Abstract

Phishing websites are a form of mimicking the legitimate ones for the purpose of stealing user's confidential information such as usernames, passwords and credit card information. Recently machine learning and data mining techniques have been a promising approach for detection of phishing websites by distinguishing between phishing and legitimate ones. The detection process in this approach is preceded by extracting various features from a website dataset to train the classifier to correctly identify phishing sites. However, not all extracted features are effective in classification or equivalent in their contribution to its performance. In this paper, we investigate the effect of feature selection on the performance of classification for predicting phishing sites. We evaluate various machine learning algorithms using a number of feature subsets selected from an extracted feature set by various feature selection techniques in order to determine the most effective subset of features that results in best classification performance. Empirical results show that using our new proposed methodology for selecting features by removing redundant ones that equally contribute to the classification accuracy, the decision tree classifier achieves the best performance with an overall accuracy of 95.40%, false positive rate (FPR) of 0.046 and false negative rate (FNR) of 0.065.
On Determining the Most Effective Subset of Features for Detecting Phishing Websites

- http://www.support-vector-machines.org/
- Rami M. Mohammad, Fadi Thabtah, and Lee McCluskey. Phishing websites features,


- Joshua S. White, Jeanna N. Matthews, and John L. Stacy. A method for the automated detection phishing websites through both site characteristics and image analysis, 2012.


Index Terms

Computer Science

Information Sciences
Keywords
phishing websites detection  machine learning  classification  feature selection