URL Classification using BoF of URL bitstream

Keiichi Shima
Hiroshi Abe
IIJ Innovation Institute Inc.

Daisuke Miyamoto
Nara Institute of Science and Technology

Tomohiro Ishihara
Kazuya Okada
Yuji Sekiya
The University of Tokyo

1 BACKGROUND

Extracting feature vectors from information entities is one of the important processes when using Machine Learning (ML) and Deep Learning (DL) techniques and one of the most difficult parts because extracting features usually requires a deep knowledge of the target information domain.

In this document, we tried to apply a mechanical approach to generate feature vectors from information entities, especially from the URL strings. At this moment, we are not sure if this approach works well for various different cases. We expect a better ML/DL method we may find in the future to extract hidden features from the mechanically generated feature vectors.

2 METHOD

To achieve mechanically generated feature vectors from a URL string, we first expand characters that consist of the URL string into a bit stream. Then, we pick 8 bits values from the head by shifting 4 bits each time. We finally get 256 different values from the string and make Bag of Features (BoF) from them. We create two BoF sets, one from the host part, the other from the path part, and generate a 512-dimensional vector from each URL string. Figure 1 shows an example of the process.

3 PRELIMINARY EVALUATION

We evaluated the possibility of this vectorising method by using actual URL strings. We downloaded latest active phishing site URLs from PhishTank[1] and mixed normal URLs observed in a research network. The total number of URLs of each URL set is 26938 and 26000 respectively. All the URLs in the URL sets were converted using the procedure explained as before. As a preliminary experiment, we used SVM for classifying these URLs into two classes, one is a phishing class, the other is non-phishing class. We marked all the URLs in the PhishTank set as a phishing class and marked the other as a non-phishing class. Precisely speaking, this is not accurate because the access list acquired from the research network may contain URLs of phishing sites. However, in this preliminary experiment, we ignored them because the ratio of accessing phishing site is quite low usually.

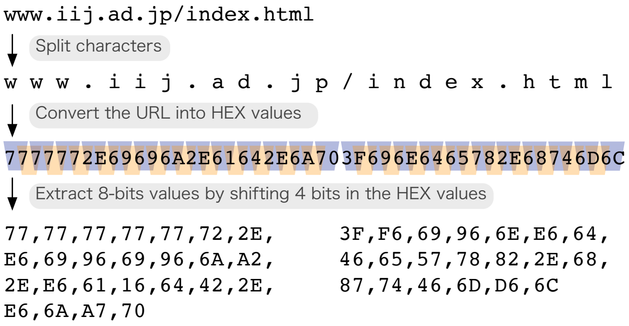
www.iij.ad.jp/index.html
Split characters
w w w . i i j . a d . j p / i n d e x . h t m l
Convert the URL into HEX values
7777772E69696A2E61642E6A703F696E6465782E68746D6C
Extract 8-bits values by shifting 4 bits in the HEX values
77,77,77,77,77,72,2E,
E6,69,96,69,96,6A,A2,
2E,E6,61,16,64,42,2E,
E6,6A,A7,70
3F,F6,69,96,6E,E6,64,
46,65,57,78,82,2E,68,
87,74,46,6D,D6,6C

Figure 1: An example process of URL vectorising

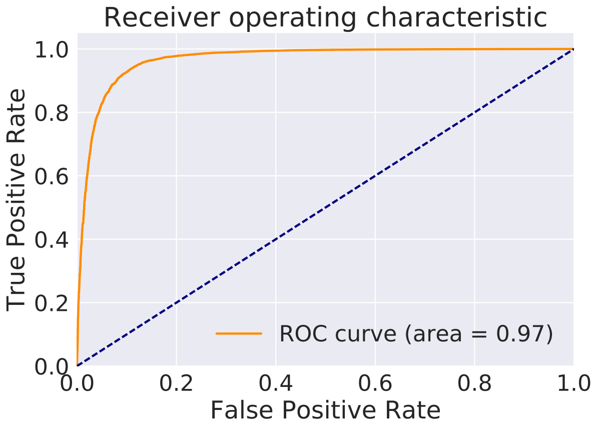

Figure 2: A preliminary result of URL classification with the proposed bitstream based URL vectorization method

The two set of URLs were mixed and the order was randomized. We used a half of them for training and evaluated the rest using the trained SVM model. Figure 2 shows the preliminary result.

4 CONSIDERATION

Although the result shows that the SVM could distinguish two classes, it doesn't mean that the method can be applied to the real data immediately. The data we have used for this preliminary experiment uses two URL data sources whose origins are completely different. If each URL set has a biased domain names or path names characteristic, the two sets may be easily distinguished even with other simpler methods. At this time, the trained SVM model generates a high ratio of false positive answers with URL data taken at different time range. We keep investigating the training model and trying different types of learning methods including DL methods to achieve better prediction results.

[1] https://www.phishtank.com/