Comparative Study of Email Spam Filtration Using Machine Leaning Algorithms

Rahul Gupta[1], Akash Raghuwanshi[2]

[1]Department of Information Technology, S R Institute of Management and Technology, Lucknow, UP, India
Email: *contacttorahulgupta[at]gmail.com*

[2]AKTU Lucknow
Email: *akash.raghuvanshi[at]gmail.com*

Abstract: ***In modern life, social network is an online platform extensively used as communication tool in order to build social relation, and email is one of them. Spam mails have become a serious matter of concern on internet in recent times. Hackers get the chance to abuse emails and steal its information for an illegal purpose. Classification of Emails presents a lot of challenges because of large number of mails. Different machine learning techniques such as K-Nearest Neighbour, Naïve Bayes, SVM and Decision tree have repeatedly been used to tackle these spam mails. Our approach is based on using the KNN algorithm - one of the simplest and efficient classification algorithms and to obtain the maximum accuracy for the best results having small processing time enough for detecting spam mails. Feature extraction is implemented using Particle Swarm Optimization (PSO) which efficiently provides good result for the proposed algorithms in this paper.***

Keywords: Spam detection, KNN, Naïve Bayes, Feature selection, PSO.

1. Introduction

Electronic-mail is an online application and is quick, inexpensive and widely used. It is extremely advantageous to businesses and organizations as it provides the efficient and productive passing of all types of data. Despite its importance email spam became the major problems of the internet world. They can ruin the personal information of individual users and financial to companies. Spam is defined as unsolicited, unrequested emails that is sent knowingly or unknowingly by someone having no relation with the user or organization. Since the spam mails are undistinguishable, user get into the trap and loose his information.

For this, many techniques have been proposed for these types of emails detection in machine learning. Classification of emails presents a lot of difficulties because of huge and different attributes in the data sets and number of emails [1]. Performance of the email classification and feature selection algorithms is affected by the quality of training datasets. This paper concentres on problems with email's spam and narrates how the spam was detected and predict the trained values mention in the dataset. Past studies were categorised mainly into three types, single based machine learning, hybrid and feature computation. In single-based ML, a specific algorithm was used to create a spam filtration method. Some of them are K-Nearest Neighbour, Naïve Bayes, Support Vector Machine (SVM), Random Forest etc. For the better accuracy of classifying the emails, some optimize technique are used, Particle Swarm Optimisation is one of them.

1.1. Spam Classification

Since, identifying spam among emails on a big scale is difficult that's why we analyse the email content and then standardize it [2]. In this dataset is collected, pre-processed and analysed exploratory, then the model is trained, compared and selected for the evaluation. Spam emails are transfer to the folder which is scrap for spam classification. The spam and ham mails are determined by the algorithm. For spam classification, KNN includes two stages: training phase and testing phase. The featured based collected message is used and the final result will be observed. If the test data is similar to the data present in spam training set, then it is determined as spam otherwise as ham.

1.2. Feature Selection

In the classification, feature selection is needed to be done in order to remove the inefficient features and select the acceptable ones. Moreover, it also makes easier to understand the calculation avoiding over fitting and enhance the accuracy in the next step [3]. Feature selection algorithm is classified on the basis of the way they process and estimate features. There are many meta-heuristic techniques like particle swarm optimization etc. Our goal is to find subsets of features that induce minimum error and to increase the accuracy of the model related to the original model accuracy. Feature subsets can be represented by a binary weight that assigns a value of 1 to features in the set and 0 to the rest of the features.

1.3. Objective

The main objective is to develop an effective, productive, efficient email spam filtration system. It focuses classifying the taken test dataset emails into spam and ham mails. It prevents our sensitive data from getting leak. So, it is to enforce the system to deliver an enhanced performance with more accuracy in less time.

1.4. Scope

Different spam classification methods are used in this paper for the taken training and test data sets to know the best algorithm to use in terms of efficiency, accuracy and time boundaries. When real time data is taken into account better outcome can be obtained. 100% accuracy is not possible in any of the algorithm but the best could be obtained to get a wide range of scope. As we are using the KNN and other techniques along with the Particle Swarm Optimisation to process a large number of mails at a time, the scope of the project gets reinforce.

2. Literature Survey

Many scrutinise have been carried out in past on email spam detection. Many scientists have described a focussed literature survey for email spam filtration by machine learning with the use of several types of algorithms such as KNN, Naïve bayes etc. These algorithms result in their own particular levels accuracy.

2.1 Data Set

Self-created and some online dataset is fetched to use in this proposed system. Online dataset is used in training our algorithm and then other for getting final result.

2.2 Existing System

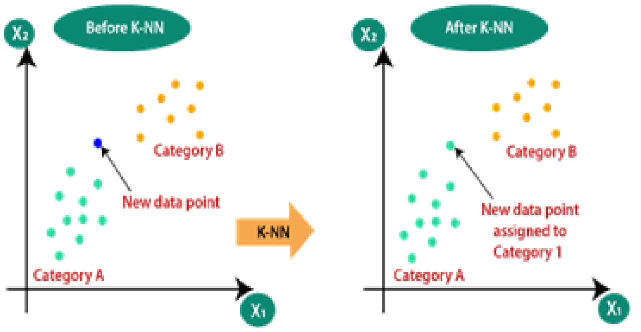
This paper proposed of the KNN and algorithms with an optimisation technique i.e., Particle Swarm Optimization. Particle Swarm Optimization (PSO) is a powerful meta-heuristic optimization algorithm which is inspired by the behaviour of swarm observed in the nature such as fishes and birds. PSO is a Simulation of a simplified social system whose elementary motive was to graphically reproduce the unpredictable choreography of a bird flock.

3. The Spam Classification Machine Learning Methods

The spam determination can be done through various methods. To avoid difficulty in identification of ham and spam mails, initially we analyse content of message and then standardize it. This includes changing all words into lower case, deleting irregular character etc [4]. In short, collection and pre-processing of data is done in order to apply algorithms. The actual mapping of email classification between training and testing sets is done after collection of datasets. Processing and training of training dataset is done to classify the test data effectively.

3.1. K - Nearest Neighbour

KNN is a simple and efficient data mining approach used for classifying data points based on the distance among them in a training dataset. KNN is an instance-based learning (also known as lazy learning) where all computation is deferred until classification. In the text classification, when a new dataset needs to be categorized, it looks for the specific K number of nearest neighbour and choose the categories with the maximum number of times as the categories of the input sample. In this, the training records are used for collection rather than an absolute category representation. It uses closeness to make classification of an individual data point and used for both regression and classification problems. Additionally, detecting the nearest neighbour can be secured by exploring traditional induction methods. To identify whether a mail is spam or not, we look at the classes of the notification that are closest to it. The estimation is a real time process. The similarities between the data points can be determine by the distance metrics which help to form decision boundaries. In this paper the distance measure used is **Euclidean distance**.



3.2 Naïve Bayes

For handling the huge datasets, such method is used which mainly has statistical approach of learning in Machine Learning field and work on Bayes Theorem. This method also handles the probability distribution issue and decide the independent quality of datasets. If a word in testing dataset occurs frequently in spam dataset, it classifies that email as spam. The posterior probability can be evaluated by:

$$P\ (N|M) = \frac{P\ (M|N)\ P\ (M)}{P\ (N)}$$

In modern times for calculating probability of occurrence of events, we mainly use the Bayes Theorem. It suspects that features are independent of each other.

Having the probability of appearing events in the past, we use this classifier which will affect the probability of coming events or the events which are dependent.

It can be used for binary and multiclass classifications. It performs well as compared to the other algorithms. Since, it estimates that all features are independent; therefore it cannot understand the relationship between features.

3.3 Support Vector Machine

This is a popular supervised learning algorithm and used for classification problem in machine learning techniques. It is totally based on the concept of decision points. It trains the model with the help labelled data to obtain an optimal plane of separation that can be used to classify the taken test data. The basic algorithmic rule model of SVM is the linear classifier with the biggest interval outlined within the feature house, that is, to resolve the separated hyperplane which might properly divide the training datasets, and then used the hyperplane as the separation plane. The two sides of the hyperplane are the two different types of samples. Here, representation based on vector form is taken. A kernel function is employed to perform the operations.

$(x_i, y_i)i = 1, 2, \ldots .., N; x_{i,}\ y_i$ said the training samples set as the sample feature vector, that is, {-1; 1}, take-1 as spam and 1 as normal mail. The separation hyperplane is:
$W^T x + b = y$
W, b are the SVM parameters in this formula.

3.4. Decision Tree

Decision tree classifier is used in areas like data extraction, text mining and machine learning. Based on several input variables and on a cascading tree structure, it creates the decision tree model and train it. The leaf nodes represent the class labels, and all internal decisions nodes corresponds to the tests on attributes. In decision tree every node represents value of an attribute. The outcome of it is depicted by the branch of the decision tree and the classes are represented by leaves. Divide and conquer method is used by the decision tree to split the problem. The process of dividing is repeated on every developed subset in iterative manner which is well known as recursive partitioning.

The algorithm's input consists of the teaching records S and attributes sets T. The algorithm works on recursively choosing the simplest attributes to divide the info and extending the leaf nodes of the tree until stopping standard meet.

This uses classification and regression tree - based algorithm with Gini index as cost function. It is given by the equation:

Gini = 1-$\sum_{i=1}^{n} (p_i)^2$

4. Proposed System

In this paper, classification of spam emails is done by implementing KNN algorithm. Also, PSO algorithm which handles a huge number of emails at one time and hence performance of classification and speed of execution increases. The outcome of the different algorithms and algorithm with optimization are compared in this proposed paper.

4.1 Particle Swarm Optimization

Particle Swarm Optimization (PSO) is an optimization technique which is based on the behaviour of the group of birds and other animals that acquire a collective behaviour. It is a computational method which is iteratively tries to improve the solution to optimize a problem. In PSO a global solution is obtained from a set of local solution, it gets closer to the best solution in every iteration as each particle share their experience. The motive is to discover the optimal solution of the fitness function explained over the search space. It performs on the basis of two main dynamic vectors which are particle position and velocity. The movement of the particle is affected by the position of the best value founded by taken particle called local optimum and the position of the best value over all particles known as global optimum. Each particle has the ability to change their trajectory as per the past knowledge and sharing properties with other particles to accomplish better solution with each increase in iteration.

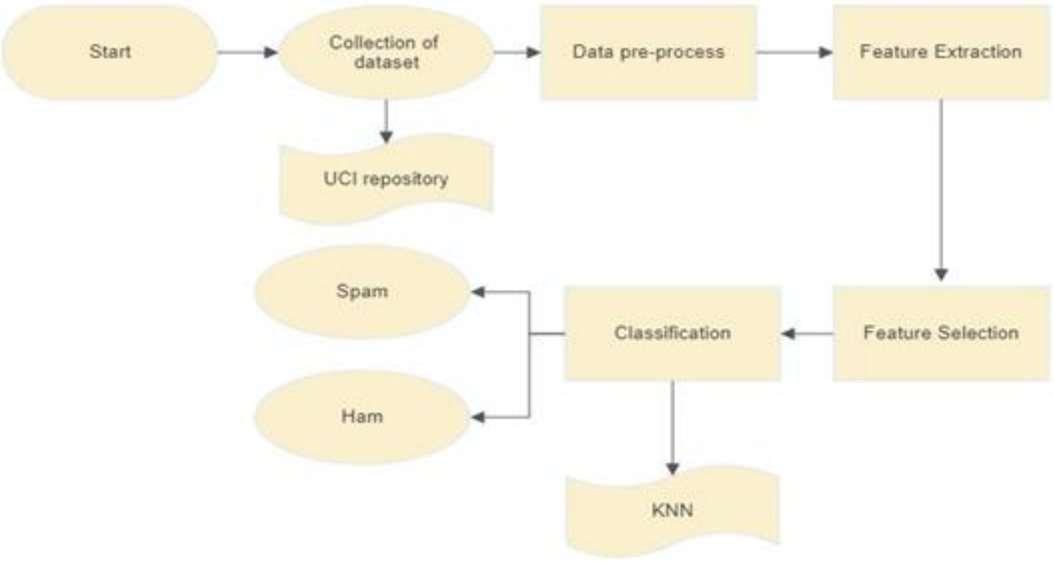
The PSO is used in many optimization tasks like estimation of the non-linear parameters; optimize the parameters of several algorithm for email spam classification and other different problems in different domains.

5. Architecture

5.1 High Level Design

Firstly, training of spam datasets is done and then at every next stage processing is done and ultimate result is given by the reducer. The retrieved features from the input datasets are the labelled words [5].

Here, the KNN classifier takes selected features as input. The high-level design of KNN is shown in given figure.



5.2. Module Description

5.2.1. Collection of Datasets

The datasets consist of data for at least one member, regarding the number of rows. The set of data for experiment is taken from UCI machine learning depository. The spam-based set of data contains 5857 emails. The particular variable for each column represented provides the method to the discrete database table and statistical data, and a certain member of the dataset is characterised by the row.

5.2.2. Data Pre-processing

In data extraction approach, it is the prime step for all the machine learning projects. For this phase, undefined dataset will be executed for the data cleaning such as labelling; stop words removal and stemming are carried out. It also includes the removal of all the white spaces, lowered the alphabet, remove the remaining punctuation, converting all the words to their root words [6]. The main advantage of processing stage is organizing the data in order to illustrate classification. Operations are done in order to remove noisy data and place only the favourable information to make the upcoming operation easy to implement. The basic requirement of data pre-processing is the training of datasets. However, most of the modernistic machine learning algorithm are capable of extracting information from dataset and reserves features in distinct way.

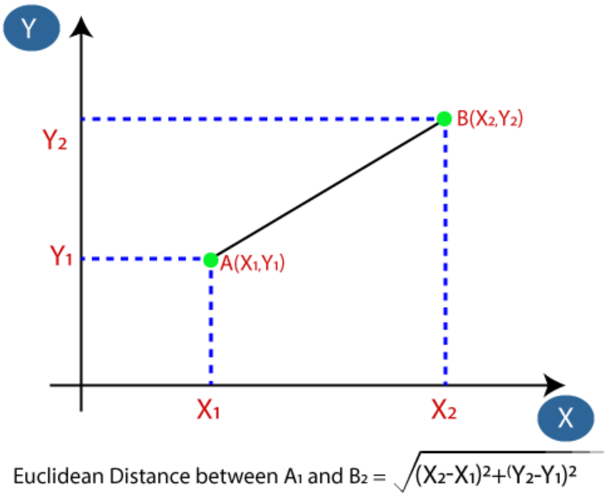
5.2.3. Feature selection

After the pre-processing, some features of emails are extracted with the help of feature extraction and represent it as feature vector. The process of converting attacks to directly classifiable form is called vectorization. In essence, each element of text is converted to a metrics of numbers, and every data of this metrics corresponds to a particular token. In our case, it is restricted to ham and spam. It is a process before classification class. Some important features are taken by this algorithm and eliminates irrelevant, redundant and noisy features for more accuracy.

5.2.4. KNN Classifier

K - Nearest Neighbour (KNN) is a type of supervised learning which is used for many applications such as image classification, data mining, an many others. This classification algorithm attempts to attain the similar set of points are taken to know their individual class as class label [7]. The class having majority of votes is selected by simple voting as the element vector class [8]. KNN emphasize k on the basis of similarities between training and testing data points. Here, **Euclidean Distance** is used to measure the distance metrices. The formula is as follow:

$$\mathbf{D\ (X,\ Y)} = \sqrt{\sum_{i=1}^{n}\ (X_i - Y_i)}^2$$



6. Detailed Description

6.1. Dataset Processing

The dataset consists of data for at least one member, regarding the number of rows.

- Characteristics of datasets are multivariant.
- Integer, real are the attributes characteristics.
- Associative tasks = Classification

Spam based dataset helps in training and testing the KNN classifier. Classification is done with the help of KNN classifier to classify data as spam or ham.

6.2. PSO for feature extraction

A subset of essential feature is taken which eliminates irrelevant and noisy features. It provides more precise data representation. By this, memory requirement storage can be obtained. Additionally, other methods for reducing features can improve performance of text classification such as pruning and clustering.

It is implemented with the help of Particle Swarm Optimization. In this, PSO is also combine with different algorithms to enhance the traits of feature selection. Here, PSO based hybrid classifier that combines PSO with other algorithms to enhance the accuracy with an appropriate feature subset.

6.3. Implementation of KNN

The spam-based dataset contains 5857 emails records in which datasets are a mixture of ham and spam messages. The dataset is divided into test and train set. Train set consist of 73% ham and spam emails where the test set contains 27% ham and spam emails [9].

Algorithm:-KNN is the simplest and best technique for classification which is used for classification and regression. It is a non-parametric technique having two phases one is Testing and other is training.

- Stage 1: Training phase-Trained datasets are store after training process.

- Stage 2: Testing phase-In this for a given input data x, its k nearest neighbours among the datasets are determined in the training set. If there are a greater number of spams among these neighbours, then classify given message as spam otherwise as ham.

Steps:-
1) Determine k = number of nearest neighbours.
2) Calculate the distance between the training samples and test data.
3) Find nearest neighbours based on the Nth shortest distance.
4) Assemble the category X of the nearest neighbours.
5) Take the larger part of the classification of closest neighbours as the prediction value of the test data.

7. Performance Analysis

Checking and comparison of results is done by training this system. This provides better accuracy [10-15]. Evaluated results are given by different classifiers to the user, and then they are compared. From this spam and ham data can be identified. The result of each classifier is represented in following graph and table. A new CSV file is made to test the train data. The overall work of the proposed system is compared with the PSO-NB different algorithms like Naïve Bayes Classifier, KNN and SVM is shown in the Table 1.

Better performance can be seen in the PSO-NB based method as clearly shown in table 1.

Table 1: Performance study of the proposed method

Performance Measure	Accuracy	Precision	Recall
KNN	75.85	75.78	75.97
NB	87.3	90.47	90.9
SVM	79.5	79.02	68.67
PSO - NB	98.53	98.68	98.57

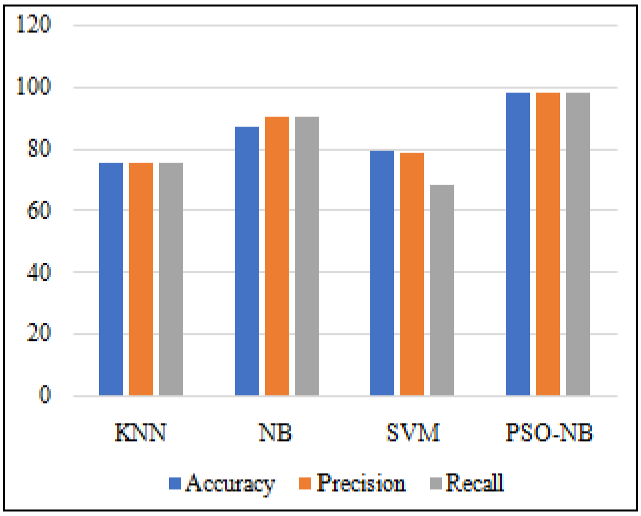


Figure: Comparative study of the proposed system with different methods

8. Conclusion

In this analysis, we analysed machine learning methods and their application to the domain of spam detection. A study of algorithms has been used for classification of emails as spam or ham is provided. The process and architecture of email spam filter were looked into. This paper analysed some of the online available datasets and metrics used to measure the effectiveness and efficiency of any spam detector. The challenges in handling the risk of spam emails were demonstrated and comparative studies available in literature were fulfilled.

The first algorithm used in this paper is based on KNN spam filter. In the proposed system the value of k = 11 is taken into account. Since it is a weak classifier, we can enhance the accuracy of the spam detection using other optimization methods. We achieved the accuracy of 75.85%. Next algorithm is based on the Naïve Bayes classification algorithm. This algorithm achieved the accuracy of 87.3% and with its optimization with Particle Swarm Optimization (PSO), we get the improved accuracy of 98.53%. On applying SVM, we achieved the accuracy of 79.5%

Declarations

Ethical Approval
Not Applicable

Competing Interests
The author declares that they have no competing interests.

Author's Contribution
Akash Raghuvanshi is responsible for the design and overall concept or framework. Akash Raghuvanshi and Awadhesh Kumar are responsible for improvement in optimization algorithm. Akash Raghuvanshi, Awadhesh Kumar and Nilesh Chandra are responsible for experimental part and data analysis. All authors read and approved the final manuscript.

Funding
Not Applicable

Availability of data and materials
Not Applicable

Consent to Publish
I, Akash Raghuvanshi, give my consent for the publication of identifiable details with in the paper to be published in the Journal.

References

[1] Bajaj, Kamini Simi, and Josef Pieprzyk. "A case study of user-level spam filtering." Proceedings of the Twelfth Australasian Information Security Conference-Volume 149. Australian Computer Society, Inc., 2014.

[2] Roy, S., Patra, A., Sau, S., Mandal, K., & Kunar, S. (2013). An Efficient Spam Filtering Techniques for Email Account. American Journal of Research, 2 (10).

[3] Brain Whit Worth, Ellizbet Whit Worth, "Spam and the social technical gap", IEEE Computer society, pp. 38-45, 2004. Brain Whit Worth, Ellizbet Whit Worth, "Spam and the social technical gap", IEEE Computer society, pp. 38-45, 2004.

[4] Nazirova, Saadat. "Survey on spam filtering techniques. " Communications and Network 3.03 (2011): 153.

[5] Idris and A. Selamat, “Improved email spam detection model with negative selection algorithm andparticle swarm optimization, ” *Appl. Soft Comput.*, vol.22, pp. 11-27, 2014.

[6] Idris, A. Selamat, N. Thanh Nguyen, S. Omatu, O. Krejcar, K. Kuca, and M. Penhaker, “A combinednegative selection algorithm-particle swarm optimization for an email spam detection system, ”*Eng. Appl. Artif. Intell.*, vol. 39, pp. 33-44, 2015.

[7] Y. Zhang, S. Wang, P. Phillips, and G. Ji, “BinaryPSO with mutation operator for feature selection using decision tree applied to spam detection, ” *Knowledge-Based Syst.*, vol. 64, pp. 22-31, 2014.

[8] D. M. Farid, L. Zhang, C. M. Rahman, M. A. Hossain, and R. Strachan, “Hybrid decision tree and naïveBayes classifiers for multi-class classification tasks, ”*Expert Syst. Appl.*, vol. 41, no. 4 PART 2, pp. 1937-1946, 2014.

[9] P. H. C. Guerra, D. Guedes, J. W. Meira, C. Hoepers, M. H. P. C. Chaves, K. Steding-Jessen, Exploring the spam arms race to characterize spam evolution, in Proceedings of the 7th Collaboration, Electronic messaging, Anti-Abuse and SpamConference (CEAS), Redmond, WA, 2010, July.

[10] E. S. M. El-Alfy and R. E. Abdel-Aal, “Using GMDH-based networks for improved spam detection and email feature analysis, ”Appl. Soft Comput. J., vol. 11, no. 1, pp. 477-488, 2011, doi: 10.1016/j. asoc.2009.12.007.

[11] N. O. Hamed, A. H. Samak, and M. A. Ahmad, “ Cloud E-mail Security: An Accurate E-mail Spam Classification Based on Enhanced Binary Differential Evolution ({BDE}) Algorithm, ” 1, vol. 5955, 2021.

[12] V. Vinitha and K. R. Dhanaraj, “MapReducemRMR: Random Forests-Based Email Spam Classification in Distributed Environment, ” 2019, pp. 241-253.

[13] S. Sumathi and G. K. Pugalendhi, “Cognition based spam mail text analysis using combined approach of deep neural network classifier and random forest, ” J. Ambient Intell. Humaniz. Comput., vol. 12, no. 6, pp. 5721-5731, 2021, doi: 10.1007/s12652-020-02087-8.

[14] F. Soleimanian, Gharehchopogh, and S. K. Mousavi, “A New Feature Selection in Email Spam Detection by Particle Swarm Optimization and Fruit Fly Optimization Algorithms, ” J. Comput. Knowl. Eng., vol. 2, no. 2, 2019, doi: 10.22067/cke. v2i2.81750.

[15] D. Dua and C. Graff, “{UCI} Machine Learning Repository.” 2017, [Online]. Available: http://archive. ics. uci. edu/ml.

[16] Faramarzi, M. Heidarinejad, S. Mirjalili, and A. H. Gandomi, “Marine Predators Algorithm: A nature-inspired metaheuristic, ” Expert Syst. Appl., vol. 152, no. 113377, pp. 1-48, 2020, doi: 10.1016/j. eswa.2020.113377.

[17] S. Mirjalili, “Dragonfly algorithm: a new meta-heuristic optimization technique for solving single-objective, discrete, and multi-objective problems, ” Neural Comput. Appl., vol. 27, no. 4, pp. 1053-1073, 2016, doi: 10.1007/s00521-015-1920-1.