ABSTRACT

E-mail is one of the popular modes of communication due to its easy accessibility and low cost. Internet has made possible, the recent explosion in research work and human knowledge. Communication can happen to any person anywhere in the world at any time at the click of the mouse. Due to the advantages of speed, time and cost effectiveness, many people use it for commercial advertisement purposes resulting in unwanted e-mails at user inboxes. Those e-mails which are useful are known as ham, while the unwanted commercial e-mails, are known as spam. Managing e-mails is a serious problem for both individuals and organizations. With the absence of appropriate counter-measures, spam e-mails could eventually undermine the usability of e-mail. Spammers collect e-mail addresses from chat rooms, websites, customer lists, newsgroups and viruses which harvest users address books. Spam e-mail are sent out in bulk quantities every day and these spam e-mails often have very similar characteristics, allowing them to be detected using various machine learning algorithms.

Generally an e-mail is represented by a vector space model in which every e-mail is considered as a vector of terms. Since there are many different terms in the e-mail, not all classifiers can handle high dimensional data. Certain terms could be powerful discriminatory terms which are required for classification, others might not be influential and could carry redundant information, confusing the classifier. The problem of curse of dimensionality further reduces the performance of the classifier. Therefore, dimensionality reduction is important to obtain the best constructed features.

Feature selection and feature extraction are important research problem in the applications related to classification including e-mail spam classification. Feature selection techniques can be applied to produce optimal subset. A filter method or wrapper method of feature selection can be employed. The statistics computed from the data are used for evaluation in the filter method whereas the predictive performance of the classifier is used in the case of wrapper method.

The thesis investigates the use of both feature extraction and feature selection techniques in order to improve the classification performance. The first chapter deals with feature extraction based on the dimensionality reduction technique namely Latent Semantic Indexing (LSI). The proposed methods include LSI based Naive Bayes (NB) classifier and LSI based Support Vector Machine (SVM) classifier. The performance of the proposed methods has been compared with the existing algorithms using evaluation measures such as precision, recall and accuracy. The drawback of the proposed method is that the feature selection is done irrespective of the classifier being used. The performance of the classifier can be improved if wrapper method is employed where the feature selection is based on the classifier that is used.

The second chapter makes use of the evolutionary algorithms such as Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), Firefly algorithm (FF) for feature selection, where the performance of a classifier is used to evaluate the quality of a feature subset. These bio-inspired computational algorithms are employed in feature subset selection and the selected significant attributes are given to the NB classifier to classify the e-mail as ham or spam. Among these proposed methods, FF-NB algorithm showed better performance in classifying the e-mails. The proposed evolutionary computational techniques suffer from premature convergence. As the number of iteration increases, they converge to the local optima which may not be the global optima.

A hybrid technique based on Group Search Optimization (GSO) and NB classifier is proposed in third chapter. GSO inspired by animal behaviour, especially animal searching behaviour, is an optimization algorithm which is based upon the group specialty. GSO algorithm includes three types of members namely Producer, Scroungers and Rangers. The framework is mainly based on Producer (finding) - scrounger (joining) model. The simulation results indicate that the proposed GSO-NB method is more effective than the existing methods.

The fourth chapter includes a technique that employs a hybrid Firefly-Group Search Optimizer algorithm (FGSO) for feature selection. The FGSO algorithm continues the classification model of FF and GSO. Angle search and step search are used simultaneously. The FGSO is an improved FF or GSO. It adopts the angle search mechanism of GSO and uses the step search mechanism of FF when the algorithm does not move forward. The results show that the proposed FGSO-NB method produces a high quality of feature subset and also overcomes the problem of local optima, thus increasing the accuracy of the NB classifier.

A multi-classifier system where the three classifiers namely Decision Tree classifier, NB classifier and Neural Network classifier are combined by using voting-based weighted rule is proposed in fifth chapter. These three classifiers have their own advantages and disadvantages. Hence the hybridization of classifiers would help in providing overall improvements by rectifying their own disadvantages by other algorithms and retaining their

advantages. The performance of the proposed FGSO based Weight based Multiple Classifier (FGSO-WMC) for e-mail spam classification has been evaluated using the measure precision, recall and accuracy. The simulation results indicate that the proposed FGSO-WMC is better than other techniques. The benchmark datasets namely UCI repository Spambase and Ling-spam corpus are used for evaluating the proposed methods. The results show that the proposed FGSO-WMC method produce improved performance for e-mail spam classification when compared to other existing and proposed methods.