Abstract

Phishing is a type of social engineering attack that aims to obtain the personal information of victims, with the method of disguising themselves as a trustworthy entity. One way to detect phishing sites is by classification of the features that characterize the site. However, several related studies actually show there are some features that are not important and not relevant. What's more, these studies use different features. This research has the aim of: creating a system that can identify the most optimum features in the classification of phishing sites. The method used is feature selection with ranking techniques using the Chi-Square formula, then followed by a classification process using the Support Vector Maching (SVM) algorithm, Multilayer Perceptron (MLP), K-Nearest Neighbors (KNN), and Random Forest. The main purpose of feature selection is to choose the best features from a collection of data features. Experiments carried out twice, namely before and after the feature selection process. The results of this study are the Random Forest algorithm produces the highest accuracy value which is 96.65% when using a dataset before feature selection and increases to 96.92% in the data after feature selection.

Keywords: feature, chi-square, website phishing, feature selection, classification