Abstract

Phishing is a serious web security problem that involves mimicking legitimate websites to deceive online users in order to steal their sensitive information. Phishing can be seen as a typical classification problem in data mining where the classifier is constructed from a large number of website’s features. There are high demands on identifying the best set of features that when mined the predictive accuracy of the classifiers is enhanced. This paper investigates features selection aiming to determine the effective set of features in terms of classification performance. We compare two known features selection methods in order to determine the least set of features of phishing detection using data mining. Experimental tests on a large number of features data set have been done using Information Gain and Correlation Features set methods. Further, two data mining algorithms namely PART and IREP have been trained on different sets of selected features to show the pros and cons of the feature selection process. We have been able to identify new knowledge in the forms of rules that show vital correlations among significant features.
References


Index Terms

Computer Science  Information Sciences
Keywords

Classification Accuracy, Website Security, Data mining, Feature Assessment, Phishing