GA-based feature subset selection in a spam/non-spam detection system

ABSTRACT

Spam has created a significant security problem for computer users everywhere. Spammers take an advantage of defrauds to cover parts of messages that can be used for identification of spam. For instance, a spammer does not need to consume much cost and bandwidth for sending junk mails even more than one hundred emails. On the other hand, from the feature selection perspective, one of the specific problems that decrease accuracy of spam and non-spam emails classification is high data dimensionality. Therefore, the reduction of dimensionality is related to decrease the number of irrelevant features. In this paper, a genetic algorithm (GA) is applied during feature selection in effort to decrease the number of useless features in a collection of high-dimensional email body and subject. Next, a Multi-Layer Perceptron (MLP) is employed to classify features that have been selected by the GA. Using LingSpam benchmark corpora as the dataset, the experimental results showed that a GA feature selector with the MLP classifier does not only decrease the data dimensionality but increase the spam detection rate as compared against other classifiers such as SVM and Naïve Bayes.

Keyword: Feature selection; Genetic algorithm; MLP; Spam detection