Chapter 86
Feature Selection Techniques for Email Spam Classification: A Survey

V. Sri Vinitha and D. Karthika Renuka

Abstract In this digital world, most of the communication is done only through the Internet. Email is widely used for exchanging information not only for personal communication but also has an important part in business communication because of its effectiveness, fastness, and cost-effective mode of communication. Spam email is the serious problem on the Internet; when users click on to the spam mail, it starts spreading viruses in the user system, consumes lot of network bandwidth and email storage space, and steals user's confidential data. Feature selection approach selects the best features from the dataset which removes irrelevant, redundant, and noisy data. The proposed paper offers email spam detection which incorporates various feature selection approaches like Information Gain, Correlation-Based Feature Selection, Genetic Algorithm, Ant Colony Optimization, Artificial Bee Colony, Particle Swarm Optimization, Cuckoo Search Algorithm, Harmony Search Algorithm, etc.; when classification is done after feature selection, it will enhance the performance of spam filtering.

Keywords Feature selection · Information gain · Genetic algorithm · Artificial bee colony · Ant colony optimization · Particle swarm optimization · Cuckoo search algorithm · Harmony search algorithm

Abbreviations

ABC	Artificial Bee Colony Optimization
BoW	Bag-of-Word
CGA	Compact Genetic Algorithm
EA	Evolutionary Algorithm

V. S. Vinitha (✉)
Bannari Amman Institute of Technology, Sathyamangalam, Erode, India

D. K. Renuka
PSG College of Technology, Coimbatore, Tamil Nadu, India

L. Ashok Kumar et al. (eds.), *Proceedings of International Conference on Artificial Intelligence, Smart Grid and Smart City Applications*,
https://doi.org/10.1007/978-3-030-24051-6_86

F-GSO	Firefly-Group Search Optimizer
HKSVM	Hybrid Kernel based Support Vector Machine
HSA	Harmony Search Algorithm
KNN	K-Nearest Neighbors
LR	Logistic Regression
MLP	Multi-Layer Perceptron
MLP-NN	Multi-Layer Perceptron Neural Network
NB	Naïve Bayes
PCA	Principal Component Analysis
PNN	Probabilistic Neural Network
PoS	Part-of-Speech
PSO	Particle Swarm Optimization
PU	Positive Unlabeled
RF	Random Forests
SCS	Stepsize Cuckoo Search
SMO	Sequential Minimal Optimization
SVM	Support Vector Machine
TF-IDF	Term Frequency – Inverse Document Frequency
TREC	Text Retrieval Conference
UCI	UC Irvine

86.1 Introduction

Feature selection [1] is a technique for discovering a minimal number of features f from the original F features of an email spam dataset. Some features are relevant for spam detection, but some are not with repetition and also it can be continuous, discrete, or nominal. It is recognizable to determine the finest attributes from the new email dataset that comprises the features with smallest number of dimensions which contributes by removing irrelevant data, reducing dimensionality to increase the accuracy and to improve the performance. This technique has optimal number of features which attain same or better results.

The features that are related to the spam detection and impact on spam detection are known as relevant features, while the rest cannot perform their role. If the features do not have any impact on spam detection, they are irrelevant features. If there are features with repetition, then they are redundant and even the features may be different. The real-world dataset used for spam detection may contain noise, irrelevant or ambiguous features where feature selection plays a vital role. Many algorithms, approaches, and methods are available for feature selection.

Feature selection algorithms [2] are predominantly characterized into three methods. They are:

- Filter method
- Wrapper method
- Embedded method.

86.1.1 Filter Method

Filter method is to be applied for choosing the significant features that have to be done before classification. It is used to determine the best features from the input spam email. It is independent of any classification algorithm and filters data based on the selective criteria. The input of the filter is the attributes of email dataset. Based on the scores obtained from various statistical tests, features which are significant in determining the outcome variable can be selected. If there are more number of features, then the filter method can be more suitable because of the high computational efficiency. Some of the filter methods are correlation-based methods, mutual information-based methods, information gain, chi-squared test, etc.

86.1.2 Wrapper Method

Wrapper method is suitable for dataset that contains less amount of attributes. For the given email dataset, it finds the suitable attributes and trains a model using them. Based on the dependencies among features, attributes can be added or removed from the subset. It provides better result when compared with the filter method. But it requires more computational resources than the filter method and more appropriate for small training datasets. Few of them are Sequential Forward Selection, Genetic Algorithms, Stepwise Regression, Backward Elimination Method, etc.

86.1.3 Embedded Method

Embedded method has been projected to incorporate the benefits of filter method and wrapper method. In this model, some good features will be selected from an email spam dataset by using the filter method. Then wrapper method is applied on those selected features to acquire the best feature. For feature selection, one of two methods, such as the subset selection or the feature ranking method, can be used. The set of possible features is selected based on the criterion, forms the optimal subset in the subset selection method, and ranks the features according to the criterion in the feature ranking process. Weighted Naïve Bayes, Sequential Forward Selection, Artificial Neural Networks, etc., come under the embedded method.

86.2 Overview of Email Spam Detection

In recent years [3], email has become a platform that is extensively used in the Internet for communication. It is an electronic messaging system used to transfer message from one user to another. In the email, spam is the major concern, which

Table 86.1 Detailed analysis of datasets used in spam email classification

Dataset	Dataset sample Size
PU	Total 7101 email (spam = 3020 and ham = 4081)
Custom	It varies from study to study
SpamBase	Total 4601 emails (spam = 1813 and ham = 2788)
Enron spam Corpus	Total 30,041 emails (spam = 13,496 and ham = 16,545)
SpamAssasin	Total 10,744 emails (spam = 3793 and ham = 6951)
TREC	Total 92,189 emails (spam = 52,790 spam and ham = 39,399)
CCERT	Total 34,360 emails (spam = 25,088 and ham = 9272)
LingSpam	Total 3252 emails (spam = 841 and ham = 2412)

transmits messages to bulk amount of beneficiaries. Spam email is also called as junk mail. Spammers usually collect these addresses from websites for spreading malware and sending phishing emails for stealing user confidential data. It consumes email storage space and wastes user time in opening and deleting the junk emails.

86.2.1 Dataset

There are various datasets available in the UCI Repository to detect spam mails. The dataset contains spam and ham mails. Some of them [4] are listed in the table (Table 86.1).

86.2.2 Evolutionary Algorithms

Evolutionary algorithm uses nature-inspired approach for optimization. It follows the behavior of living organisms to solve the problem and is inspired by the concepts in Darwinian Evolution and modern genetics. Evolutionary algorithm is intended for resolving a problem more quickly which will consume more time for thorough processing. There are four inclusive steps in EA, which are Selection, Mutation, Crossover, and Accepting. In EA, appropriate members will persist and increase, while irrelevant members will perish and not contribute for further generations. From the given population of individuals, natural selection is made by environmental pressure to rise the appropriateness of the population. Fitness measure is applied to the appropriate solutions which were randomly created. With this measure, the better possible solutions are applied with recombination or mutation and given to the next generation. Recombination operator is enforced on parents and results in the children. Mutation operator is enforced on one candidate to produce another new candidate. Depending on their fitness measure in the next generation, recombination and mutation replace old ones with a new candidate. This process will be iterated until the best solution is found.

86.2.3 Classification Algorithms

After selecting the finest attributes from an email spam dataset, Machine Learning algorithms are used for performing classification. Various algorithms are available to classify the non-spam and spam emails. Classification algorithms include Decision Tree, AdaBoostj48, Naïve Bayes (NB), Support Vector Machine (SVM), Random Forests (RF), Neural Networks, and Multi-Layer Perceptron (MLP). Then the performance is evaluated by using different metrics such as accuracy, sensitivity, and specificity.

86.3 Related Works

Some of the research works done on predicting the spam email by employing various Feature Selection techniques to select the best features for performing classification are deliberated in the following section:

Email is the instantaneous method for exchanging information through the Internet. Spam mails contain mischievous code to steal personal information about the user and also to infect user's system through spreading viruses. In order to reduce the consequences of spam mail, Vrinda Sharma [5] proposed a Term Frequency and Inverse Document Frequency (TF-IDF) and Information Gain for efficient feature selection. Then, the result of these two feature selection is applied on four classification algorithms, namely, Support Vector Machine, Naïve Bayes, K-Nearest Neighbors (KNN), and Random Forest. It is tested on different datasets such as DBWorld E-Mails, LingSpam, and Enron dataset using classification algorithms like Naïve Bayes, KNN, Random Forests, and SVM. Random Forests and SVM provide a better result, but SVM takes more time. But Naïve Bayes and KNN are improved in terms of accuracy and time.

Spam is a major concern on the today's Internet. To classify the spam emails, four feature selection techniques and Machine Learning algorithms are used for classification. Reshma Varghese [6] recommended Bag-of-Word (BoW)s, Bigram Bag-of-Word (BoW)s, Part-of-Speech (PoS) Tag, and Bigram PoS Tag for extracting the features. The Naïve Bayes score is used to eliminate the rare features. Enron dataset is taken as the input. Features are selected by Information Gain and form matrix using Term Frequency – Inverse Document Frequency (TF-IDF). For classification, AdaBoostJ48, Random Forest, and Popular Linear Support Vector Machine (SVM), called Sequential Minimal Optimization (SMO), are used and yield an accuracy of 0.932, 0.911, and 0.750 for Adaboost, Random Forest, and SMO. Adaboost provides good results with ensemble model.

Email is one of the fastest modes of communication used on a daily basis by millions of people. However, the number of email users has increased resulting in dramatic increase in spam mails over the past few years. P.U. Anitha [7] proposed an efficient spam classification technique using Naïve Bayes classifier and Compact

Genetic Algorithm (CGA) by using SpamBase and LingSpam datasets. It contains training and testing phases. During the training phase, best features are selected using hybrid Cuckoo Search and Genetic Algorithm. After selecting the best features, classification is done by using Naïve Bayes algorithm. Performance was compared with existing techniques like Particle Swarm Optimization (PSO). The comparison indicates that the proposed system using hybrid optimization provides better accuracy.

Email is one of the important ways of exchanging information. Spam is serious concern in today's Internet. So there is a need to filter the spam emails. Issam Dagher [7] recommended spam filtering using Kernel Principal Component Analysis. It is implemented using a Public Corpus extracted from the University of California-Irvine Machine Learning Repository. The best features are extracted using PCA. For classification using Support Vector Machine and Naïve Bayes, different training and testing sets are used. The spam mails are correctly classified for more number of trials and it takes less time comparable to PCA. Kernel PCA provides the best performance in terms of accuracy. The accuracy of the Bayes detector was high, but it takes more time for classifying large number of features.

Spam is the major problem faced by most of the email users, as it consumes large amount of email storage and steals all users' personal data. Therefore, a filter is needed to block these spam emails. In the dataset, not all the features are relevant for spam classification. Thus suitable features should be extracted for further processing. Mehdi Zekriyapanah Gashti [8] chose various datasets such as SpamBase, LingBase, and PU1 and applied the Harmony Search Algorithm (HSA) to select the best features. Selected features help to improve the accuracy of its predictions. Then, Decision Tree is used for classifying the selected features. The proposed model on SpamBase dataset provides an accuracy of about 95.25% which is better than SVM, J48, MLP, and NB. And also, the accuracy of proposed model on LingSpam and PU1 dataset provides better result than LR, NB, and SVM.

Email has established a significant role in exchanging information because of its fastest and cost-effective way of communication. It plays a vital role in both personal and business communication. The rapid growth of email has generated several issues. From past decades, spam emails start spreading tremendously. These spam emails spread malware in user system and steal personal data by sending phishing emails. So, there is a need for efficient filter to classify the spam and ham mails. Harjot Kaur [9] proposes MLP for classifying spam emails. MLP takes more execution time and degrades the performance of algorithm. So in future work, refined MLP along with N-Gram-based feature selection is used to remove noise and outliers in the dataset and for selecting the best features from the corpus.

The communication tool is attacked by intruders for sending unwanted spam emails. Several spam filtering techniques exist, but still the problem survives. Masoome Esmaeili [10] addresses this issue by implementing the Bayesian method and PCA to filter these spam emails from the user inbox. Forty spam and fifty non-spam emails are considered in the training step and extracted the features and saved them with their frequencies in a local dictionary. Then, they were classified by

using the Bayesian method and compared its result with various feature selection techniques. The ratio method was applied on the original dictionary in the preprocessing step to remove the irrelevant features. Then GA) was applied on modified dictionary and obtained 97.76% with 3400 features.

In this digital world, spam causes serious problem to the Internet users. In this paper, T. Kumaresan [11] suggested a modified Cuckoo Search called Stepsize Cuckoo Search (SCS) and Support Vector Machine for spam email classification. SCS is used which not only speeds up the convergence of the algorithm but also allows us to find the optimal features from the SpamBase dataset. Then the classification is done by using the Support Vector Machine. For the effectiveness of classification, three different kernels, such as linear, polynomial, and quadratic, are used. The proposed system is evaluated by different metrics like precision, recall, and accuracy, and the result shows that it provides better result when compared with the existing classification technique.

Due to cost-effective communication, email is widely used for personal and business communication for transferring messages. Spam has become a major problem because it causes unnecessary traffic and security threats. Several techniques have been deployed to block these spam emails. Shashi Kant Rathore [12] proposes the hybrid Bayesian algorithm and swarm intelligence for recognizing spam mails. Best features from the LingSpam dataset are selected by using swarm intelligence and the classification is done by using the Naïve Bayes algorithm. This approach takes static values of probabilities for each token. So, an automated trained filter can also be maintained by including Nature-based optimization techniques such as Artificial Bee Colony and Spider Monkey Optimization. From this, the best tokens can be classified to recognize the spam mails.

In this e-world, email stands out for communication in the Internet. Because of its popularity, it is misused by people for sending unwanted messages to large number of recipients. These emails are called as spam emails. Spam email lessens the productivity, consumes extra storage in the mailbox, takes up a lot of time for opening and deleting the mails, spreads viruses, and steals user's data through phishing emails. So, there is a need to block the spam mails from entering the user's inbox. Shradhanjali [13] suggested a novel method using Support Vector Machine and feature extraction. The proposed system obtains an accuracy of about 98% with the test datasets.

Spam email causes severe problem to the Internet community that threatens network bandwidth and productivity of the users. T. Kumaresan [14] recommended a framework using S-Cuckoo and Hybrid Kernel based Support Vector Machine (HKSVM). At first, textual features are selected from the LingSpam dataset using Term Frequency and for images, correlogram and wavelet moment are taken. The optimal features are selected using hybrid S-Cuckoo Search. After selecting the features, classification is done using HKSVM. Then, the performance is analyzed by using evaluation metrics such as precision, recall, and accuracy. Experimental result shows that the proposed HKSVM provides better result when compared to other SVM-based models (Table 86.2).

Table 86.2 Performance analysis of email spam detection using feature selection and classification techniques

Author	Feature selection used	Classification techniques used	Dataset	Evaluation metrics
Vrinda Sharma et al. [5]	TF-IDF and information gain	Naïve Bayes, SVM, KNN, and Random Forests	DBWorld E-mails, LingSpam and Enron6 datasets	Accuracy: NB – 0.8524, KNN – 0.8196, RF – 0.9852, SVM – 1
Reshma Varghese et al. [6]	Information gain	Adaboostj48, random forest, and sequential minimal optimization (SMO)	Enron spam dataset	Accuracy: Adaboost – 0.932, Random Forest – 0.911, SMO – 0.750
Mehdi Zekriyapanah Gashti [8]	Harmony search algorithm	Decision tree	SpamBase, LingSpam, and PUI datasets	Accuracy: SpamBase – 0.9525, LingSpam – 0.9980, PUI dataset – 0.9712
Harjot Kaur et al. [9]	N-gram-based feature selection	Multi-layer perceptron neural network (MLP-NN) and Support Vector Machine (SVM)	Enron dataset	Accuracy: SVM – 0.6466, MLP – 0.7809
Shradhanjali et al. [13]	Features that matched the word from dictionary are extracted and are mapped using vocab file	Support vector machine	Apache public corpus	Accuracy – 0.98
Dhanaraj Karthika Renuka et al. [15]	F-GSO algorithm	Decision tree rule, Naïve Bayes, and neural network	SpamBase dataset	Accuracy – 0.9883
Masoome Esmaeili et al. [10]	Principal component analysis	Naïve Bayes	Forty spam and fifty non-spam emails	Accuracy – 0.9687

(continued)

Table 86.2 (continued)

Author	Feature selection used	Classification techniques used	Dataset	Evaluation metrics
Issam Dagher et al. [7]	(PCA)	Support vector machine, Naïve Bayes	University of California-Irvine Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection)	Accuracy: SVM – 0.9621, NB – 9695
T. Kumaresan et al. [14]	Stepsize-Cuckoo Search (SCS)	SVM	LingSpam Dataset	Accuracy – 0.96
T. Kumaresan et al. [14]	S-Cuckoo Search	Hybrid kernel based support vector machine (HKSVM)	LingSpam dataset	Accuracy – 0.97235
Shashi Kant Rathore et al. [12]	ABC (Artificial Bee Colony Optimization), SMO (Spider Monkey Optimization)	Naïve Bayes	LingSpam dataset	Accuracy – 0.92
Mohammad Zavvar et al. [16]	Hybrid particle swarm optimization algorithms and artificial neural network	Support vector machine	SpamBase dataset	Accuracy – 0.8742
D.Karthika Renuka et al. [17]	Ant Colony optimization	Naïve Bayes	SpamBase dataset	Accuracy – 0.84
Masurah Mohamad et al. [18]	Term frequency inverse document Frequency (TF-IDF)	Rough set theory	169 emails comprising of texts and images	Accuracy – 0.848
S. Kumar et al. [19]	Particle swarm optimization (PSO)	Probabilistic neural Network (PNN)	90 emails	Accuracy – 0.90
Sorayya Mirzapour Kalaibar et al. [20]	Genetic algorithm	Bayesian network and KNN classifiers	SpamBase dataset	Accuracy: Bayesian network – 0.935, KNN – 0.886

86.4 Conclusion

This paper provides an overview of various feature selection techniques that can be used for email spam detection. In the input email dataset, all the attributes are not relevant for detecting the spam mails. Some features are relevant, but some are irrelevant. So, there is a need to eliminate the irrelevant features that lessen the

execution time and provide better accuracy. Feature selection is used to select the best attributes from the spam email dataset. It reduces the dimension of the input and after that, it uses classification techniques to classify the spam emails, which helps in improving the performance of spam detection.

Acknowledgments Our sincere thanks to the University Grants Commission (UGC), Hyderabad, for granting the funds to carry out this work.

References

1. Xue B, Zhang M, Browne WN, Yao X (2016) A survey on evolutionary computation approaches to feature selection. IEEE Trans Evol Comput 20(4):606–626
2. Jain D, Singh V (2018) Feature selection and classification systems for chronic disease prediction: a review. Egypt Inform J Elsevier 19(3):179–189
3. Bhuiyan H, Ashiquzzaman A, Juthi TI, Biswas S, Ara J (2018) A survey of existing E-mail spam filtering methods considering machine learning techniques. Global J Comp Sci Technol C Softw Data Eng 18(2):21–29
4. Mujtaba G, Shuib L, Raj RG, Majeed N, Al-Garadi MA (2017) Email classification research trends: review and open issues, IEEE Access, pp 9044–9064
5. Sharma V, Poriye M, Kumar V (2017) Various classifiers with optimal feature selection for-email spam filtering. Int J Comput Sci Commun 8(2):18–22
6. Varghese R, Dhanya KA (2017) Efficient feature set for spam email filtering. IEEE 7th international advance computing conference, pp 732–737
7. Dagher I, Antoun R (2017) Ham – Spam Filtering using Kernel PCA. Int J Comput Commun 11:38–44
8. Mehdi Zekriyapanah Gashti (2017) FHSA. Eng Technol Appl Sci Res 7(3):1713–1718
9. Kaur H, Prince Verma E (2017) E-mail spam detection using refined MLP with feature selection. Int J Modern Educ Comput Sci 9:42–52
10. Esmaeili M, Arjomandzadeh A, Shams R, Zahedi M (2017) An anti-spam system using naive Bayes method and feature selection methods. Int J Comput Appl 165(4):1–5
11. Kumaresan T, Palanisamy C (2017) E-mail spam classification using S-cuckoo search and support vector machine. Int J Bio-Inspired Comput 9(3):142–156
12. Rathore SK, Yada S (2017) A hybrid Bayesian approach with ABC to recognition of email SPAM. Int J Comput Sci Mob Comput 6(5):459–466
13. Shradhanjali, Verma T (2017) E-mail spam detection and classification using SVM and feature extraction. Int J Adv Res Ideas Innov Technol 3(3)
14. Kumaresan T, Saravanakumar S, Balamurugan R (2017) Visual and textual features based email spam classification using S-cuckoo search and hybrid kernel support vector machine. Clust Comput 22:33–46. Springer Publication
15. Renuka DK, Visalakshi P (2017) Weighted-based multiple classifier and F-GSO algorithm for email spam classification. Int J Business Intelligence Data Mining 12(3):274–298
16. Zavvar M, Rezaei M, Garavand S (2016) Email spam detection using combination of particle swarm optimization and artificial neural network and support vector machine. Int J Modern Education Computer Science (IJMECS) 8(7):68–74
17. Karthika Renuka D, Visalakshi P, Sankar T (2015) Improving E-mail spam classification using ant Colony optimization algorithm. Int J Comput Appl 22–26

18. Mohamad M, Selamat A (2015) An evaluation on the efficiency of hybrid feature selection in spam email classification. IEEE international conference on computer, communication, and control. Technology 227–231
19. Kumar S, Arumugam S (2015) A probabilistic neural network based classification of spam mails using particle swarm optimization feature selection. Middle-East J Sci Res 23(5):874–879
20. Kalaibar SM, Razavi SN (2014) Spam filtering by using genetic based feature selection. Int J Comput Appl Technol Res 3(12):839–843