An Experimental Study for Assessing Email Classification Attributes Using Feature Selection Methods

Authors : Issa Qabaja, Fadi Thabtah

Abstract : Email phishing classification is one of the vital problems in the online security research domain that have attracted several scholars due to its impact on the users payments performed daily online. One aspect to reach a good performance by the detection algorithms in the email phishing problem is to identify the minimal set of features that significantly have an impact on raising the phishing detection rate. This paper investigate three known feature selection methods named Information Gain (IG), Chi-square and Correlation Features Set (CFS) on the email phishing problem to separate high influential features from low influential ones in phishing detection. We measure the degree of influentially by applying four data mining algorithms on a large set of features. We compare the accuracy of these algorithms on the complete features set before feature selection has been applied and after feature selection has been applied. After conducting experiments, the results show 12 common significant features have been chosen among the considered features by the feature selection methods. Further, the average detection accuracy derived by the data mining algorithms on the reduced 12-features set was very slight affected when compared with the one derived from the 47-features set.

Keywords : data mining, email classification, phishing, online security

Conference Title : ICSRD 2020 : International Conference on Scientific Research and Development

Conference Location : Chicago, United States

Conference Dates : December 12-13, 2020