**Associative Classification**

**Of**

**Phishing URL**

By: Ivan Golovine 500813431

April 11, 2021

Ryerson University

**Introduction**

Within this report I explore different ways to classifying Phishing URL. The data was found on UCI machine learning repository, and was already in the arff format which I placed inside weka. The data includes 1354 URL websites, amongst these URL there are 548 legitimate websites, 703 phishing URL, and 104 suspicious URL. Using associative classification, I determine which algorithm fits the best when trying to find whether a website is legitimate, phishing or suspicious. There are 9 attributes being used to determine what the website represents and all of them are nominal.

**Testing Approach**

The 5 algorithms that I chose to test and compare are J-48, Decision Stump, Multilayered Perceptron, Naïve Bayes, and K-nearest neighbor. While the 2 selection algorithms I picked to compare the best result against is OneR algorithm and the JRip algorithm. In order to determine if something is a Phishing website 9 attributes are being used, which are SFH, pop up window, SSL final state, Request URL, URL of Anchor, Web traffic, URL length, Age of domain, and having an IP address. I compare the way the algorithms categorize the URL, and determine which one is the best by looking at the amount of correctly classified instances, its success rate and which categories the algorithms are the best at determining.

**5 Algorithm Results**

**J-48**

**Observation =** The J-48 algorithm was size 88 with 55 leaves, and when run using 10 fold cross-validation managed to correctly classify 1228 instances resulting in 90.76% accuracy. It was really good at finding websites which might be suspicious getting 96 out of 104. Alongside that, it was effective with legitimate websites, achieving 489 out of 548 and with phishing website where it got 643 out of 703.

=== Run information ===

=== Summary ===

Correctly Classified Instances 1228 90.7613 %

Incorrectly Classified Instances 125 9.2387 %

Kappa statistic 0.8363

Mean absolute error 0.0831

Root mean squared error 0.2213

Relative absolute error 22.199 %

Root relative squared error 51.1815 %

Total Number of Instances 1353

=== Confusion Matrix ===

a b c <-- classified as

96 2 5 | a = 0

10 489 49 | b = 1

9 50 643 | c = -1

**Decision Stump**

**Observation =** The decision stump is pretty good at deciding if the website is a phishing or legitimate website. However, due to it going directly from the starting node to the end node it eliminates the possibility of it choosing suspicious. Because of that it greatly reduces the correctly classified instances and the success rate is only 81.74%.

=== Run information ===

Classifications

SFH = 1 : -1

SFH != 1 : 1

SFH is missing : -1

=== Summary ===

Correctly Classified Instances 1106 81.7443 %

Incorrectly Classified Instances 247 18.2557 %

Kappa statistic 0.6558

Mean absolute error 0.2092

Root mean squared error 0.3236

Relative absolute error 55.9213 %

Root relative squared error 74.8342 %

Total Number of Instances 1353

=== Confusion Matrix ===

a b c <-- classified as

0 60 43 | a = 0

0 465 83 | b = 1

0 61 641 | c = -1

**Multilayer Perceptron**

**Observation =** The multilayer perceptron performs very well, correctly classifying 1201 instances and having 88.77% accuracy. Going through the hidden layer allows it moderately accurately identify all 3 possible results.

=== Run information ===

=== Summary ===

Correctly Classified Instances 1201 88.7657 %

Incorrectly Classified Instances 152 11.2343 %

Kappa statistic 0.8005

Mean absolute error 0.0856

Root mean squared error 0.2526

Relative absolute error 22.8726 %

Root relative squared error 58.4176 %

Total Number of Instances 1353

=== Confusion Matrix ===

a b c <-- classified as

87 10 6 | a = 0

11 478 59 | b = 1

13 53 636 | c = -1

**Naïve Bayes**

**Observation =** The naïve bayes algorithm performs very well when you separate the attributes and try to identify a particular result coming from that attribute. However when you use all 9 attributes together the amount of correctly identified instances is only 1138 and the accuracy is 84.11%.

=== Run information ===

Naive Bayes Classifier

Class

Attribute 0 1 -1

(0.08) (0.4) (0.52)

=========================================

SFH

1 44.0 84.0 642.0

-1 47.0 341.0 60.0

0 15.0 126.0 3.0

[total] 106.0 551.0 705.0

popUpWidnow

-1 49.0 363.0 123.0

0 41.0 187.0 414.0

1 16.0 1.0 168.0

[total] 106.0 551.0 705.0

SSLfinal\_State

1 63.0 132.0 559.0

-1 11.0 237.0 63.0

0 32.0 182.0 83.0

[total] 106.0 551.0 705.0

Request\_URL

-1 85.0 285.0 250.0

0 20.0 213.0 191.0

1 1.0 53.0 264.0

[total] 106.0 551.0 705.0

URL\_of\_Anchor

-1 59.0 336.0 218.0

0 19.0 52.0 99.0

1 28.0 163.0 388.0

[total] 106.0 551.0 705.0

web\_traffic

1 33.0 235.0 175.0

0 22.0 217.0 237.0

-1 51.0 99.0 293.0

[total] 106.0 551.0 705.0

URL\_Length

1 58.0 90.0 214.0

-1 22.0 230.0 182.0

0 26.0 231.0 309.0

[total] 106.0 551.0 705.0

age\_of\_domain

1 76.0 256.0 496.0

-1 29.0 294.0 208.0

[total] 105.0 550.0 704.0

having\_IP\_Address

0 91.0 499.0 611.0

1 14.0 51.0 93.0

[total] 105.0 550.0 704.0

=== Summary ===

Correctly Classified Instances 1138 84.1094 %

Incorrectly Classified Instances 215 15.8906 %

Kappa statistic 0.7084

Mean absolute error 0.1389

Root mean squared error 0.2777

Relative absolute error 37.1236 %

Root relative squared error 64.2273 %

Total Number of Instances 1353

=== Confusion Matrix ===

a b c <-- classified as

13 45 45 | a = 0

18 493 37 | b = 1

10 60 632 | c = -1

**K-nearest neighbor**

**Observation =** The K-nearest neighbor is good at predicting when it needs to know if its phishing or legitimate with 1178 correctly classified instances and 87.07% accuracy. It falls short when trying to identify the more obscure cases which could result in a suspicious website. The clustering also causes the other instances to occasionally fall into the wrong result.

=== Summary ===

Correctly Classified Instances 1178 87.0658 %

Incorrectly Classified Instances 175 12.9342 %

Kappa statistic 0.7607

Mean absolute error 0.1321

Root mean squared error 0.2459

Relative absolute error 35.302 %

Root relative squared error 56.8557 %

Total Number of Instances 1353

=== Confusion Matrix ===

a b c <-- classified as

26 37 40 | a = 0

3 498 47 | b = 1

3 45 654 | c = -1

**5 Algorithm Comparison**

The best performing algorithm is J-48 which managed to correctly classify 1228 instances resulting in 90.76% accuracy, and is good overall at identifying every type of result. Decision Stump performs the worst overall with 1106 correct instances at 81.74%. It does so poorly because it does not include suspicious websites. If you only consider phishing and legitimate websites it performs on par with some of the other algorithms. The multilayer perceptron is a good option, correctly classifying 1201 instances with 88.77% accuracy. However, its results are just a worse version of J-48 and it does not have significant advantages over others in anything. On the other hand, Naïve Bayes performance, it a bit underwhelming overall, achieving 1138 correct instances and the accuracy 84.11%. But unlike multilayer perceptron it can be used to better refine the testing you do, looking how a specific attributes effects the results. Finally, K-nearest neighbor is okay at identifying instances correctly with 1178 and 87.07% accuracy. However, the fact that it has a good chance to overlook the possibility of the website being suspicious, makes it less than ideal if you want to get more accurate results.

**2 Attribute Selection Algorithms**

**OneR**

**Observation =** The most important attribute it OneR is SFH, which achieves 1106 correct instances at 81.74% accuracy.

=== Run information ===

=== Classifier model (full training set) ===

SFH:

1 -> -1

-1 -> 1

0 -> 1

(1106/1353 instances correct)

=== Summary ===

Correctly Classified Instances 1106 81.7443 %

Incorrectly Classified Instances 247 18.2557 %

Kappa statistic 0.6558

Mean absolute error 0.1217

Root mean squared error 0.3489

Relative absolute error 32.5287 %

Root relative squared error 80.6764 %

Total Number of Instances 1353

=== Confusion Matrix ===

a b c <-- classified as

0 60 43 | a = 0

0 465 83 | b = 1

0 61 641 | c = -1

**JRip**

**Observation =** TheJRip algorithm performs very well on the dataset, achieving 1221 correct instances and 90.24% accuracy. By performing incremental pruning, and ending up with 17 rules it achieves good results in all possible results. The Jrip algorithms most important attributes are SFH and Request URL.

=== Run information ===

=== Summary ===

(popUpWidnow = -1) and (SSLfinal\_State = 0) and (URL\_Length = 1) => Result=1 (12.0/5.0)

=> Result=-1 (682.0/35.0)

Correctly Classified Instances 1221 90.2439 %

Incorrectly Classified Instances 132 9.7561 %

Kappa statistic 0.8282

Mean absolute error 0.1012

Root mean squared error 0.2381

Relative absolute error 27.052 %

Root relative squared error 55.0684 %

Total Number of Instances 1353

=== Confusion Matrix ===

a b c <-- classified as

100 1 2 | a = 0

13 489 46 | b = 1

11 59 632 | c = -1

**Accuracy of with or without Data mining Algorithms**

The best performing algorithms use attribute selection algorithms with that being J-48 and JRip. When you compare JRip to something like Multilayer Perceptron or K-nearest neighbor, the results are a lot more consistent throughout. With rules and pruning the distribution of achievable results increases, and the amount of mistakes the algorithm makes is reduced.

**References**

*Liang Chen, Jiaying Peng, Yang Liu, Jintang Li, Fenfang Xie, and Zibin Zheng. 2020. Phishing*

*Scams Detection in Ethereum Transaction Network. ACM Trans. Internet Technol. 21, 1, Article 10 (February 2021), 16 pages. DOI:https://doi.org/10.1145/3398071*

*UCI machine Learning REPOSITORY: Website Phishing data set. (n.d.). Retrieved April 12,*

*2021, from* [*https://archive.ics.uci.edu/ml/datasets/Website+Phishing*](https://archive.ics.uci.edu/ml/datasets/Website+Phishing)