Novel Email Spam Classification using Integrated Particle Swarm Optimization and J48

Harpreet Kaur
Department of Computer Science & Engineering,
Amritsar College of Engineering & Technology,
Amritsar, Punjab

Ajay Sharma
Department of Computer Science & Engineering,
Amritsar College of Engineering & Technology,
Amritsar, Punjab

ABSTRACT
E-mails have become an integral part of both private and professional lives and can also be studied as formal papers in communication between users. Several activities such as spam detection and classification, subject classification, etc. can be done by email's data mining and analysis. Review has shown that the use of unsupervised filtering to filter the input data set is ignored by the most of the existing researchers. The use of hybridization of data mining techniques is ignored in order to improve the accuracy rate further for detection of fraudulent emails. Most of the existing techniques are limited to some significant features of emails therefore utilising more features may provide more significant results. The overall objective of this work is to propose an integrated particle swarm optimization based J48 algorithm to enhance the accuracy rate further.

Keywords
Email, Spam Email Classification, Particle swarm optimization, j48, Unsupervised Filter

1. INTRODUCTION
The past 10 years have witnessed considerable increase in usage of internet and the trend seems to be continuously growing in future. Now Internet has been proved as a vital part of our routine life. With increase in the usage of internet, e-mail turns out to be a powerful tool for gaining ideas besides exchanging information. E-mail is being utilized by wide range of users. This is because in present times almost all activities from personal to business are being carried out by emails. Emails proves to be upper hand due to various positive factors which includes negligible delays in transmission, enhanced security to the information being sent and overall low costs. On the other hand there are several issues that become barrier in professional procedure of email and spam e-mails or unsolicited bulk emails (UBS) tops the list. It is deeply rooted in the internet. These unsolicited emails are directed to large number of users arbitrarily without wastage of any resources. These emails can include harmful malwares or phishing links, making email users vulnerable to several security breaching attacks and can even crash the mail servers.

2. RELATED WORK
Bharati et al. [2010] [2] describes that data mining is a method which finds useful habits from large amount of data. The paper talks about data mining algorithms and techniques and states how adapted data mining technology results in the improvement in the businesses and lead to extraordinary results.

Cantú-Paz et al. [2001] [3] presents the use of more capable and efficient nature inspired evolutionary methods to data mining augmented with individual interaction to handle conditions for which concept descriptions are abstract and hard to define, hence not really quantifiable in a complete sense. Finally, it concludes some ideas for the methods for future implementations.

Bai, Qinghai [2010] [4] presents a heuristic optimization method called as a Particle swarm optimization (PSO) based on swarm intelligence. The algorithm proves to be upper hand due to lesser number of parameters and hence easy to implement. The central notion of principle of algorithm and various plusses and drawbacks are summarized in the paper.

Khan, Amreen et al. [2010] [5] describes the utilization of Particle Swarm optimization for cluster analysis. The efficiency of Fuzzy C-means clustering produce increment in performance and maintains additional diversity within the swarm and additionally permits the particles to be sturdy to trace the dynamic atmosphere.

Sarkar, Sunita et al. [2013] [6] paper presents a short survey on PSO application in data cluster. Cluster with swarm-based algorithms (PSO) is rising like an alternate to a lot of standard clustering approach. PSO is a population-based random research algorithmic rule with the aim of mimics the potential of swarm.

Jindal et al. [2007] [10] suggests that the reviews provided by users to the vendors holds an upmost importance to the makers as well as new customers. However they also lead to spam due to fake positive or negative reviews. Authors performed spam detection using Duplicates detection and Spam classification to review spam.

Michal et al. [2012] [16] suggests two totally dissimilar algorithms in favour of spam detection. The first one works on Bayesian filter. However this filter is enhanced by exploiting information compression algorithms just during the case if Bayesian filter fails to make decision. The second algorithmic rule is centred on document classification algorithmic rule making use of Particle Swarm optimization. The paper concludes by depicting their results.

Vyas et al. [2015] [18] considers completely another techniques for classification exploiting WEKA to filter spam mails. Appreciable accuracy and least time among another technique have been shown by Naive bayes technique. The paper presents a comparative revision of all procedure into the terms of accurateness as well as time occupied is produced.

Qian et al. [2010] [20] put forwards the development of online unsupervised spam learning and detection scheme. The learning algorithm is efficient in mining repeated occurrences of terms that are generated by templates and rarely seen in spam. The results are comparable to those of the de-facto supervised-learning-based filtering systems like Spam Assassin (SA), signifying that unsupervised e-mail spam filtering can be efficiently used to detect net spam.

3. DATASET DESCRIPTION

Spambase Dataset is collected from the UCI Machine Learning Repository, which has the data of 4601 E-mail messages. Every instance of the Spambase dataset has 58 attributes. The majority of the attributes signifies the frequency of the particular words or else characters into the e-mail to correspond towards the instances.

- Word freq w: 48 attributes is used to define the frequency as well as percentage of the words within e-mails.
- Char freq c: 6 attributes defines the frequency of the character c as well as percentage of character in emails.
- Char freq cap: 3 attributes define the longest length, average length and the entire number of capital letters.
- Spam class: target attribute declared that the e-mail is spam or not.

4. DECISION TREE

A decision tree also defined as predictive machine-learning model so that it can be used for making decision of target value (dependent variable) of variety of attribute values of obtainable information. The internal nodes of decision tree act as several attributes, the branches within the nodes gives the feasible values of the attributes and the terminal nodes gives the finishing value (classification) of the non-independent variable.

The predicted attribute is identified as non-independent variable which is determined by or depend upon the values of each and every attributes. The attributes that are used to conclude the values of the dependent variable are categorized like independent variables within the dataset.

The most powerful approach within knowledge discovery as well as data mining is the decision tree. This defines the technology of researching large as well as complex information within the direction of discovers the useful patterns. This process is more efficient, cost-effective and accurate. Decision tree is extremely successful tool in a lot of areas like data and text mining, information extraction, machine-learning as well as pattern recognition.

J48 creates the decision tree that depends upon the attribute values of the available training data towards classification of new item. So whenever it analyses all items of training set, it recognize the various attributes to discriminate the variety of instances more clearly. This feature differentiates all the instances which are used to classify them the best is supposed to contain the highest information achieve.

A tree defines the root node, leaf nodes that are used to signify as any classes and internal nodes that are used to signify as a test conditions which be applied on top of attributes as shown within fig 1.

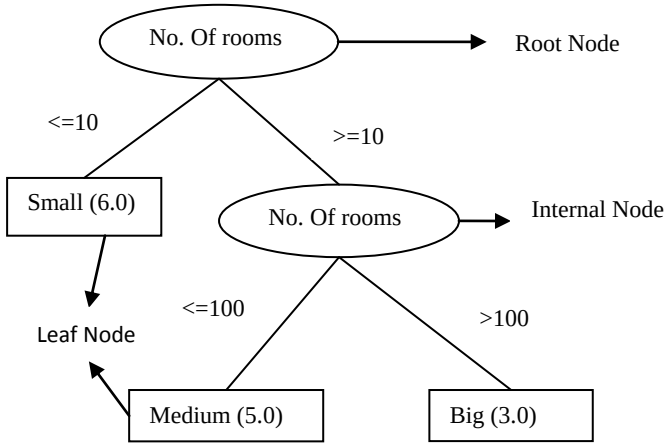

Fig 1: Decision Tree

Decision tree offers a lot of advantages in data mining, as given below

- It can easily learn by the end user.
- It can easily operate on various input data like Nominal, Numeric and textual.
- It can be performed on the training data with missing attributes values.
- The performance of the tree is very high as compare to the efforts.

5. PROPOSED METHODOLOGY

Proposed methodology completely depends upon the data mining approach used for classification of emails as spam and non-spam emails from spambase dataset. This methodology will utilise Particle Swarm Optimization (PSO) algorithm to evaluate the emails as spammed or not in more accurate manner. Also to improve the accuracy rate further, the unsupervised filtering is also used as pre-processing tool while classifying spam e-mails. Methodology suggests different section of data mining procedure as: data selection, data pre-processing, data classification and result analysis. [Fig 2]

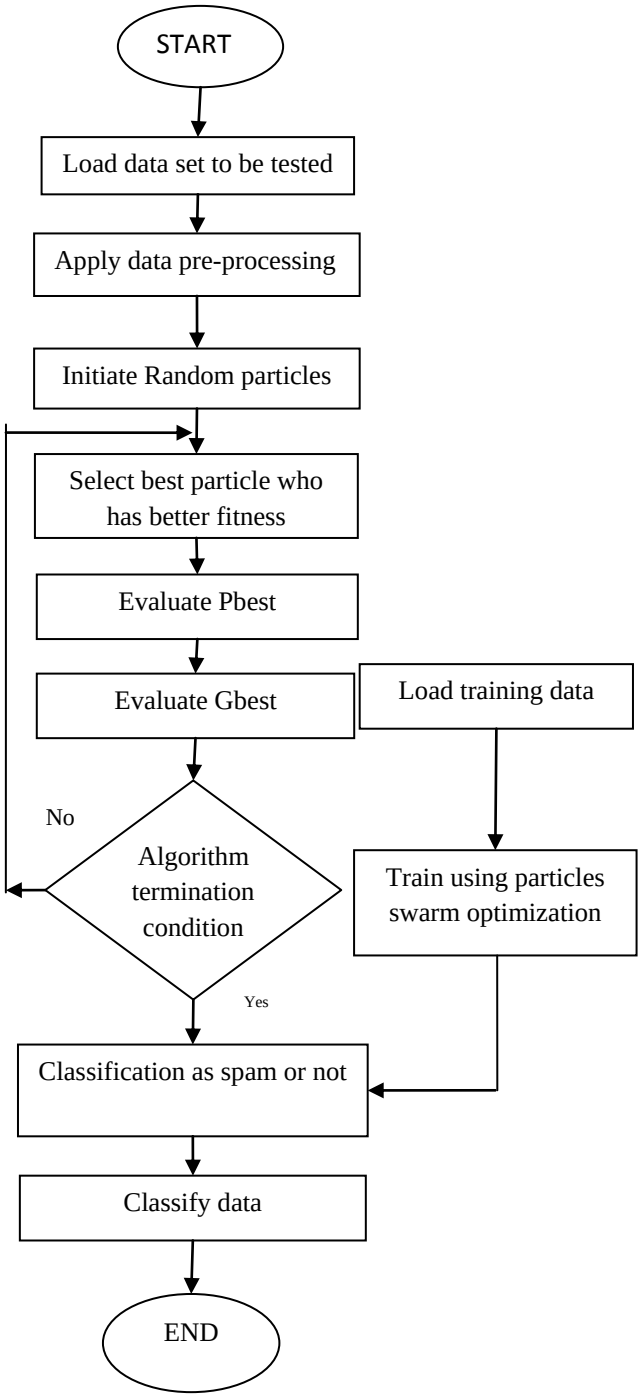

Fig 2: Proposed Architecture

5.1 Dataset

Spambase dataset is loaded as a data source. This dataset contain 4601 emails in which 1813 are spam and 2788 are non-spam emails. Pre-processing task is implemented on this dataset to extract the appropriate information.

5.2 Data Pre-processing

Real world data is normally incomplete, loud, and inconsistent because of its large size. Various tasks are performed in data pre-processing for data cleaning and data reduction.

5.3 Particle swarm optimization

PSO is a technique based on the movement and intelligence of swarms. This technique is developed by James Kennedy (social-psychologist) and Russell Eberhart (electrical engineer) in 1995. Numbers of agents (particles) are used that lie in the group and move in the region of in a search space looking for the best solution with some velocity. These particles are not static. When particles move in the search space they change their velocity. Particles move in the search space to find:-

a) Pbest (local maxima)-best fitness for individual particle.

b) Gbest (global minima)-best fitness in all the particles.

Genetic algorithm operates on some kind of generations; similarly PSO operates in iterative manner, means some kind of iteration is applied on each particle. In each iteration, every particle gets one chance to move. The particles will move by the magnitude of their velocity. In higher velocity, better steps are taken by the particles and in lower velocity; very small steps are taken by the particles.

Initially we take random approach where the particles are located at different or distinctive region in this fitness landscape and assigning some kind of velocity to each particle.

First, move the particles in this fitness depending upon the velocity:-

$$\boldsymbol{Pi^{t+1} = Pi^{t} + Vi^{t+1}} \quad \textbf{(1)}$$

Where,

Pi – Particle i

$Pi^{(t)}$ – current position of i^{th} particle at the number of iteration t

$Pi^{(t+1)}$ – modified position of i^{th} particle at the number of iteration t+1

$Vi^{(t)}$ – current velocity of i^{th} particle at the number of iteration t

$Vi^{(t+1)}$– modified velocity of i^{th} positionat the number of iteration t+1

So, position of any particle is modified or updated by the magnitude of velocity at any position. All the particles move towards the global minima.

Velocity of the particle can be modified in such a way, that particles move from one place to another, so because of this reason velocity of the particle is modified and current position is also modified. In this way, particles jump towards the global minima. Each and every particle stuggle to update its current position by using the following information:-

a) Current position

b) Current velocity

c) Distance between current position and Pbest

d) Distance between current position and Gbest

Particles modified their position by using following expression:

$$\boldsymbol{Vi^{t+1} = Vi^{t} + C_1 Rand_1 (Pbest_i - Pi^{t}) + C_2 Rand_2 (Gbest - Pi^{t})} \quad \textbf{(2)}$$

Where,

Vi^{t}- Defines the velocity of i^{th} particle at iteration t,

C_i – constants used for the speed up the process,

Rand() – Define as a random number between (0,1),

Pbest – pbest for the particle i,

Gbest –gbest for all the group of particles

5.4 Data classification

In Data mining, classification technique is used to classify the large dataset and split into different classes. Here we are working with j48 for the classification of spam and non-spam e-mails from the spambase dataset.

5.5 Results

In this section, we analyses the result and define the correctly classified and incorrectly classified instances, accuracy and confusion matrix. Different types of parameters are followed as:

Correctly classified Instance: It is used to define ability of algorithm, how much classify the instances correctly.

$$Correctly\ classified\ instances = TP + TN \quad (3)$$

Incorrectly Classified Instances: It is used to define the instances which are incorrectly classified.

$$Incorrectly\ classified\ instances = FP + FN \quad (4)$$

Kappa statistic: The kappa instance or value is a metric which is used to compare the observed accuracy with an expected accuracy.

$$Kappa\ Instance = \frac{observed\ \ accuracy\ -expected\ \ accuracy}{1-expected\ \ accuracy} \quad (5)$$

Mean absolute error: $\frac{Incorrectly\ \ Classified\ \ Instances}{Total\ \ Number\ \ of\ \ Instances}$ (6)

Root mean squared error: $\sqrt{mean\ absolute\ error}$ (7)

Relative absolute error: $\frac{Absolute\ \ error}{true\ \ value} * 100$ (8)

Root relative squared error:$\sqrt{relative\ absolute\ error}$ (9)

True Positive (TP): It is used to represent the instances that are truly classified.

False Positive (FP): It is used to represent the instances that are falsely classified.

$$True\ Positive\ Rate(TPR) = \frac{TP}{TP + FN} \quad (10)$$

$$False\ Positive\ Rate(FPR) = \frac{TN}{FP + TN} \quad (11)$$

ROC Area: Receiver operating characteristic (ROC) is define as an area under the curve. ROC curve is a graphical plot that represents the performance of the binary classifier system. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold setting.

Precision (P): It is used to define the fraction of retrieved instances in the dataset that are relevant.

$$P = \frac{TP}{TP + FP} \quad (12)$$

Recall (R): It is used to evaluate the fraction relevant instances that are retrieved.

$$R = \frac{TP}{TP + FN} \quad (13)$$

Accuracy (ACC):

$$\frac{TP + TN}{TP + TN + FP + FN} \quad (14)$$

Here, TP is true positive, TN is true negative, FP is false positive and FN is false negative. TPR is also defined as a sensitivity which is used to evaluate recall. FPR is also known as fall-out.

Table 1. Classification Analysis

Correctly classified instances	4524
Incorrectly classified instances	77
Kappa statistic	0.9649
Mean absolute error	0.0305
Root mean squared error	0.1235
Relative absolute error	6.3914%
Root relative squared error	25.2815%
Coverage of cases (0.95 level)	99.1306%
Mean rel. region size (0.95 level)	53.521%

Table 1 shows the classification of instances based on different parameters using Particle Swarm Optimization and j48.

Table 2. Confusion Matrix Analysis

Parameters	Class 0	Class1	Aggregate
TP rate	0.989	0.974	0.983
FP rate	0.026	0.011	0.02
Precision	0.983	0.983	0.983
Recall	0.989	0.974	0.983
F-Measure	0.986	0.979	0.983
ROC Area	0.995	0.995	0.995

Table 2 shows the aggregate of different parameters on the bases of class 0 and class 1. Class 0 represents the e-mails as a non- spam and class 1 represents the e-mails as a spam. These parameters are used to analysis the confusion matrix. Confusion matrix are extremely valuable used for evaluating the performance of classifier, as they provide an specific table layout which is used to represent the distribution of correct and incorrect classified instances.

6. CONCLUSION

Review has shown that use of unsupervised filtering to filter the input data set is ignored by most of the existing researchers. The use of hybridization of data mining techniques is ignored in order to improve the accuracy rate further for Detection of fraudulent emails. Most of the existing techniques are limited to some significant features of emails therefore utilising more features may provide more significant results. In order to overcome these issues, this research work has proposed a novel technique. The proposed technique will utilise Particle Swarm Optimization algorithm to evaluate the emails as spammed or non-spam in more accurate manner. Also to improve the accuracy rate, further the unsupervised filtering is also used as pre-processing tool while classifying e-mail spam.
In order to obtain the objectives in near future Matlab and WEKA knowledge flow will be used.

7. REFERENCES

[1] Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar, "Introduction to data mining," Vol. 1. Boston: Pearson Addison Wesley, 2006.

[2] Bharati, Mrs, and M. Ramageri "Data mining techniques and applications," (2010).

[3] Bai, Qinghai, "Analysis of particle swarm optimization algorithm," Computer and information science 3.1 (2010): 180.

[4] Khan, Amreen, N. G. Bawane, and Sonali Bodkhe, "An analysis of particle swarm optimization with data clustering-technique for optimization in data mining," (IJCSE) International Journal on Computer Science and Engineering 2.07 (2010): 2223-2226.

[5] Sarkar, Sunita, Arindam Roy, and Bipul Shyam Purkayastha,"Application of Particle Swarm Optimization in Data Clustering: A Survey," International Journal of Computer Applications 65.25, (2013).

[6] Shahreza, M. Lotfi, et al., "Anomaly detection using a self-organizing map and particle swarm optimization," Scientia Iranica 18.6 (2011): 1460-1468.

[7] Rini, Dian Palupi, Siti Mariyam Shamsuddin, and Siti Sophiyati Yuhaniz. "Particle swarm optimization: technique, system and challenges."International Journal of Computer Applications 14.1 (2011): 19-26.

[8] Al-Kadhi, Mishaal Abdullah, "Assessment of the status of spam in the Kingdom of Saudi Arabia," Journal of King Saud University-Computer and Information Sciences 23.2 (2011): 45-58.

[9] Kumar, R. Kishore, G. Poonkuzhali, and P. Sudhakar,"Comparative study on email spam classifier using data mining techniques," Proceedings of the International MultiConference of Engineers and Computer Scientists,Vol. 1, 2012.

[10] Jindal, Nitin, and Bing Liu,"Review spam detection," Proceedings of the 16th international conference on World Wide Web. ACM, 2007.

[11] Günal, Serkan, et al., "On feature extraction for spam e-mail detection,"Multimedia content representation, classification and security, Springer Berlin Heidelberg, 2006, 635-642.

[12] Alsmadi, Izzat, and Ikdam Alhami,"Clustering and classification of email contents," Journal of King Saud University-Computer and Information *Sciences* 27.1 (2015): 46-57.

[13] Rathi, Megha, and Vikas Pareek, "Spam Mail Detection through Data Mining-A Comparative Performance Analysis," International Journal of Modern Education and Computer Science 5.12 (2013): 31.

[14] Kumar, R. Kishore, G. Poonkuzhali, and P. Sudhakar, "Comparative study on email spam classifier using data mining techniques," Proceedings of the International MultiConference of Engineers and Computer Scientists, Vol. 1, 2012.

[15] Elssied, Nadir Omer Fadl, Othman Ibrahim, and Waheeb Abu-Ulbeh,"AN IMPROVED OF SPAM E-MAIL CLASSIFICATION MECHANISM USING K-MEANS CLUSTERING," Journal of Theoretical & Applied Information Technology 60.3 (2014).

[16] Pérez-Díaz, Noemí, et al.,"Rough sets for spam filtering: Selecting appropriate decision rules for boundary e-mail classification," Applied Soft Computing 12.11 (2012): 3671-3682.

[17] Salehi, Saber, and Ali Selamat,"Hybrid simple artificial immune system (SAIS) and particle swarm optimization (PSO) for spam detection," Software Engineering (MySEC), 2011 5th Malaysian Conference in. IEEE, 2011.

[18] Sharma, Amit Kumar, and Renuka Yadav, "Spam Mails Filtering Using Different Classifiers with Feature Selection and Reduction Technique,"Communication Systems and Network Technologies (CSNT), 2015 Fifth International Conference on. IEEE, 2015.

[19] Vyas, Tarjani, Payal Prajapati, and Somil Gadhwal, "A survey and evaluation of supervised machine learning techniques for spam e-mail filtering,"Electrical, Computer and Communication Technologies (ICECCT), 2015 IEEE International Conference on. IEEE, 2015.

[20] Qian, Feng, et al., "A case for unsupervised-learning-based spam filtering,"ACM SIGMETRICS Performance Evaluation Review, Vol. 38, No. 1. ACM, 2010.