International Journal of Computer Theory and Engineering, Vol.3, No.1, February, 2011
1793-8201

A Novel Approach towards Image Spam Classification

M.Soranamageswari, Dr.C.Meena

Abstract—**The volume of unsolicited commercial mails has grown extremely in the past few years because of increased internet users. This unsolicited mails termed as spam occupy large storage space and bandwidth. Therefore designing an efficient spam filter is a challenging issue ahead for the future generation. Here we use gradient histogram is a key feature that can be exploited to improve the categorization capability. The gradient values are valuated for each pixel of an image. These obtained features are then normalized for efficient spam classification. These normalized features are then applied as input for feed forward back propagation neural network (BPNN) model. The experiments are conducted for different training/testing rule for the BPNN. The performance measure in terms of accuracy is determined for the proposed method using various training rule of BPNN.**

Index Terms—**Feature Extraction, Feed Forward Back Propagation Neural Networks, Gradient Histogram, and Training/Testing rule.**

I. Introduction

Direct Marketers are very much attracted by the communication using internet because of increasing number of internet users and the low cost of e-mail. As a result the volume of unsolicited commercial mails has grown enormously in the past few years. These unsolicited bulk emails are coined by the term "Spam". Spam email with advertisement text embedded in images generally known as image spam, which poses a great challenge to Anti-spam filters in detecting these spam emails [15]. E-mail management has become a vital and growing problem for individuals and organizations as it is prone to misuse. The invasion of image spam into email has created problem for spam classifiers.

In the past, spam filtering required the manual construction of pattern matching rule sets [6]. Contemporary spam filtering program indulgence spam detection as a text classification problem utilizing machine-learning algorithms such as neural networks and naive Bayesian classifiers to learn spam characteristics. Among these, Bayesian-based pproaches [4, 14] have accomplished outstanding accuracy and have been widely used [20]. OCR-based modules can be used against image spam, to tolerate the analysis of the semantic content embedded into images. The performance of the OCR-based technique in detecting image spam is explained in [9]. The main limitation of this OCR-based spam classification technique is that it requires more processing time.

Filters are available to combat these unsolicited annoyances. But spammers continually develop new techniques to avoid detection by filters. Spammer use different randomization techniques to embed the text in an image which probably fool the spam detectors. Current and comprehensive list of spam techniques are available in [10]. A number of image spam identification and classification techniques have been proposed [1, 19, 7] including image processing and computer vision techniques [17, 9]. Spam filters are realistically effective now, but are slowly becoming less effective with innovative forms of image spam and audio-based spam which will in time prove to be a confront for today's most widely used Bayesian filters. Spammers have tailored again and again to effective adversaries and countermeasures. This is likely to continue in the future. Therefore it is necessary to develop a spam classification technique that results in low false positive rate for the future generation.

This paper discusses an novel image spam classification technique. In this technique the features of an image is extracted using histogram. In our proposed method gradient histogram is extracted from an image. The obtained feature point of an image is then processed using Artificial Neural Network (ANN). In particular, this paper utilize feed forward back propagation neural network for classifying the image spam from those of legitimate mail ("ham"). The experiment is conducted on combination of spam and ham images obtained from Spam Archive Data set to prove the accuracy of the proposed method in image spam classification.

The remainder of this paper is organized as follows. Section II provides an overview of relevant research in image spam classification using feature extracted from histogram. Section III describes our approach of image feature extraction using histogram for efficient image spam classification. Section IV provides details of the experiments and performance of our approach in classifying the image spam. Section V concludes the paper with a critical discussion.

II. Background Study

Many discussions have been carried out previously on image spam detection. This section of paper provides an overview of relevant research work in image spam classification.

Aradhye et al. proposed an efficient image spam

M. Soranamageswari is with LRG Govt Arts College for Women, Tirupur. (email : swarnakannappan@rediff.com).

Dr. C. Meena is with the Avinashilingam University for Women in India. (E-mail: meena_cc@avinuty.ac.in;

classification technique in [1]. The method distinguishes spam images from other general categories of e-mail images based on extracted overlay text and color features. The method is fast and efficient for categorization of spam e-mails containing imagery (or links to images) for the purposes of filtering or categorizing the communication. In addition, the method can also be used for monitoring of outbound e-mails by corporations to detect communications including proprietary or company confidential material. The method comprises of three steps. The text regions in the image are first extracted. The consequent step defines a small number of consistent spam-indicative features from the image, using in part the extracted text regions. Subsequently, support vector learning is applied to make a spam–non-spam decision for each image. The method proposed in [1] for spam image categorization detects vertically oriented edge transitions and connected components of similar intensity in a grayscale image, and links those that are attuned in size and relative position to form lines of text.

A multi-modal framework was put forth by Zhang et al. in [21] for revealing common sources of spam images. A multimodal framework put forth in [21] clusters spam images so that ones from the same spam source/cluster are grouped together. For this reason, text recognition and visual feature extraction are performed. The text content and foreground illustrations embedded in image spam co-operate a key role in identifying the connection between spam images. This multi-modal framework reveals the origin of the spam images using the following three steps. First step being Image segmentation, followed by feature extraction and similarity calculation and the final step is spam image clustering. The color-code histogram of foreground illustration areas is built to describe the color composition of foreground illustrations in an image.

Octet histogram spam mail filtering technique was described by Chun-Chao Yeh and Nai-Wei Yeh in [5]. They developed a near-duplicated mail detection scheme for spam filtering. The detection scheme is based on octet histogram of mail context. The higher dimension the feature vector is the more precision can be achieved with the higher cost of computation. A feature vector resulting from the octet histogram is related with the mail to represent the mail. Then, a similarity comparison is made between the new mail and a database consisting of possible candidates, based on the feature vectors. They considered different strategies to reduce the feature dimension. They focused on the strategy based on duplicate mail detection. Each type of strategies, either based on content-information in a mail or similarity between mails, has its weakness. Since both types of strategies take advantages of different spam properties, they can be a complement to each other.

Bowling et al. in [12] discussed an approach using artificial neural networks for image spam classification. They proposed a method for identifying image spam by training an artificial neural network. The process is accomplished in three steps. First step is image preparation. In the second step they trained the artificial neural network with their training data. Finally, they made a test on the network, which contains "unknown" images to see how well it has learned to identify spam versus ham. They created an Artificial Neural Network (ANN) with 22,500 inputs, two hidden layers of 50 or 75 nodes each, and one output node. The input nodes are the pixels of an image. The output layer is the -1 or 1 indicating ham or spam.

An approach based on low level image processing technique was presented by Biggio et al. in [2]. They proposed an approach based on low-level image processing techniques to detect one of the main characteristics of most image spam, namely the use of content obscuring techniques to defeat OCR tools. Their approach consists in developing techniques to perceive the presence of noisy text into an image. Such techniques are projected to be used by a specific module of a spam filter, whose output could be either a crunchy label indicating the presence or absence of noisy text, or a real number indicating the "amount" of noise in a proper scale. They discussed a method to detect such kind of noise and to measure its amount.

Fdez-Riverola et al. in [8] described an approach for spam classification. They showed how a previously successful instance-based reasoning e-mail filtering model can be improved in order to better track concept drift in spam domain. Their approach is based on the definition of two complementary techniques able to select both terms and e-mails representative of the current situation. The enhanced system is evaluated against other well-known successful lazy learning approaches in two scenarios, all within a cost-sensitive framework. Their experimental results revealed that the instance-based reasoning systems can offer a number of advantages tackling concept drift in dynamic problems, as in the case of the anti-spam filtering domain.

Ian Stuart et al. in [11] put forth an approach for that can distinguish legitimate e-mail from spam. In contrast to earlier approaches, this approach feature set uses descriptive characteristics of words and messages similar to those that a human reader would use to identify spam. This alternative approach using a neural network (NN) classifier on a corpus of e-mail messages from one user is tested for different image data set. The results of their work are compared to previous spam detectors that have used Naïve Bayesian classifiers.

An effective e-mail classifying and cleansing method was described by Bin Cui et al. in [3].Their method deals with both fields having pre-defined semantics as well as variable length free-text fields for obtaining higher accuracy. First, they presented a new model based on the Neural Network (NN) for classifying personal E-mails. Second, they proposed the use of Principal Component Analysis (PCA) as a preprocessor of Neural Networks to decrease the data in terms of both size as well as dimensionality so that the input data become more classifiable and faster for the convergence of the training process used in the NN model.

Thus this section provides a clear overview of the relevant research works carried out in literature for image spam classification using histogram technique and neural networks.

III. Feature Extraction for Image Spam Classification

Feature extraction plays critical role in image spam classification. Under this section of the paper, we explain the proposed methodology in extracting the feature points from an image and the feed forward back propagation neural network, which effectively pretends as a classifier for filtering the image spam.

A. Gradient Histogram

The gradient histogram of an image describes the orientation properties of the image. The technique counts occurrences of gradient orientation in localized portions of an image. This method is similar to that of edge orientation histograms, scale-invariant feature transform descriptors, and shape contexts, but differs in that it computes on a dense grid of uniformly spaced cells and uses overlapping local contrast normalization for improved performance. Gradient computation is the critical task in this histogram. Image pre-processing thus provides modest collision on performance [13]. Instead, the first step of calculation is the computation of the gradient values.

In each of the cells, local gradient histogram features are extracted. Denote L(x; y) the result of convolving the image I(x; y) with a smoothing filter, employed for denoising purposes. First, the horizontal and vertical gradient components Gx and Gy are determined as

$$G_x = L(x+1, y) - L(x-1, y)$$
$$G_y = L(x, y+1) - L(x, y-1)$$

Alternatively to the smoothing plus the gradient computation, a Gaussian derivative filter can be employed. In any case, the gradient magnitude 'm' and direction 'θ' are then obtained for each pixel with coordinates (x; y) as

$$m(x, y) = \sqrt{G_x^2 + G_y^2}$$
$$\theta(x, y) = \frac{1}{2}\tan^{-1}\left[\frac{G_y}{G_x}\right]$$

In the evaluation of the gradient value of the image the foremost step is to resize the image into size represented by NXN where N=256 in our case. Gaussian filter is used as smoothing filter. The orientation angle θ is considered to be 360. The gradient value of an image is calculated at each pixel with region of interest being 256X256. The gradient value of an image I (i, j) is evaluated at each pixel assuming the number of bins to be 5. These extracted gradient values are then used as input for feed forward back propagation neural networks algorithm. Before providing the obtained feature vectors as inputs to BPNN they are normalized such that the mean of the image is 0 and the standard deviation is 1.

B. BPNN Model

BPNN is a training model used for classification of image spam using the optimum feature vectors extorted from an image. Back propagation algorithm [16, 17] is one of the well-known algorithms in neural networks. It is a supervised learning technique for training the neural network. This recognizes different classes of data and test how well network has learned from the previous set of data. The network consists of one input layer with four neurons and two hidden layers with twenty neurons and one output layer with a single neuron. BPNN present a training sample to the neural network. It then compares the network's output to the desired output of that sample. Calculate the error in each output neuron. For each neuron, calculate what the output should have been, and a scaling factor, how much lower or higher the output must be adjusted to match the desired output. This is termed as local error.

BPNN adjusts the weights of each neuron to minimize the local error. Learning rate and momentum are the key attributes for the performance of the BPNN in classification of the spam messages. The learning rate and momentum used for this network is 0.9 and 0.9 respectively. In addition the maximum number of epochs is 200. The model is tested using the standard rule of 80/20, where 80% of the samples are randomly used for training and 20% is used for testing. Similarly the experiments are repeated using the rule 90/10, where 90% of the samples are randomly utilized for training and remaining 10% is used for testing. BPNN is also trained for 70/30, 60/40, and 50/50 environment. Though the BPNN algorithm performs gradient descent on the total error and the weights adjusts to minimize the local error thereby improving the performance of BPNN in classifying the spam messages.

IV. Experimental Results

A Gradient histogram based feature point extraction method for Image spam classification system is built using MATLAB. The experimental system developed, reports a new image spam classification method using Gradient histogram as a image feature.

A. Image Spam Dataset

Email corpora are difficult to construct due to the private nature of email communication. In many spam classification assessments, duplicate or highly analogous emails are included to imitate the real world nature of spam. To calculate the performance of the proposed approach we used a spam archive data set [9] partly. The Spam Archive images were taken from the Spam Archive data provided by Giorgio Fumera's group and used in this paper. This spam archive data set contains combination of personal image ham and personal image spam. In total, the images considered to this proposed work is with 5087 images combined of 3209 spam and 1878 ham images, which consist of JPEG, GIF, PNG and BMP images.

B. Performance Measures

We decided to validate our system by performing 5 tests for each category for training set and the corresponding testing set. The average of the results is considered for evaluating the performance measures. The number of correctly identified spam is termed as true positive, number of correctly identified ham denoted as true negative, number of spam images misidentified as ham is false negative and ham images misidentified as spam represents false positive. As false positives are generally considered to be more harmful than false negatives, the goal is to ensure that low false alarm rate is the first priority, while at the same time

minimizing the rate of false negatives as much as possible. We also evaluate the performance in terms of Accuracy (A), Precision (P) and Recall (R).

We have measured the processing system of our method on an Intel core 2 Duo machine and it is developed by using MATLAB. The proposed spam classification system validate on the aforementioned database. Here the personal ham and spam are combined together to measure the performance of proposed technique. Experiments show that our method is efficient in classifying image spam messages with feature extracted from gradient histogram based approach.

The extracted feature is normalized using Mean value for Standard deviation (i.e. mean for values of 0 and standard deviation for values 1) technique so that the classification accuracy is improved in an efficient manner. The average classification accuracy of around 93.7% was obtained on considering the rule of 90/10, where 90% of the samples are randomly utilized for training and remaining 10% is used for testing in a BPNN. Similarly the experiments are carried out for 80/20, 70/30, 60/40, 50/50 rule based training. Therefore the experimental results showed that our proposed system performs more effectively in the classification of the image spam. Table 1 shows the result of our experiment with training/testing ratio 90/10. In the similar manner, table 2, table 3, table 4, table 5 show the results with training/testing rationale 80/20, 70/30, 60/40 and 50/50 respectively. Table 6 represents the performance measures. Figure 1 shows the comparison of Accuracy for the different training rule of BPNN. Figure 2 shows the comparison of Precision and Recall for the gradient histogram based feature extraction for image spam classification, using BPNN for different training rules. Figure 1 show that the accuracy of our proposed approach with training rule 90/10 is better than the accuracy obtained using the same approach with other lower training/testing set in image spam classification.

TABLE 1.EXPERIMENTAL RESULT FOR TRAINING/TESTING RULE 90/10

Accuracy	Spam precision(SP)	Spam Recall(SR)
95.26	0.87	0.91
93.27	0.85	0.96
91.29	0.89	0.88
93.28	0.86	0.94
94.26	0.87	0.91

TABLE 2.EXPERIMENTAL RESULT FOR TRAINING/TESTING RULE 80/20

Accuracy	Spam precision(SP)	Spam Recall(SR)
92.41	0.90	0.93
89.35	0.82	0.89
93.10	0.85	0.95
93.50	0.87	0.90
92.01	0.87	0.90

TABLE 3.EXPERIMENTAL RESULT FOR TRAINING/TESTING RULE 70/30

Accuracy	Spam precision(SP)	Spam Recall(SR)
90.92	0.83	0.93
93.04	0.87	0.93
93.04	0.87	0.93
90.81	0.84	0.91
90.02	0.87	0.93

The experimental results obtained for various training/testing rule for BPNN shows that the increase in the training set reduces the false positive (i.e. low false positive rate is obtained).

TABLE 4.EXPERIMENTAL RESULT FOR TRAINING/TESTING RULE 60/40

Accuracy	Spam precision(SP)	Spam Recall(SR)
91.29	0.85	0.91
92.08	0.86	0.92
91.24	0.85	0.90
91.72	0.86	0.94
91.29	0.85	0.91

TABLE 5.EXPERIMENTAL RESULT FOR TRAINING/TESTING RULE 50/50

Accuracy	Spam Precision(SP)	Spam Recall(SR)
91.46	0.87	0.89
87.64	0.81	0.85
89.45	0.79	0.91
71.70	0.47	0.66
89.96	0.82	0.90

TABLE 6PERFORMANCE MEASURES FOR VARIOUS TRAINING/TESTING SET OF BPNN

Training/Testing ratio	Accuracy	Spam precision(SP)	Spam Recall(SR)
90/10	93.7	0.87	0.94
80/20	92.07	0.87	0.92
70/30	91.47	0.84	0.92
60/40	91.72	0.85	0.91
50/50	86.04	0.74	0.84

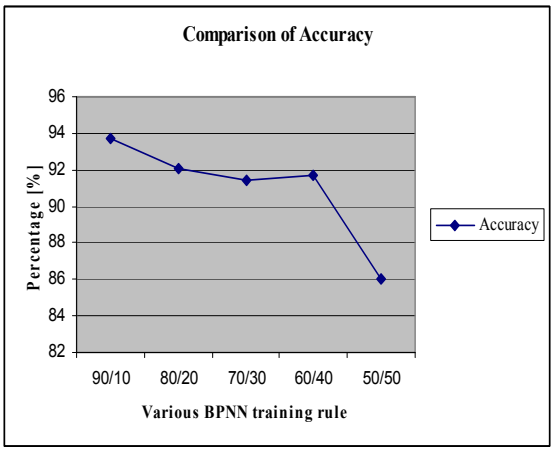

Figure 1. Comparison of Accuracy

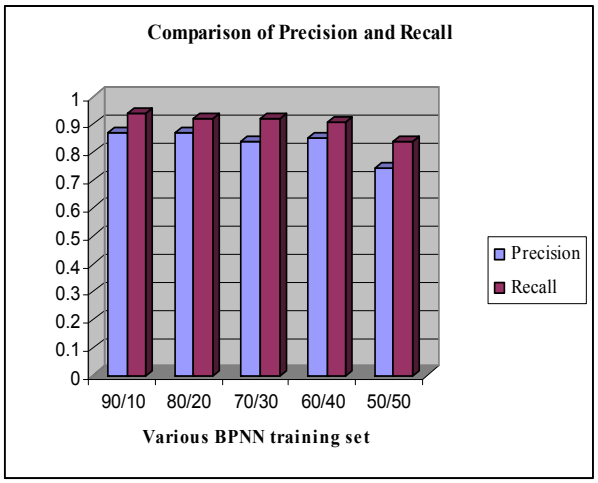

Figure 2. Comparison of precision and recall for various BPNN training set.

V. CONCLUSION

In this paper, we presented a new method for filtering image spam using gradient histogram as a key technique in Feed forward back propagation algorithm. In this technique gradient histogram is effectively utilized to extract the feature points from an image. Experimental results indicates that gradient histogram based image spam classification method provide good result and 90/10 training and testing yield better performance than the other pair of training and testing sets. Future work include, except this approach considered in this paper, a variety of back propagation network can be proposed for image spam filtering. And also in addition to gradient histogram other image features may aim at classifying the image spam e-mails.

REFERENCES

[1] H. B. Aradhye, G. K. Myers, and J. A. Herson, "Image analysis for efficient categorization of image-based spam e-mail," Eighth International Conference on Document Analysis and Recognition, (ICDAR'05), vol. 2, pp. 914-918, 2005.

[2] Battista Biggio, Giorgio Fumera, Ignazio Pillai, Fabio Roli, "Image Spam Filtering by Content Obscuring Detection," Fourth Conference on Email and Anti-Spam (CEAS 2007), 2007.

[3] Bin Cui, Anirban Mondal, Jialie Shen, Gao Cong and Kian-Lee Tan, "An Effective E-mail Classification via Neural Networks," Database and Expert Systems Applications, vol. 3588, pp. 85-94, 2005.

[4] Blosser and D. Josephsen, "Scalable Centralized Bayesian Spam Mitigation with Bogofilter," in USENIX LISA, 2004.

[5] Chun-Chao Yeh and Nai-Wei Yeh, "Octet-Histogram Based Near-Duplicate Mail Detection for Spam Filtering," IEEE, 2005.

[6] L. F. Cranor, B. A. LaMacchia, "Spam!" Communications of the ACM, vol. 41, pp. 74-83, 1998.

[7] M. Dredze, R. Gevaryahu, and A. Elias-Bachrach, "Learning Fast Classifiers for Image Spam," Fourth Conference on Email and Anti-Spam (CEAS), 2007.

[8] F. Fdez-Riverolaa, E. L. Iglesiasa, F. Díazb, J. R. Méndeza, and J. M. Corchadoc, "Applying lazy learning algorithms to tackle concept drift in spam filtering," Expert Systems with Applications, Elsevier, vol. 33, no. 1, pp. 36-48, 2007.

[9] G. Fumera, I. Pillai, and F. Roli, "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), vol. 7, pp. 2699-2720, 2006.

[10] J. Graham-Cumming, "The Spammer's Compendium," http://www.jgc.org/tsc.html.

[11] Ian Stuart, Sung-Hyuk Cha and Charles Tappert, "A Neural Network Classifier for Junk E-mail," vol. 3163, pp. 442-450, 2004.

[12] Jason R. Bowling, Priscilla Hope, and Kathy J. Liszka, "Spam Image Identification Using an Artificial Neural Network."

[13] Jose A. Rodrıguez and Florent Perronnin, "Local gradient histogram features for word spotting in unconstrained handwritten documents," Proceedings of ICFHR 2008.

[14] Kang Li and Zhenyu Zhong, "Fast Statistical Spam Filter by Approximate Classifications," in Proceedings of ACM SIGMETRICS, St. Malo, France, pp. 347 – 358, June 2006.

[15] Ngo Phuong Nhung, "An Efficient Method for Filtering Image-Based Spam E-mail," Book on Computer Analysis of Images and Patterns, vol. 4673, pp. 945-953, 2007.

[16] Y. Nong, S. Vilbert and Q. Chen, "Computer Intrusion Detection through EWMA for Auto-Correlated and Uncorrelated Data," IEEE Transactions on Reliability, vol.52, pp. 75-82, 2003.

[17] Russell, S. and P. Norvig, "Artificial Intelligence: A Modern Approach," 2nd Edition, Prentice Hall, Inc, 2003.

[18] Z. Wang, W. Josephson, Q. Lv, M. Charikar, and K. Li, "Filtering Image Spam with Near-Duplicate Detection," Fourth Conference on Email and Anti- Spam (CEAS), 2007.

[19] C.-T. Wu, K.-T. Cheng, Q. Zhu, and Y.-L. Wu, "Using visual features for anti-spam filtering," IEEE International Conference on Image Processing, vol. 3, pp. 509–512, 2005.

[20] Yan Gao, Ming Yang, Xiaonan Zhao, Bryan Pardo, Ying Wu, T. N. Pappas, and A. Choudhary, "Image Spam Hunter," ICASSP 2008, IEEE 2008.

[21] C. Zhang, Wei-Bang Chen, Xin Chen, Richa Tiwari, Lin Yang, and Gary Warner, "A Multimodal Data Mining Framework for Revealing Common Sources of Spam Images," Journal of Multimedia, vol. 4, no. 5, pp. 313-320, 2009.

M. Soranamageswari: The author received Master degree of Computer Application in 1996 from Avinashilingam University for Women and M.Phil degree in Computer Science in 2003 from Manonmaniam Sundaranar University, India. She is serving as a Asst.Professor of Computer Science Department at LRG Govt Arts College for Women, Tirupur. She is currently working toward the Ph.D degree at Avinashilingam University for women in India.

Her research interest includes Image Processing, Machine Learning and Network Security.

Dr.C.Meena: The co-author received the Master degree of Computer Application in 1990 from Madurai Kamaraj University and Ph.D degree in 2006 from Bharathiyar University,India. She is acting as a Head of Computer Center at Avinashilingam University for Women in India. She is a member in Research review committee in Mother Teresa University for women, Kodaikanal and Anna University, Chennai. She has more than 10 National and International Publications.

Her research interests include pattern recognition, image processing and their applications in Biometrics