit	0	0.598579	0.613043	0	0	0	0
number of digits	0	4.85968	5.03354	0	0	0	0
number of unique characters	6.950048	6.366785	6.45528	6.56	7.297197	8.271304	6.985988
vowel ratio	0.23088	0.130036	0.124689	0.203364	0.231119	0.249719	0.230883
consonant ratio	0.76912	0.257581	0.246118	0.796636	0.768881	0.750281	0.769117
hex character ratio	0.231033	1	1	0.203364	0.23091	0.686999	0.23119
digit ratio	0	0.612384	0.629193	0	0	0	0
digit to letter ratio	0	2.370278	2.4458	0	0	0	0
longest consonant sequence	2.666951	1.207815	1.17795	3.637727	2.873873	5.506018	2.67977
longest vowel sequence	1.095724	0.706927	0.688199	1.159091	1.133897	1.698442	1.106299
longest digit sequence	0	2.463588	2.555901	0	0	0	0
is md5 like	0	0	0	0	0	0	0
shannon entropy	2.702089	2.573619	2.595386	2.565509	2.762371	2.8493	2.724961
gini coefficient	0.833914	0.817652	0.820429	0.812027	0.839674	0.842223	0.838035
classification error of characters	0.774944	0.738718	0.743478	0.727818	0.779689	0.757372	0.777015
2-gram avg	3.995813	3.26597	3.21483	4.276618	4.0002	4.259616	3.999733
2-gram med	3.672643	2.723628	2.710037	3.904387	3.670477	3.986937	3.669589
2-gram std	4.009412	3.324262	3.260116	4.332198	4.022293	4.27844	4.020556
3-gram avg	2.243522	1.37741	1.332719	2.707717	2.258867	2.777769	2.252602
3-gram med	1.82022	1.011853	1.003343	2.100205	1.812403	2.288951	1.820128
3-gram std	2.256355	1.379969	1.329796	2.845758	2.290003	2.935201	2.27966
4-gram avg	0.685708	0.318793	0.313842	1.062565	0.694671	1.189373	0.691402
4-gram med	0.492578	0.166341	0.148259	0.647858	0.488218	0.725164	0.483508
4-gram std	0.732142	0.354289	0.347281	1.206975	0.755186	1.408491	0.74403
5-gram avg	0.081152	0.022526	0.026261	0.215911	0.086751	0.286458	0.090098
5-gram med	0.029193	0.003088	0.00466	0.06653	0.027855	0.034347	0.035805
5-gram std	0.097579	0.032985	0.038936	0.261777	0.108294	0.399925	0.106388

Table C.1: Mean of features of hard-to-detect families (1)

features	**pushdo**	**pushdotid**	**pykspa**	**pykspa2**	**pykspa2s**	**ramdo**	**shifu**
domain length	10.07401	10	8.999569	10.42025	10.49157	16	7
TLD length	2.323978	2.601667	3.372718	3.040222	3.003414	3	3.333333
TLD hash	0.441704	0.594525	0.758127	0.682634	0.685721	0.445644	0.687899
is first character digit	0	0	0	0	0	0	0
number of digits	0	0	0	0	0	0	0
number of unique characters	8.182422	8.568667	7.524309	8.317614	8.331426	9.431167	6.189618
vowel ratio	0.424688	0.23065	0.257181	0.367682	0.366035	0.462865	0.237237
consonant ratio	0.575312	0.76935	0.742819	0.632318	0.633965	0.537135	0.762763
hex character ratio	0.291362	0.230967	0.230814	0.255585	0.260731	0.229688	0.23405
digit ratio	0	0	0	0	0	0	0
digit to letter ratio	0	0	0	0	0	0	0
longest consonant sequence	1.818484	3.522	2.959071	3.169903	3.286145	3.537667	2.209781
longest vowel sequence	1.337809	0.931833	1.284287	1.530513	1.542771	2.9725	1.032175
longest digit sequence	0	0	0	0	0	0	0
is md5 like	0	0	0	0	0	0	0
shannon entropy	2.933335	3.027193	2.806493	2.913664	2.914695	3.066342	2.566784
gini coefficient	0.857807	0.8691	0.844199	0.851592	0.851456	0.865775	0.820459
classification error of characters	0.787605	0.806233	0.779541	0.781408	0.780551	0.79001	0.762824
2-gram avg	4.377646	4.089354	4.029325	4.154213	4.171653	4.155329	3.999295
2-gram med	4.204043	3.696417	3.698659	3.903499	3.92615	3.9158	3.719933
2-gram std	4.322867	4.160746	4.056976	4.145483	4.164694	4.165057	3.988712
3-gram avg	2.940111	2.467285	2.314029	2.547324	2.587115	2.589714	2.239481
3-gram med	2.542861	1.929182	1.848191	2.099199	2.122312	2.044059	1.812094
3-gram std	3.006977	2.58302	2.372752	2.628357	2.674587	2.772912	2.227591
4-gram avg	1.427354	0.87531	0.735835	0.938742	0.96745	0.933584	0.684202
4-gram med	0.999477	0.57064	0.501467	0.697052	0.71663	0.61786	0.51491
4-gram std	1.534131	0.982562	0.8116	1.025968	1.065584	1.156311	0.707839
5-gram avg	0.395933	0.159077	0.105004	0.173678	0.184921	0.184608	0.090337
5-gram med	0.192436	0.046316	0.03406	0.079827	0.081093	0.008151	0.035675
5-gram std	0.450371	0.195397	0.131109	0.239425	0.250835	0.2841	0.104722

Table C.2: Mean of features of hard-to-detect families (2)

features	**simda**	**szribi**	**tempedrevetdd**	**tofsee**	**torpig**	**vawtrak**	**virut**
domain length	7.130751	8	8.253996	7	8.26571	8.990741	6
TLD length	3.530169	3	3.250444	2.5	3	2.888889	3
TLD hash	0.719612	0.393414	0.645706	0.49486	0.677889	0.363414	0.393414
is first character digit	0	0	0	0	0	0	0
number of digits	0	0	0	0	0.163546	0	0
number of unique characters	6.259439	6.204417	7.301066	3.791358	7.358926	7.725556	5.40251
vowel ratio	0.437219	0.328143	0.232423	0.153968	0.261543	0.306285	0.369251
consonant ratio	0.562781	0.671857	0.767577	0.846032	0.718462	0.693715	0.630749
hex character ratio	0.246712	0.232942	0.254798	0.434921	0.292304	0.264194	0.249187
digit ratio	0	0	0	0	0.019995	0	0
digit to letter ratio	0	0	0	0	0.02282	0	0
longest consonant sequence	1	2.779728	2.802842	1.833333	4.027674	2.605185	1.854573
longest vowel sequence	1	1.491317	1.015098	0.388889	1.261132	0.948519	1.313189
longest digit sequence	0	0	0	0	0.163546	0	0
is md5 like	0	0	0	0	0	0	0
shannon entropy	2.574948	2.517056	2.809293	1.854472	2.815733	2.869123	2.381057
gini coefficient	0.821504	0.80682	0.84944	0.711313	0.849381	0.853377	0.798017
classification error of characters	0.762037	0.727775	0.79158	0.674427	0.791052	0.791626	0.744752
2-gram avg	4.309746	4.077502	4.047395	3.90265	4.196661	4.26805	4.082903
2-gram med	4.158152	3.715581	3.692302	3.711107	3.80759	4.024792	3.812697
2-gram std	4.205455	4.103492	4.08359	3.758309	4.260772	4.239277	4.031833
3-gram avg	2.847946	2.346676	2.348031	2.019052	2.519342	2.75185	2.354338
3-gram med	2.517519	1.860385	1.911838	1.847306	1.905808	2.350426	2.083137
3-gram std	2.822546	2.423099	2.399559	1.780181	2.622004	2.78457	2.290521
4-gram avg	1.280851	0.756468	0.772518	0.634786	0.779857	1.25519	0.759883
4-gram med	1.061892	0.47941	0.565337	0.518647	0.533367	0.916434	0.607425
4-gram std	1.251724	0.824542	0.828781	0.630986	0.850573	1.318435	0.744181
5-gram avg	0.320581	0.111085	0.127861	0.11462	0.114584	0.317789	0.110924
5-gram med	0.222781	0.049217	0.061583	0.071652	0.044804	0.170643	0.110924
5-gram std	0.32558	0.130324	0.147249	0.120489	0.135658	0.347494	0.103441

Table C.3: Mean of features of hard-to-detect families (3)

C.2 Median

features	**conficker**	**ekforward**	**infy**	**mydoom**	**nymaim**	**padcrypt**	**proslikefan**
domain length	8	8	8	10	8	16	8
TLD length	3	2	3	3	3	3	2
TLD hash	0.691369	0.473572	0.161907	0.691369	0.627068	0.445644	0.473572
is first character digit	0	1	1	0	0	0	0
number of digits	0	5	5	0	0	0	0
number of unique characters	7	6	6	7	7	8	7
vowel ratio	0.2	0.125	0.125	0.2	0.2	0.25	0.222222
consonant ratio	0.8	0.25	0.25	0.8	0.8	0.75	0.777778
hex character ratio	0.2	1	1	0.2	0.2	0.6875	0.222222
digit ratio	0	0.625	0.625	0	0	0	0
digit to letter ratio	0	1.666667	1.666667	0	0	0	0
longest consonant sequence	2	1	1	4	3	5	3
longest vowel sequence	1	1	1	1	1	2	1
longest digit sequence	0	2	2	0	0	0	0
is md5 like	0	0	0	0	0	0	0
shannon entropy	2.75	2.5	2.5	2.646439	2.807355	2.858459	2.75
gini coefficient	0.84375	0.8125	0.8125	0.82	0.84375	0.84375	0.84375
classification error of characters	0.8	0.75	0.75	0.7	0.8	0.75	0.777778
2-gram avg	4.042543	3.116489	3.064992	4.327997	4.04663	4.274082	4.044403
2-gram med	3.663748	2.720986	2.720159	3.908807	3.660771	3.957942	3.660771
2-gram std	4.092477	3.232688	3.129757	4.407019	4.103113	4.283019	4.098187
3-gram avg	2.247428	1.30103	1.267172	2.808717	2.267954	2.80448	2.253257
3-gram med	1.724276	0.977724	0.954243	2.044343	1.716003	2.286681	1.724276
3-gram std	2.319063	1.244823	1.214988	2.978355	2.358573	2.966147	2.334307
4-gram avg	0.571429	0.2	0.2	0.993751	0.577236	1.195687	0.583577
4-gram med	0.477121	0	0	0.69897	0.477121	0.69897	0.477121
4-gram std	0.560081	0.4	0.4	1.109921	0.619441	1.444321	0.58893
5-gram avg	0	0	0	0	0	0.166667	0
5-gram med	0	0	0	0	0	0	0
5-gram std	0	0	0	0	0	0.30103	0

Table C.4: Median of features of hard-to-detect families (1)

features	**pushdo**	**pushdotid**	**pykspa**	**pykspa2**	**pykspa2s**	**ramdo**	**shifu**
domain length	10	10	9	10	10	16	7
TLD length	2	3	3	3	3	3	4
TLD hash	0.464847	0.487298	0.795062	0.691369	0.691369	0.445644	0.795062
is first character digit	0	0	0	0	0	0	0
number of digits	0	0	0	0	0	0	0
number of unique characters	8	9	7	8	8	9	6
vowel ratio	0.416667	0.2	0.25	0.333333	0.333333	0.4375	0.285714
consonant ratio	0.583333	0.8	0.75	0.666667	0.666667	0.5625	0.714286
hex character ratio	0.272727	0.2	0.222222	0.266667	0.266667	0.25	0.285714
digit ratio	0	0	0	0	0	0	0
digit to letter ratio	0	0	0	0	0	0	0
longest consonant sequence	2	3	3	3	3	3	2
longest vowel sequence	1	1	1	1	1	3	1
longest digit sequence	0	0	0	0	0	0	0
is md5 like	0	0	0	0	0	0	0
shannon entropy	2.947703	3.121928	2.807355	2.921928	2.921928	3.07782	2.521641
gini coefficient	0.861111	0.88	0.857143	0.86	0.86	0.867188	0.816327
classification error of characters	0.8	0.8	0.8	0.8	0.8	0.8125	0.714286
2-gram avg	4.397538	4.124403	4.074559	4.153978	4.16127	4.170354	4.053258
2-gram med	4.221453	3.686636	3.699144	3.891872	3.891872	3.968483	3.73416
2-gram std	4.330831	4.231466	4.125214	4.190905	4.207453	4.183097	4.068802
3-gram avg	2.98982	2.499515	2.336059	2.569257	2.597059	2.61913	2.24005
3-gram med	2.585461	1.848189	1.763428	2.031408	2.047275	2.058805	1.70757
3-gram std	3.060023	2.66097	2.443998	2.66727	2.694956	2.800379	2.272609
4-gram avg	1.441302	0.734686	0.640978	0.79588	0.837273	0.911363	0.511883
4-gram med	1	0.477121	0.477121	0.69897	0.812913	0.69897	0.5
4-gram std	1.53473	0.830218	0.699854	0.955814	1	1.136271	0.5
5-gram avg	0.285714	0	0	0	0	0.083333	0
5-gram med	0	0	0	0	0	0	0
5-gram std	0.389076	0	0	0	0	0.276385	0

Table C.5: Median of features of hard-to-detect families (2)

features	**simda**	**szribi**	**tempedrevetdd**	**tofsee**	**torpig**	**vawtrak**	**virut**
domain length	7	8	8	7	8.5	9	6
TLD length	4	3	3	2.5	3	3	3
TLD hash	0.795062	0.393414	0.620353	0.49486	0.691369	0.393414	0.393414
is first character digit	0	0	0	0	0	0	0
number of digits	0	0	0	0	0	0	0
number of unique characters	6	6	7	4	7	8	6
vowel ratio	0.428571	0.375	0.25	0.142857	0.25	0.363636	0.333333
consonant ratio	0.571429	0.625	0.75	0.857143	0.75	0.636364	0.666667
hex character ratio	0.285714	0.25	0.25	0.428571	0.25	0.272727	0.166667
digit ratio	0	0	0	0	0	0	0
digit to letter ratio	0	0	0	0	0	0	0
longest consonant sequence	1	3	3	2	4	2	2
longest vowel sequence	1	1	1	0	1	1	1
longest digit sequence	0	0	0	0	0	0	0
is md5 like	0	0	0	0	0	0	0
shannon entropy	2.521641	2.5	2.75	1.950212	2.807355	2.913977	2.584963
gini coefficient	0.816327	0.8125	0.84375	0.734694	0.857143	0.859504	0.833333
classification error of characters	0.714286	0.75	0.777778	0.714286	0.777778	0.8	0.833333
2-gram avg	4.33964	4.091717	4.077296	3.947858	4.22875	4.346886	4.134757
2-gram med	4.205434	3.766413	3.668317	3.701827	3.854625	4.149573	3.838723
2-gram std	4.216087	4.127457	4.131914	3.797579	4.253701	4.309228	4.107387
3-gram avg	2.902202	2.349601	2.342258	1.959518	2.510417	2.904113	2.367822
3-gram med	2.563481	1.748188	1.816241	1.748188	1.851258	2.426511	2.047275
3-gram std	2.898187	2.454354	2.447388	1.806235	2.60326	2.960387	2.342061
4-gram avg	1.249198	0.6	0.623249	0.5	0.666667	1.248761	0.666667
4-gram med	1	0.30103	0.477121	0.5	0.477121	1	0.60206
4-gram std	1.20972	0.695114	0.728487	0.481599	0.748331	1.333791	0.551331
5-gram avg	0.176091	0	0	0	0	0.154902	0
5-gram med	0	0	0	0	0	0	0
5-gram std	0.150515	0	0	0	0	0.156967	0

Table C.6: Median of features of hard-to-detect families (3)

Bibliography

[1] Ethem Alpaydin. *Introduction to Machine Learning*. MIT press, 2020.

[2] Manos Antonakakis, Roberto Perdisci, Yacin Nadji, Nikolaos Vasiloglou, Saeed Abu-Nimeh, Wenke Lee, and David Dagon. From Throw-Away Traffic to Bots: Detecting the Rise of DGA-Based Malware. In *Presented as part of the 21st {USENIX} Security Symposium ({USENIX} Security 12)*, pages 491–506, 2012.

[3] Johannes Bader. Implementation of Suppobox DGA. `https://gist.github.com/baderj/477f04c2c6f11661f403`. Accessed: 2020-04-26.

[4] Johannes Bader. Mydoom DGA. `https://github.com/baderj/domain_generation_algorithms/tree/master/mydoom`. Accessed: 2020-05-13.

[5] Johannes Bader. The DGA of Ranbyus. `https://johannesbader.ch/blog/the-dga-of-ranbyus/`. Accessed: 2020-04-26.

[6] Johannes Bader. The DGA of Simda. `https://johannesbader.ch/blog/the-dga-of-simda-shiz/`. Accessed: 2020-04-26.

[7] Leo Breiman. Random Forests. *Machine learning*, 45:5–32, 2001.

[8] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

[9] Checkpoint. VolatileCedar Technical Report. `https://www.checkpoint.com/downloads/volatile-cedar-technical-report.pdf`. Accessed: 2020-04-26.

[10] Corinna Cortes and Vladimir Vapnik. Support-Vector Networks. *Machine learning*, 20:273–297, 1995.

[11] Ryan R Curtin, Andrew B Gardner, Slawomir Grzonkowski, Alexey Kleymenov, and Alejandro Mosquera. Detecting DGA domains with recurrent neural networks and side information. In *Proceedings of the 14th International Conference on Availability, Reliability and Security*, pages 1–10, 2019.

[12] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A Fast Incremental Gradient Method With Support for non-Strongly Convex Composite Objectives. In *Advances in neural information processing systems*, pages 1646–1654, 2014.

[13] Fidelis. Vawtrak DGA Round 2. `https://www.fidelissecurity.com/threatgeek/archive/vawtrak-dga-round-2/`. Accessed: 2020-05-13.

[14] Yoav Freund and Robert E Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer, 1995.

[15] Jerome H Friedman. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of statistics*, pages 1189–1232, 2001.

[16] Jason Geffner. End-To-end Analysis of a Domain Generating Algorithm Malware Family. *Black Hat USA*, 2013.

[17] GovCERT.ch. Gozi ISFB - When A Bug Really Is A Feature. `https://www.govcert.ch/blog/gozi-isfb-when-a-bug-really-is-a-feature/`. Accessed: 2020-04-26.

[18] Talos Group. Threat Spotlight: Dyre/Dyreza: An Analysis to Discover the DGA. `https://blogs.cisco.com/security/talos/threat-spotlight-dyre`. Accessed: 2020-04-26.

[19] David G Kleinbaum and Mitchel Klein. *Logistic Regression*. Springer, 2002.

[20] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. In *Proceedings of the 26th Annual Network and Distributed System Security Symposium*, NDSS 2019, February 2019.

[21] Wes McKinney. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010.

[22] Tom M Mitchell. Generative and discriminative classifiers: Naive bayes and logistic regression. *Machine learning*, pages 1–17, 2010.

[23] Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to Linear regression Analysis*, volume 821. John Wiley & Sons, 2012.

[24] Kevin P Murphy et al. Naive Bayes Classifiers. *University of British Columbia*, 18:60, 2006.

[25] The pandas development team. pandas-dev/pandas: Pandas version 1.0.1. `https://doi.org/10.5281/zenodo.3509134`, 2020.

[26] Constantinos Patsakis and Fran Casino. Exploiting Statistical and Structural Features for the Detection of Domain Generation Algorithms. *arXiv preprint arXiv:1912.05849*, 2019.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[28] Mayana Pereira, Shaun Coleman, Bin Yu, Martine DeCock, and Anderson Nascimento. Dictionary Extraction and Detection of Algorithmically Generated Domain Names in Passive DNS Traffic. In *International Symposium on Research in Attacks, Intrusions, and Defenses*, pages 295–314. Springer, 2018.

[29] Daniel Plohmann. DGArchive. *URL https://dgarchive. caad. fkie. fraunhofer. de*, 2015.

[30] Daniel Plohmann, Khaled Yakdan, Michael Klatt, Johannes Bader, and Elmar Gerhards-Padilla. A Comprehensive Measurement Study of Domain Generating Malware. In *25th {USENIX} Security Symposium ({USENIX} Security 16)*, pages 263–278, 2016.

[31] Stefano Schiavoni, Federico Maggi, Lorenzo Cavallaro, and Stefano Zanero. Phoenix: DGA-Based Botnet Tracking and Intelligence. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*, pages 192–211. Springer, 2014.

[32] Dennis Schwarz. Bedep's DGA: Trading Foreign Exchange for Malware Domains. `https://web.archive.org/web/20160311083109/http://www.arbornetworks.com/blog/asert/bedeps-dga-trading-foreign-exchange-for-malware-domains`. Accessed: 2020-04-23.

[33] Raaghavi Sivaguru, Chhaya Choudhary, Bin Yu, Vadym Tymchenko, Anderson Nascimento, and Martine De Cock. An Evaluation of DGA Classifiers. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 5058–5067, 2018.

[34] Brett Stone-Gross, Marco Cova, Lorenzo Cavallaro, Bob Gilbert, Martin Szydlowski, Richard Kemmerer, Christopher Kruegel, and Giovanni Vigna. Your Botnet is My Botnet: Analysis of a Botnet Takeover. In *Proceedings of the 16th ACM conference on Computer and communications security*, pages 635–647, 2009.

[35] Sandeep Yadav, Ashwath Kumar Krishna Reddy, AL Narasimha Reddy, and Supranamaya Ranjan. Detecting Algorithmically Generated Malicious Domain Names. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, pages 48–61, 2010.

[36] Bin Yu, Daniel L Gray, Jie Pan, Martine De Cock, and Anderson CA Nascimento. Inline DGA Detection with Deep Networks. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 683–692, 2017.