Email Image Spam Detection Using Fast Support Vector Machine and Fast Convergence Particle Swarm Optimization

T. Kumaresan, P. Subramanian, D. Stalin Alex, M.I. Thariq Hussan, B. Stalin

Abstract: ***Today's internet scenario, spam email is a major problem of internet users. The spam field has two different types namely email text spam and image spam. Now a day's email spam filters are available in market, but the filters are capable to detect the text based spam only. Spammers are using intelligent ways to bypass the spam filters like embedding the spam text in an image so that spam filters are not able to detect the image spam. This paper analyses the various attributes of image spam with the careful attention given with the existing system. The proposed method uses the fast convergence particle swarm optimization technique which uses the diversity location of each particle by presenting a new classifier. Experimental results show that proposed method has achieved better accuracy than the other existing methods.***

Index Terms: Image spam detection, Email spam, Support Vector Machine (SVM), Standard Particle Swarm Optimization (PSO), Fast Convergence Particle Swarm Optimization (FCPSO).

I. INTRODUCTION

Email spam is exiting from the past few years to promote some services or product. Most of the spam mails are being sent to various internet users by spammers to market their products [1].

Although there are so many filters are available to detect the spam still spammers are successfully sending their mails to many users. One of the main reasons for this includes spammers continuously changing their way of sending mails [2]. Email spam can be detected by using some unique features for example if a mail have some words like "work from home" "earn money without investment" are probably a spam mail. A good classifier can detect these spam mails by using these words as features. Spam classification accuracy is based on many factors like selecting best features set and classifiers [3].

Since the text based spam is easy to detect by using the features listed above spammers have changed their way of sending spam text through images [4]. While spam text is being sent through an image then the filters assumes these as original mail [5]. To avoid these kinds of image spam there is a need of good classifier as well as optimization algorithms [6]. The need of optimization algorithm is to select the optimal feature set [7]. Since the classification accuracy is based on best features only.

This paper proposed a classification technique by using support vector machine [8] as well as fast convergence particle swarm optimization. Support vector machine is known to be good classifier particularly for two class problems [9]. The main drawback in particle swarm optimization is its convergence speed [10].

The speed of the convergence is increased in PSO by identifying the diversity location of each particle available in the set [11]. This fast convergence particle swarm optimization identifies the best feature [12] set and then these features are given to the support vector machine classifier for further classification. Ling spam [1] and spam archive [3] data sets have been taken for experimental analysis.

II. PROPOSED FEATURE EXTRACTION AND CLASSIFICATION

The mails which are sent to huge number of peoples without their interest is often referred as spam mail. An Image in a mail with spam content is called as image spam. These image spams are very difficult to detect by text based spam filters. The existing spam filters are assumes these images as legitimate pictures sharing between internet users. Most of the email messages can be categorized in the following two categories.

1. An email has text and image which is not a spam.
2. An email has image with embedded spam text.

The first category often called as ham and the second one is spam.

A. Support Vector Machine

Support vector machine [6] is a well-known classifier for two class problems. It is based on statistics learning. It will identify the optimal hyper plane with maximum margin to divide the classes.

Maximize

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(X_i, X_j)$$

Subject to

$$\sum_{i=1}^{n} \alpha_i y_i = 0\,, \quad 0 \le \alpha_i \le C, i = 1,2,\dots,n \qquad (1)$$

Most widely used RBF function has been used here

Revised Manuscript Received on May 05, 2019.

Dr.T.Kumaresan, Professor, Dept. of CSE, Sri Indu College of Engineering and Technology, Hyderabad, Telangana, India.
Dr.P.Subramanian, Professor, Dept. of CSE, Sri Indu College of Engineering and Technology, Hyderabad, Telangana, India.
Dr.D.Stalin Alex, Professor, Dept. of IT, Guru Nanak Institute of Technology, Hyderabad, Telangana, India.
Dr.M.I.Thariq Hussan, Professor and Head, Dept. of IT, Guru Nanak Institutions Technical Campus, Hyderabad, Telangana, India.
Dr.B.Stalin, Assistant Professor, Department of Mechanical Engineering, Anna University, Regional Campus Madurai, Madurai, Tamilnadu, India. (E-mail: stalin1312@gmail.com)

$$K(X, X_j) = \exp\left(-\gamma \left\| X_i - X_j \right\|^2, \gamma > 0 \right) \quad (2)$$

A test model Y is classified by the following formula

$$y = sign\left(\sum_{i=1}^{n} \alpha_i y_i K(X_i, X)\right),$$
$$sign(a) = \begin{cases} +1, & if\, a > 0 \\ -1, & otherwise \end{cases} \quad (3)$$

B. Fast Convergence Particle Swarm Optimization

Fast Convergence Particle Swarm Optimization [13] plays a vital role in feature selection. Since the classification accuracy is based on best features set only. The main draw with the particle swarm optimization is that its limitations with dimensions. It is good when there are low dimension problems. If there are high dimension problems it will not perform well. The fast convergence particle swarm optimization is the solution for the high dimension problems. It identifies the diversity location of each particle so that the convergence speed is improved.

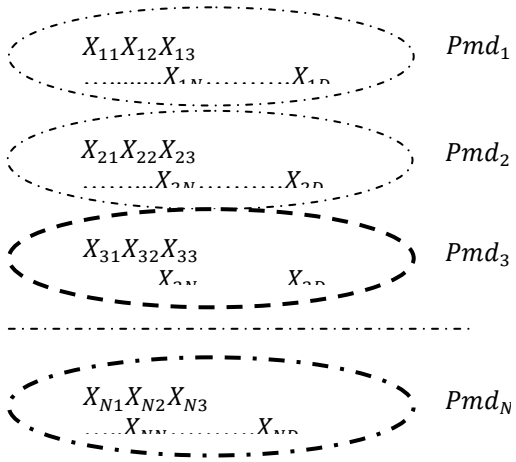

Fig. 1: Mean Dimension Evaluation

$$Pmd_i = (x_{i1} + x_{i2} + \cdots + x_{iD})/D \quad (4)$$

C. Feature Analysis

Feature selection plays a vital role in spam email classification. Here the basic features along with some unique features of images have been taken for the analysis.

Table 1: Image Features

Feature	Description
f1	The image width in header
f2	The image height in header
f3	Aspect ratio
f4	GIF image
f8	Anisotropy
f9	Edge frequency
f10	Contrast

The feature vector f1 indicates the image width in header and f2 denotes the image height in the header, f3 denotes the aspect ratio and f4 denotes the GIF image, f8, f9 and f10 indicates the texture feature of Anisotropy, Edge frequency and Contrast.

III. EXPERIMENTAL RESULTS

The experiment is done by using MATLAB. This work includes maximum kinds of images those are grouped arbitrarily from spam archive data set provided by Fumera *et al* [14]. This spam archive data set contains combination of personal ham and spam images. In total, the images considered to this proposed work is about 5089 images combined of 3208 spam and 1881 ham images, which consist of JPEG, GIF, PNG and BMP images.

The following three metrics are used in this work.

$$Accuracy(A) = \frac{TP + TN}{TP + TN + FN + FP}$$

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

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

Where TP is true positive TN is true negative FP is false positive and FN is false negative.

The fast convergence particle swarm optimization first extracts all the possible features from the image and then it identifies the optimal feature vectors. Finally based on the optimum features classification is done by support vector machine. For result comparisons this work has been implemented by using one more classifier namely K-nearest neighbor. Following table shows the accuracy and execution time obtained using the proposed technique. From the results it is clear that the proposed technique has outperformed well.

Table 2: Accuracy and Execution Time for Proposed Fast SVM

Approach	Accuracy (%)			Execution Time (Seconds)		
	k-NN	SVM	Proposed Fast SVM	k-NN	SVM	Proposed Fast SVM
File properties	73.6%	79%	83.6%	59	42	32
File feature f1-f6	78.1%	82.5%	89.4%	50	38	34
Image feature f7-f10	82.4%	85.9%	88.5%	53	40	28
All features	88.1%	89.5%	92.4%	56	46	25

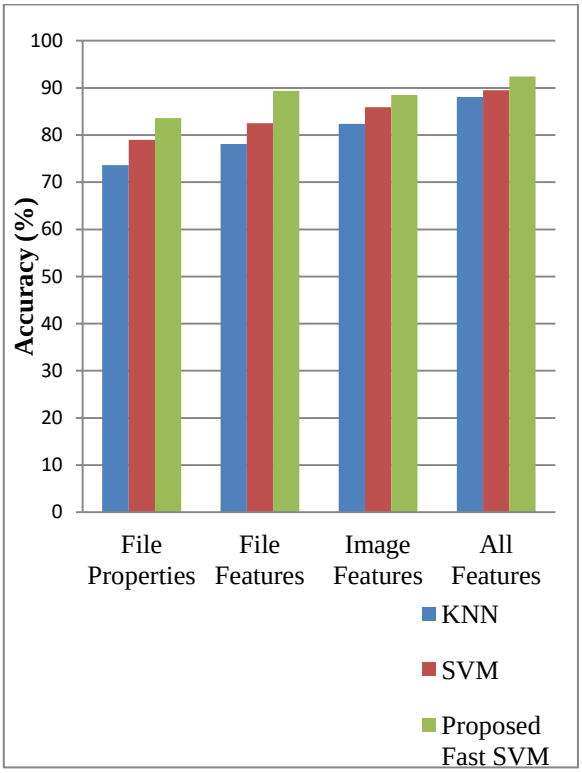

Fig. 2: Accuracy for Proposed Fast SVM

Figure 2 and 3 shows the accuracy and execution time obtained from proposed method of Fast SVM. From the figure it is clear that it gives better accuracy with minimum execution time.

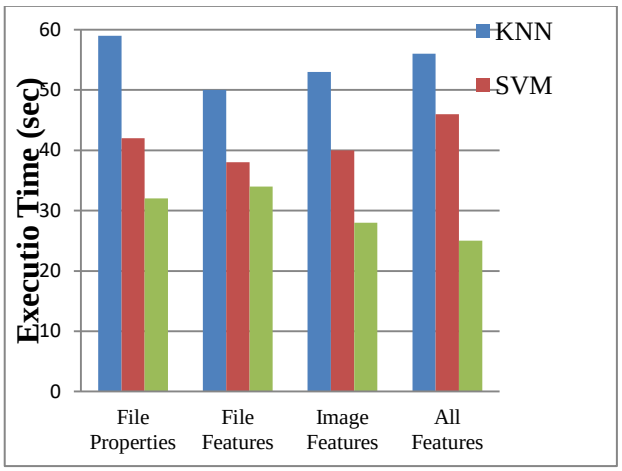

Fig. 3: Execution Time for Proposed Fast SVM

IV. CONCLUSION

Email spam is one of the important treat to the internet users. Though there are so many spam filters still most of the mail boxes are being filled with junk mails. To detect the email spam we have proposed an efficient method using fast convergence particle swarm optimization and support vector machine. Experimental result shows that the proposed method achieved the better accuracy with minimum execution time.

ACKNOWLEDGMENT

We thank the management of Sri Indu College of Engineering and Technology, Guru Nanak Institute of Technology and Guru Nanak Institutions Technical Campus, Hyderabad for providing us enough resources to prepare and experiment our work. We are thankful to our faculty members and staffs for their support.

REFERENCES

1. T. Kumaresan and C. Palanisamy (2017), E-mail spam classification using S-Cuckoo search and support vector machine. *International Journal of Bio-Inspired Computation* 9(3), pp. 142-156.
2. T. Kumaresan, S. Sanjushree and C. Palanisamy (2014), Image spam detection using color features and K-Nearest neighbor classification, *International Journal of Computer, Information, Systems and Control Engineering 8(10)*, pp. 1746-1749.
3. T. Kumaresan, S. Saravanakumar and R. Balamurugan (2017), Visual and Textual Features Based Email Spam Classification Using S-Cuckoo Search and Hybrid Kernel Support Vector Machine. *Cluster Computing*, Springer, DOI : https://doi.org/10.1007/s10586-017-1615-8
4. Hayati, Pedram, and Vidyasagar Potdar, "Evaluation of spam detection and prevention frameworks for email and image spam': a state of art", In Proceedings of the 10th International Conference on Information Integration and Web-based Applications & Services, 2008, pp. 520-527.
5. Mehta, Bhaskar, Saurabh Nangia, Manish Gupta, and Wolfgang Nejdl, "Detecting image spam using visual features and near duplicate detection", In Proceedings of the 17th international conference on World Wide Web, 2008, pp. 497-506.
6. Vapnik, Vladimir Naumovich, and Vlamimir Vapnik, *Statistical learning theory*, New York: Wiley, Vol. 2, 1998.
7. Hsia, Jen-Hao, and Ming-Syan Chen, "Language-model-based detection cascade for efficient classification of image-based spam e-mail", In Multimedia and Expo, 2009. ICME 2009. IEEE International Conference, 2009, pp. 1182-1185.
8. Attar, Abdolrahman, Reza Moradi Rad, and Reza Ebrahimi Atani (2013), A survey of image spamming and filtering techniques. *Artificial Intelligence Review 40(1)*, pp. 71-105.
9. Zhong, Jian, YiLu Zhou, and Wei Deng (2013), Filtering image-based spam using multifractal analysis and active learning feedback-driven semi-supervised support vector machine. *In Conference Anthology, IEEE*, pp. 1-5.
10. Caruana, Godwin, Maozhen Li, and Yang Liu (2013), An ontology enhanced parallel SVM for scalable spam filter training. *Neurocomputing 108*, pp. 45-57.
11. Gleeson, Matt, David Hoogstrate, Sandy Jensen, Eli Mantel, Art Medlar, and Ken Schneider (2004). 'Method and apparatus for filtering email spam based on similarity measures', U.S. Patent Application 10/846,723.
12. Zhou, Bing, Yiyu Yao, and Jigang Luo 2010. 'A three-way decision approach to email spam filtering', In Advances in Artificial Intelligence, pp. 28-39. Springer Berlin Heidelberg.
13. Sahu, Amaresh, Sushanta Kumar Panigrahi, and Sabyasachi Pattnaik (2012), Fast Convergence Particle Swarm Optimization for Functions Optimization. *Procedia Technology 4*, pp. 319-324.
14. G. Fumera, I. Pillai, and F. Roli (2006), Spam Filtering based on the Analysis of Text Information Embedded into Images. *Journal of Machine Learning Research (special issue on Machine Learning in Computer Security) 7*, pp. 2699-270.
15. Sankar, S.P., Vishwanath, N., Lang, H.J., and Karthick, S. "An effective content based medical image retrieval by using abc based artificial neural network (ANN)", Current Medical Imaging Reviews, vol. 13, no. 3, pp. 223-230, 2017. DOI: 10.2174/1573405612666160617082639
16. Arul Teen, Y.P., Nathiyaa, T., Rajesh, K.B., and Karthick, S. "Bessel Gaussian Beam Propagation through Turbulence in Free Space Optical Communication", Optical Memory and Neural Networks (Information Optics), vol. 27, no. 2, pp. 81-88, 2018. DOI: 10.3103/S1060992X18020029
17. Sathish, T. "Performance measurement on extracted bio-diesel from waste plastic", Journal of Applied Fluid Mechanics, vol. 10, pp. 41-50, 2017.
18. Sathish, T., Jayaprakash, J. "Optimizing Supply Chain in

Reverse Logistics", International Journal of Mechanical and Production Engineering Research and Development, Vol. 07, pp. 551-560, 2017.
19. Sathish, T., Periyasamy, P. "Modelling of HCHS system for optimal E-O-L Combination section and Disassembly in Reverse Logistics", Applied Mathematics and Information science, Vol. 13, No. 01, pp. 1-6, 2019.
20. Sathish, T., Muthulakshmanan, A. "Design and simulation of connecting rods with several test cases using AL alloys and high Tensile steel", International Journal of Mechanical and Production Engineering Research and Development, Vol. 08,Issue 1, pp. 1119-1126, 2018.
21. Sathish, T., and Karthick, S. "HAIWF-based fault detection and classification for industrial machine condition monitoring", Progress in Industrial Ecology, vol. 12, no. 1-2, pp. 46-58, 2018

AUTHORS PROFILE

Dr. T. Kumaresan, as a professor of CSE department in Sri Indu College of Engineering and Technology, Hyderabad. He has completed his Ph.D in Anna University Chennai. He has more than 11 years of experience. His areas of interest include data mining, image processing, wireless networks and cloud computing.

Dr. P. Subramanian working as a professor of CSE department in Sri Indu College of Engineering and Technology, Hyderabad. He did his Ph. D in St. Peter's University, Chennai. He has more than 15 years of experience. His area of interest includes wireless networks, Data mining and Image processing.

Dr.D.Stalin Alex, Professor and Head of Department of Information Technology, Guru Nanak Institute of Technology, Hyderabad, India. He received his PhD in Image Processing from Anna University, Chennai, India. He is a member of ISTE. He has published 9 papers in SCOPUS Indexed journals. He filed a patent in the area of Image Processing. Area of Interest: Image Processing, Video Processing.

Dr.M.I.Thariq Hussan is working as a Professor and Head, Department of Information Technology in Guru Nanak Institutions Technical Campus, Hyderabad, India. He was awarded his Doctoral Degree in the Faculty of Information and Communication Engineering from Anna University, Chennai, Tamilnadu. He received 'Best Teacher Award-2018' from the Institute for Exploring Advances in Engineering accredited by EA-JAS. He is having 2 years of industry and 17 years of teaching experience. His research areas include Sequential Pattern Mining and Mobile Computing.

Dr.B.Stalin received B.E. Degree in Mechanical Engineering from the University of Madras, Tamilnadu, India and M.E. Degree in Manufacturing Engineering from the Anna University, Tamilnadu, India. He obtained his Ph.D. in Mechanical Engineering discipline at Anna University, Chennai, Tamilnadu, India. His current research interests include Materials Characterization, Mechanical Properties, Composite Materials, Optimization Techniques, and Manufacturing Engineering.