Volume 9, Issue 4
August 2020

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

A Machine Learning Approach Based Spam Filtering With the Use of Neural Network

[1]Ajinkya Gulunjkar and [2]Dr. Rakesh Rathi

[1]Department of Information Technology, Government Engineering College Ajmer, Rajasthan, India

[2]Department, Computer Science and Information technology, Govt. Engineering College Ajmer, Rajasthan, India

E-mail: [1]aj.gulu@gmail.com

ABSTRACT

The increase in the number of unnecessary spam emails has made it important to create more efficient and robust spam filters. Recent machine learning techniques are used to detect and delete spam messages effectively. Spammers ' growing propensity to compromise legal accounts has grown as an evasive and more efficient way of propagating spam. The methods for identifying vulnerability accounts to identify spammers adhering to this behaviour have now become increasingly relevant. We have been working on hybridizing the evolutionary and neural network paradigm in this article. We have seen that our proposed model is more efficient than the existing research methodology with 83% accuracy.

Keywords: *Email, spam filtering, GA-RWN, data mining, CSDMC2010 dataset*

1. INTRODUCTION

Email is a big improvement in traditional communication networks, as it utilizes technology effectively, efficiently, securely and economically. The massive distribution of unnecessary, harmful email messages known as spam emails is a significant boost to online communications. One critical challenge is the creation of adequate filters to catch this e-adequately and achieve high efficiency. To resolve this problem, researchers in Machine Learning (ML) have established various approaches [1]. During machine learning, support vector machines (SVMs) contributed significantly to the development of spam filtering. Based on the Vector Machine support, various schemes were drawn up by Text Classification Method (TC). The collection of kernels as they freely influence the partitioning of emails in the consistency space [2] is a crucial issue when using SVM. The spam filtering is explained here in fig. 1.

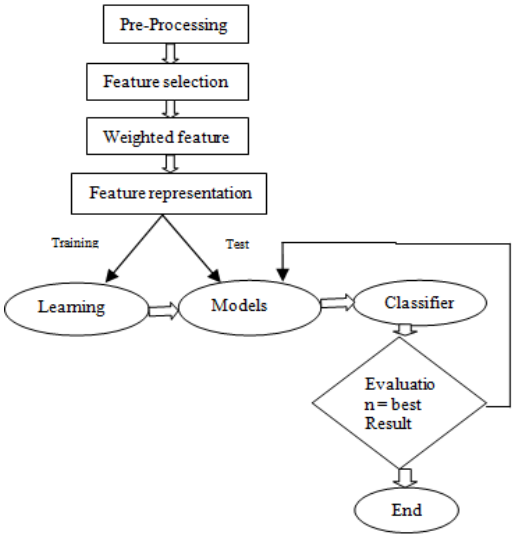

Fig. 1. Process of Spam Filtering

Various forms of e-mail filtering work, some of them acquired accuracy, some continue. The summary of the researcher suggests that e-mail filtering is an e-mail processing mechanism based on such parameters. As different email filtering mechanisms exist, incoming and outgoing filtering is known Incoming Filtering is the mechanism by which an Internet address message is read and the outgoing filters are read by the local user. Spam filtering, which takes place using an anti-spam system, is perhaps the most efficient and useful scanning of communications. Spammers are innovative and use complex mechanisms that continuously adjust to avoid spam control and spam filtering [3, 4]. Spammers are a threat. Spam filtering is an unwelcome message prevention mechanism that avoids entrance into the user's inbox. Many anti-spam solutions have since been developed to avoid unwanted large-scale correspondence. Several antifoam strategies range from false negatives (missing spam) to false positives (rejecting suitable e-mails) that serve as a dissuasion for the most successful antifoaming in the system [5]. Therefore, the primary need for web users is an insightful and efficient spam filtering system.

2. RELATED WORK

The global research community's interest in email spam filtering is growing rapidly. We present similar reviews in this section that have been published in this area in the literature. This approach is used such that the problems to be discussed are presented and the gaps with our current analysis are highlighted.

In [6],a short survey was submitted to explore the disparity between the probability of using message filters and the possible information.

Techniques to postulate spam email detection to promote the successful use of spam filtering technology. The report did not provide explanations of learning algorithms, the techniques for modeling, the publicly accessible data and the framework for e-spam. The criteria used for previous studies in

the evaluation of other possible techniques are also not provided.

[7] Reviewed the various techniques used in handling unnecessary spam communications. The article also categorizes e-spam in many hierarchical directories and regulates automatically the tasks necessary to respond to an email. However, some of the vulnerabilities found in the review paper include machine learning, the e-spam architecture, previous algorithms, and the simulation environment, many of which have not already been listed. For this report, the attention is on the aspects of the spam filtering system, titled "Spake filters and email services." Then he put in place a framework to connect multiple filters to an advanced filtration model with the learning ensemble's algorithm. The paper also clarified in an e-mail program the notions of Operable E-Mail (OE). Moreover, OE's success of an e-mail assistant and other smart technology on the world's largest e-mail network was demonstrated [8]. Moreover, as it has been written more than ten years ago, the survey paper did not contain recent publications

Cormack [9] the previously suggested spam filtering algorithms have been checked up to 2008 with a special focus on systems performance. The study focuses on the interaction between spam filtering and other spam filtering schemes in communication and storage devices. This article also addressed the characterization of e-spam which includes user information preferences and the involvement of spam seive in a wide and dynamic communication network. However, certain essential characteristics of spam filtering have not been taken into account in the study. This includes the configuration of the systems, the simulation environment and the analyses in terms of performance analysis

They research issues related to email spam were discussed by Sanz, Hidalgo, and Perez[10] how they affect consumers and what can be done between consumers and providers. It also mentions the legal, financial and technological acts used in e-mail spam communication. The researchers found out that the content processing filters are widely used based on technical measurements and thus have shown an acceptable proportion of precision and consistency. The study clarified how many machine-learning methods used to process email spam are structured and continued. Nevertheless, since it was written in 2008, this summary did not include current research publications in this category. There were also no comparative analyzes of different content filters

A brief report on the methods of screening spam images via email was submitted [11]. The research focused on e-mail filtration methods for the conversion of text to image-based approaches. Spam and spam filtering, which have been designed to limit the amount of innovation and innovations, have improved. The research did not, however, deal with computer processing methods, modeling software, dataset corpus, and e-mail spam filtering methods.

[12] Presented a comprehensive analysis of some of the common spam filtering approaches based on content. This article mainly focused on spam filtering algorithms. They examined core concepts, policies, efficiency, and trends of spam filtering. We explored the basics of spam filtering, the changing essence of spam, the Spammers ' Guide to Avoid Spam Filters from E-Mail Services Providers (ESPs), and specific approaches to combat the challenge of spam learning.

[13] A comprehensive analysis of E-mail Spam Filters used for the detection of irregularities has eliminated the need to identify e-mail spam and only operates with a single app group representation. This analysis demonstrates the first Spam Screening approach focused on an exception, an advanced tool, which used the data minimum technologies to decrease the transmission period while retaining the level of identification, and discusses how acceptable it is to pick spam-free email or spam as a display of normality.

3. PROPOSED ALGORITHM

The whole work builds on the selection, modeling, implementation, and optimization of apps. The selection of functions was the key factor in model creation where a new version based on PSO and RWN was produced.

The E-mail Feature Extraction Tool (EMFET 4) is used to convert e-mail messages into functionality that can be processed using a 212-based machine learning algorithm. Our team has planned the EMFET 213 as an open-source platform that's available to the public. 140 214 functionalities were derived by using EMFIT. These features can be separated into 49 header elements, 2 215 attachments, and 89 payload functions.

4. RESEARCH METHODOLOGY

Spam filtering seeks to reduce the number of unsolicited emails to the lowest possible amount. E-mail filtering is the retrieval of e-mails that are rearranged to a certain degree. Mail filters typically are used to manage incoming messages, spam-filtering, malicious code, trojans and malware-identification and deleting e-mails. Each important protocol, such as the SMTP, influences how the email functions. The most commonly used Email User Agents (MUA) are Mutt, Elm, Eudora, Microsoft Outlook, Pine, Mozilla Thunderbird, IBM fax, K-mail, and Balsa. There are e-mail users that can help e-mail users. Spam filtering can be set up in key positions of both consumers and servers. [14].

This work is based on feature selection, modeling, evaluation, and assessment. Throughout model growth, feature selection played a significant role in creating a new PSO-and RWN-based model.

In 1995 PSO algorithm was suggested by Kennedy and Eberhart, a metaheuristic algorithm suitable for optimizing nonlinear continuous functions. The algorithm was influenced by the idea of swarm intelligence, which is often seen in groups of animals including flocks and shoals.

A description of the actions of a flock is discussed to illustrate how the PSO motivated the creation of the

Volume 9, Issue 4
August 2020

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

optimization algorithm to overcome complex mathematical problems. A swarm of birds flying over a place must find a landing point and in this situation, it is a complicated problem to determine which point the entire swarm is expected to land, as it is related to many problems that are to maximize the availability of food and minimize the risk of predator activity. Here, the movement of birds can be interpreted as a choreographer; the birds move synchronously for a certain time to the best position and the whole flock at once [15].

The goal of segmentation problems is to create an X= vector variable [x1x2x3... xn] which, in keeping with the suggested optimization, reduced or greatly increased the formulation of the function F(X). The variable vector X is called vector position; this vector describes a variable model and is a vector dimension n where n describes the number of variables estimated in an issue, that is to say, the latitude and the longitude of the issue of the point to land determination by a flock. The f(X) function is also called a Fitness or Objective function, a function that can determine whether a position X was good or poor, i.e. whether the bird feels it is good a given landing point after the animal finds it and, in this case, this assessment is performed using several parameters for survival.

There is a positioning vector when considering a swarm of P particles$X_i^t = (\mathrm{x_{i1}x_{i2}x_{i3}} \ldots . \mathrm{x_{in}})^{\mathrm{T}})^{\mathrm{T}}$ and a velocity vector $V_i^t = (\mathrm{v_{i1}v_{i2}v_{i3}} \ldots . \mathrm{v_{in}})^{\mathrm{T}})^{\mathrm{T}}$ At each of the i particles that form it at iteration. These vectors are modified according to the following equations by dimension j:

$$V_{\mathrm{ij}}^{\mathrm{t+1}} = wV_{\mathrm{ij}}^{\mathrm{t}} + \mathrm{c_1}r_{\mathrm{i}}^{\mathrm{t}}(pbestij - X_{\mathrm{ij}}^{\mathrm{t}}) + \mathrm{c_2}r_2^{\mathrm{t}}(gbestj - X_{\mathrm{ij}}^{\mathrm{t}}) \quad (1)$$

And $$X_{\mathrm{ij}}^{\mathrm{t+1}} = X_{\mathrm{ij}}^{\mathrm{t}} + V_{\mathrm{ij}}^{\mathrm{t+1}} \quad (2)$$

where i = 1,2,…,P and j = 1,2,…,n.

Equality. (1) Implies that in the movement of a particle there are three separate participants in iteration, hence there are three words to be explored further. In the interim, Eq. (2) changes the position of the particle. The variable w is the continuous weight of inertia and is a positive constant value for the classical PSO version. This attribute is difficult to reconcile global search (where lower values are specified), often called a manipulation. This is known as tracking (if higher values are set) and local search. Concerning the parameter, one should note that this is one of the principal differences between classical PSO versions and other PSO versions.

The training method of 176 single layer feedforward neural networks (SLFNs) [16] has been initially developed in 1992 by RWNs. The goal of the RWN is not only to widespread the SLFN but also to feed multi-179 networks in which a node can be a sub-with extra unnecessary node elements. The learning speed of the RWN 181 is incredibly high and increases the generalization performance compared to 180 typical step-down methods used mainly for SLFN teaching. In comparison, unlike typical 182 techniques like backpropagation, which have to manually set parameters (learning rate, 183 epochs, etc.), the RWN doesn't require an intensive human input).

This dataset used in this is CSDMC2010 SPAM corpus. This data collection includes a range of e-mails as data for training and testing. Due to this dataset was used for a competition; it doesn't label the testing data but only training data. The below flow diagram shows the research methodology of the proposed model.

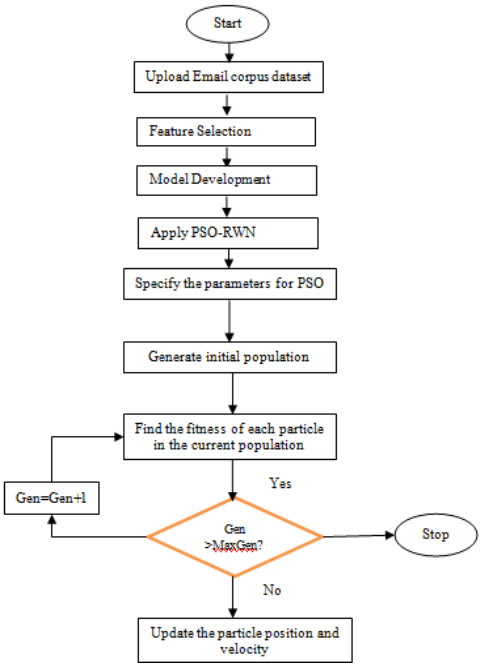

Fig. 2. Proposed Flow Diagram

5. RESULT AND DISCUSSION

The research methodology opted for this work is implemented in MATLAB 2018. For scientific computing, it is a highly skilled language. This incorporates calculation, visualization, and programming into a simple use environment in which common mathematical notation describes problems and solutions. The below figures illustrate the result of both algorithms and the number of features selected through an existing and proposed algorithm.

Fig. 3. Selected features based on the GA_RWN

Volume 9, Issue 4
August 2020

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X


```
Command Window
  Selected Features
  ------------------------------------
  1.    MinCharacterDiversitySubject
  2.    RatioDigitToAllSubject
  3.    RatioUpperLowerSubject
  4.    CountCapWordSubject
  5.    SingleLetterWordsSubject
  6.    TextPlainUnique
  7.    FromYahoo
  8.    ReplayGoogle
  9.    ToHotmail
  10.    ToMIL
  11.    ReplayGOV
  12.    VocabularyRichness
  13.    SMOG-I
  14.    ComplexWords
  15.    TFISFWithoutStopwords
  16.    MaxLongestRepeatingCharacter
  17.    MinCharacterDiversity
  18.    UniqueHTMLAnchor
  19.    RatioDigitToAll
  20.    CountofParagraphs
  21.    HTMLAnchor
  22.    HTMLNotAnchor
  23.    SenBegUpper
  24.    CountUpperChars
  25.    SingleQuote
  ------------------------------------
```

Fig. 4. Selected features based on the PSO_RWN

```
  9.    ToGoogle
  10.    ToMIL
  11.    ReplayAOL
  12.    WordLengthWithoutStopwords
  13.    SMOG-I
  14.    ComplexWords
  15.    TFISFWithoutStopwords
  16.    MaxLongestRepeatingCharacter
  17.    MinCharacterDiversity
  18.    RatioDigitToAll
  19.    CountofParagraphs
  20.    FunctionWords
  21.    HTMLAnchor
  22.    HTMLNotAnchor
  23.    SenBegUpper
  24.    CountUpperChars
  25.    CountWord
  ------------------------------------
  Percentage Correct Classification   : 80.522845%
  Percentage Incorrect Classification : 19.477155%
  Mean Square Error : 0.149791
  Accuracy = 0.805228 %
  Precision Spam = 90.776038 %
  Precision Ham = 57.492355 %
  Recall Spam = 82.748992 %
  Recall Ham = 73.509286 %
  G-Mean = 0.779924 %
>>
```

Fig. 5. Result parameters of GA-RWN

```
  9.    ToHotmail
  10.    ToMIL
  11.    ReplayGOV
  12.    VocabularyRichness
  13.    SMOG-I
  14.    ComplexWords
  15.    TFISFWithoutStopwords
  16.    MaxLongestRepeatingCharacter
  17.    MinCharacterDiversity
  18.    UniqueHTMLAnchor
  19.    RatioDigitToAll
  20.    CountofParagraphs
  21.    HTMLAnchor
  22.    HTMLNotAnchor
  23.    SenBegUpper
  24.    CountUpperChars
  25.    SingleQuote
  ------------------------------------
  Percentage Correct Classification   : 83.937824%
  Percentage Incorrect Classification : 16.062176%
  Mean Square Error : 0.130827
  Accuracy = 0.839378 %
  Precision Spam = 92.852280 %
  Precision Ham = 63.914373 %
  Recall Spam = 85.250000 %
  Recall Ham = 79.923518 %
  G-Mean = 0.825438 %
>>
```

Fig. 6. Result parameters of the proposed algorithm

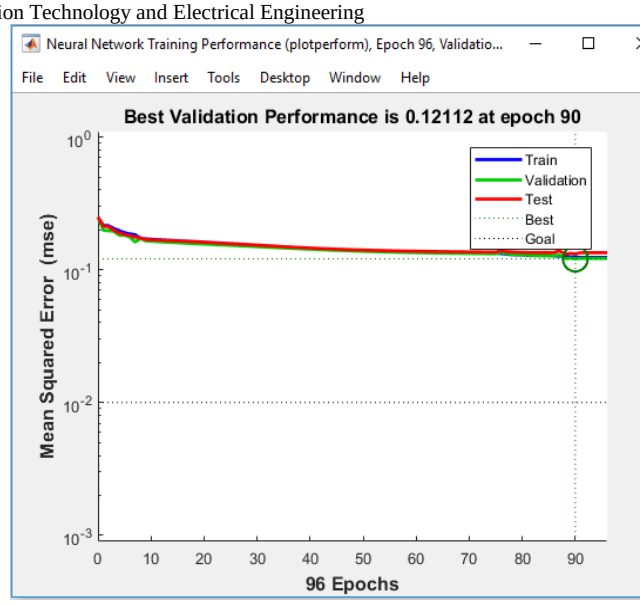

Fig. 7. Best Validation performance at the iteration 90

TABLE I
COMPARISON OF BOTH RESEARCH METHODOLOGY

Parameters/Algorithms	GA-RWN	PSO-RWN
Accuracy	80%	83%
Precision ham	57.49%	63.91%
Precision spam	90.77%	92.85%
Recall ham	73.50%	79.92%
Recall spam	82.74%	85.25%

6. CONCLUSION

We explored machine learning and its implementation in the spam filtration area in this study. Here we have applied hybridization of optimization algorithm and Neural Network, Particle Swarm Optimization (PSO) and (Random Weight Network). The shortcomings of the machine learning algorithms were pointed out and comparative analyzes of the computer training approaches in literature were performed in the effective management of danger of spam. The literature statistic and quantity we studied suggests a substantial improvement in this field in general. In general, significant progress must still be made.

In the future, we will try to work on any other email spam filtering using the same approach, and also we will try to implement supervised and unsupervised machine learning algorithms over the same dataset.

REFERENCES

[1] [1] Yanti Rosmunie Bujang, Husnayati Hussin "Should We Be Concerned with Spam Emails? A Look at Its Impacts and Implications", presented at the 5th International Conference on Information and Communication Technology, IEEE 2013, p 01.

[2] Izabella Miszalska, Wojciech Zabierowski, Andrzej Napieralski, "Selected Methods of Spam Filtering in Email", CADSM'2007, February 20-24, 2007, Polyana, UKRAINE,p 02.

[3] Liu Ming, Li Yunchun, Li Wei "Spam Filtering by Stages" presented at the International Conference on Convergence Information Technology, IEEE 2007, pp 01-04.

[4] "Spam email percentage in the mailbox" McAfee Managed Mail Protection Always On, Automatic Mail Protection. Available: Website:http://www.mcafee.com/us/resources/misc/web-protectioninfographic.pdf [Accessed: Feb 5, 2014].

[5] Ayad Al-Hajj, "Cyber Crimes: Threats and Protection", presented at the International Conference on Networking and Information Technology, IEEE 2010, pp 01-02.

[6] Cynthia Dhinakaran and Jae Kwang lee, Dhinaharan Nagamalai, "Reminder: please update your details": Phishing Trends", presented at the First International Conference on Networks & Communications, IEEE 2009, p 01.

[7] Asoke K. Talukder, Vedula Bhaskar Rao, Vaibhav Kapoor, and Devashish Sharma," Artificial Hygiene: A Critical Step towards Safety from Email Viruses", presented at the IEEE India Annual Conference 2004. Indian, IEEE 2004, p 01.

[8] Robin Gandhi, Anup Sharma, William Mahoney, William Sousan, Qiuming Zhu, And Phillip Laplante, " Dimensions of Cyber Attack", IEEE, IEEE Technology And Society Magazine 1932- 4529/11/$26.00©2011IEEE Magazine, 2011.

[9] *C.P. Lueg, From spam filtering to information retrieval and back: seeking conceptual foundations for spam filtering, Proc. Assoc. Inf. Sci. Technol. 42 (1) (2005)*.

[10] X.L. Wang, Learning to classify email: a survey, in 2005 International Conference on Machine Learning and Cybernetics (Vol. 9, pp. 5716-5719), IEEE, Aug 2005.

[11] W. Li, N. Zhong, Y. Yao, J. Liu, C. Liu, Spam filtering and email-mediated applications, in: Paper presented at the International Workshop on Web Intelligence Meets Brain Informatics, 2006.

[12] G.V. Cormack, Email spam filtering: a systematic review, Found. Trends Inf. Retr. 1 (4) (2008) 335–455. *E.P. Sanz, J.M.G. Hidalgo, J.C.C. Perez, Email spam filtering, Adv. Comput. 74 (2008) 45–114*.

[13] S. Dhanaraj, V. Karthikeyan, A study on e-mail image spam filtering techniques, in: Paper presented at the International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME), 2013.

[14] A. Bhowmick, S.M. Hazarika, Machine Learning for E-Mail Spam Filtering: Review, Techniques and Trends, arXiv:1606.01042v1 [cs.LG] 3 Jun 2016, 2016, pp. 1–27.

[15] C. Lauren, X. Ugarte-Pedrero, I. Santos, B. Sanz, J. Nieves, P.G. Bringas, Study on the effectiveness of anomaly detection for spam filtering, Inf. Sci. 277 (2014) 421–444

AUTHOR PROFILES

Ajinkya Gulunjkar has received B.tech Degree in stream of Computer Science from Government Engineering College Ajmer, Rajasthan, India and is currently pursuing M.Tech Degree in Information Technology from Government Engineering College Ajmer, Rajasthan, India.

Dr. Rakesh Rathi, is currently working as Head of Department of Computer Science and Information Technology of Government Engineering College Ajmer, Rajasthan, India.