Malicious URL detection via machine learning

A WhoisXML API User Success Story

The problem

The protection against malicious websites is an important task in cybersecurity. A common way of identifying such sites is the use of blacklists which contain a large set of URLs considered as dangerous. There are various techniques for compiling such lists, and there is obviously a need for methods to verify if a suspicious site is really dangerous. Also, as the number of malicious sites is extremely high and changes frequently in time, it is virtually impossible to create perfect blacklists. Hence, methods for identifying malicious sites is of significant practical importance and it is a subject of relevant scientific interest, too.

A machine learning approach

When a site is suspected to be malicious, Whois data can serve as a good basis for an investigation. In Ref. [1] a machine-learning approach was introduced. The idea was to use the Whois information available on a given set of URLs to design a classifier to decide if a site is malicious.

The idea is to consider *features* of URLs. These include lexicographical information such as, e.g. the length of the URL, or whether it contains the word "Login" which frequently appears in malicious ones. Another set of features come from the Whois data, as malicious URLs frequently come from certain geolocations, they exists for a very limited time only, etc.

Mathematically these features can be modeled as directions in a multidimensional space, i.e. a dimension for each feature set. In this space each URL is represented by a point. The space is then split into two parts, one in which most of the malicious ones reside, while the points of benigns are in the other part. Supervised learning consists in finding the geometrical object separating the two part as accurately as possible, using a large “training set” of URLs, already labeled as malicious/non-malicious. If this is done, one has a *classifier* at hand: an arbitrary URL can be identified as malicious or benign with some accuracy, depending on the part of the hyperspace in which its point resides.

In Ref. [1] the subsets are assumed to be separated by a *hyperplane*, the multi-dimensional version of a plane. The finding of this plane is a large-scale optimization problem.

Recently Astorino et al. [2] have published a method in the same spirit. Their methodology is aimed to be suitable for very large datasets, too. In addition, *spherical separation*, in which a multidimensional spherical surface is considered instead of the hyperplane, is utilized as it appears to have benefits over hyperplanes. They consider a limited number of features in order to prevent the explosion of the size of the sample space. They have adopted a low complexity algorithm which possibly yields a somewhat less accurate characterization, but it is computationally efficient and thus suitable for large datasets.

Results

In Ref. [2] the details of the methods are described and its successful operation is demonstrated on various real-life datasets. The biggest considered dataset consists of 11975 URLs, 5090 malicious and 6885 benign.

The research was carried out by a research group of University of Calabria, Italy. The implementation of the method needed a large amount of accurate Whois data. Queries had to be implemented in the programming languages preferred by the researchers and had run quickly to obtain the relevant information. This was realized using the services of Whois XML API.

References

[1] Ma J, Saul LK, Savage S, and Voelker GM. Beyond blacklists: learning to detect malicious web sites from suspicious URLs. KDD’09, pages 1245–1253, Paris, France, June 28–July 1 2009.

[2] A. Astorino, A. Chiarello, M. Gaudioso, and A. Piccolo. Malicious URL detection via spherical classification. *Neural Computing and Applications*, Jun 2016. doi:10.1007/s00521-016-2374-9