Phishing websites Detection Using classification Mining Techniques

By

Asem Ismail

Supervised

Dr. Yousef Elsheikh

ABSTRACT

Most people today rely on the Internet and online activities like online shopping, online banking, online booking, and many more. Some websites are replica of the original sites so they are phished or faked website and designed to trick unskilled computer users to steal their account login credentials, credit card numbers or other personal information. This usually happens when sending e-mail messages that contain a spoofed or phishing URL to victim users. These messages are usually confusing to victim users, such as terminating a potential account or an imaginary alert to illegal transactions. This leads to phishing being a malicious technique used to steal victim information.

This leads us to seek solutions that prevent and eliminate such threats in light of the widespread use of the Internet and its applications today. Data mining techniques are among the potential solutions to this problem by predicting whether the websites being browsed real or fake. Data mining is a field of study that aims to find useful information in large databases in order to help decision-makers make the right decisions about the tasks assigned to them in their work. However, data mining involves many tasks, such as classification, association rules and clustering. In this thesis, we will focus on the task of classification, which is concerned with forecasting, assigning or predicting unseen instances to their predefined classes, however the significance of the thesis lies in, for example, predicting websites being browsed are either phishing or not.

Through this thesis, we will be able to make a comparative assessment of the most common binary classification methods used in machine learning, namely decision tree, random forest and Neural Network. Three different scenarios have been proposed to determine the features (website properties) through which we will predict whether those websites we are browsing are safe or not. In addition, this thesis seeks to employ the ensemble classification learning approach to see if its use allows us to obtain better predictive performance than using any of the individual classification learning techniques mentioned above, thereby improving the detection and protection of phishing websites. To do this, three experiments carried out, as follows: The first experiment was using all the features available in the data set adopted from Phish Tank (www.phishtank.com) for phishing websites without reducing their number and then applying the Ensemble classifier to determine the accuracy of the prediction performed. The remaining two experiments were using techniques that helped reduce the number of features and thus maintain the important and influential features, once using the chi-square and once using the PCA and then applying the ensemble classifier to each of them to determine the accuracy of the prediction in each experiment. In all three experiments, the results showed that employing the ensemble classifier was more accurate in predicting the status of the websites compared to what observed in the literature regarding the use of any of the three above-mentioned learning classifiers individually. Moreover, our use of PCA technique to determine the most effective features in predicting website status has had a

better effect than using Chi-square technique in describing the status of websites we are browsing safe or not. 1.