

ANTI IMAGE SPAM FILTER USING UNIFIED FEATURE FUSION AND HYBRID SUBSET SELECTION ALGORITHM

S.Dhanaraj ¹ and Dr.V.Anuratha ²

¹Assistant Professor,

Department of Information Technology,

Sree Saraswathi Thyagaraja College,

Pollachi, Coimbatore District , Tamilnadu, India

²Assistant Professor,

PG Department of Computer Science,

Sree Saraswathi Thyagaraja College, Pollachi,

Coimbatore District , Tamilnadu, India

January 16, 2018

Abstract

The explosive growth in the usage of electronic mail is also increasing the need for an efficient anti-image spam filter that can help Internet Service Providers, administrators and users. The current market need is to have filters that can detect spam emails inspite of the numerous innovative ways spammers use to create image spam emails and which can stop them from entering the inbox. In general, an anti spam filter identifies spam/ham mails in two main steps. The first is feature extraction and the second is classification of ham and spam mails. In this paper, the usage of multiple features is analyzed. An intermediate step that performs fusion of these multiple features and constructs optimal feature subset from this fused set is also presented. This is

performed using an amalgamation of techniques that combines Genetic Algorithm with hybrid multiple filter-based and wrapper-based feature selection algorithm. The Support Vector Machine classifier is used to classify email images as spam or ham. Experimental results show that the proposed filter is efficient in terms of precision, recall, f-measure and accuracy when compared to the conventional system.

Key Words : Anti-Image Spam Filter, Filter-based Feature Selection Algorithm, Genetic Algorithm, Multiple Feature, SVM Classifier, Wrapper-based Feature Selection Algorithm.

1 INTRODUCTION

Electronic Mail (Email) is one of the primary communication media for information exchange and distribution, due to the various merits offered like being fast, geographically independent and easy to use. An unwanted consequence of the e-mails is the Spam mail, also called as junk mails or unsolicited bulk mails, which is defined as unsolicited messages sent by email in the form of commercial advertisements that may hide malwares as scripts or other harmful executable file attachments. Spam mails are considered as one of the biggest and most serious issue of emailing, which has been growing steadily since 1990s and is expected to be the most crucial security threat of this century (McCafe Lab Threats Report, 2016 and Shcherbakova et al., 2015). Some examples of image spam mails are shown in Figure 1

Figure 1 : Image-based Spam Examples



Out of the various techniques used during the creation of spam mails, image spam mails are predominantly used by spammers. Image spam mails have the ability to evade conventional text-based

spam filters by storing the textual part of the message in image format (GIF, JPEG, PNG or BMP) (Wang et al., 2007). Image spam mails are difficult to detect due to the various techniques used by spammers for spam image creation. Several filtering techniques like Bayesian filter, image analysis algorithm, statistical approaches and machine learning algorithms have been proposed (Dinh et al., 2015; Alsmadi and Alhami, 2015). A detailed review of the various anti-spam techniques proposed was presented in our previous work (Dhanaraj and V. Karthikeyani). From this review, it was understood that even though several effective spam filtering software that aim to eliminate 100% of the images spam have been developed, this dream is still not a reality and research in this field is still very active. To reduce the gap between this reality and dream, this paper proposes feature extraction and selection algorithm combined with machine learning algorithm to improve the accuracy of image spam mail detection.

The proposed image spam filter enhances spam identification in 3-steps algorithm. The first step is feature extraction where four feature sets are extracted, instead of single set of features used by the existing filters. The second step performs feature subset selection and fusion using multiple filtering algorithm, Genetic Algorithm (GA) and wrapper-based algorithm. The final step uses the SVM classifier to identify image spam mails. The rest of the paper is arranged as follows. Section 2 presents the methodology of the proposed filter. Section 3 evaluates the performance of the proposed method in their effectiveness in detecting image spam. Section 4 concludes the work with future research directions.

2 ANTI-IMAGE SPAM FILTER

Given the recent upsurge in image-based spam, the proposed system is developed to classify the images, based on image features into spam or ham. The proposed anti-spam filter is termed as Anti image spam filter using unified Feature fusion and Subset selection algorithm and SVM classifier (AF2S) in this paper. AF2S addresses the task of recognizing image spam as a two-class classification problem between spam and ham images, in a given feature space. The proposed AF2S, as mentioned earlier, identifies spam

and ham mails in four main steps, namely, feature extraction, feature fusion, optimal feature subset selection and spam/ham classification. Details regarding the methods used in each of these steps are presented in the following sub-sections.

2.1 Feature Extraction

Extraction of important and useful features from large heterogeneous spam imagery data set is a crucial task in spam detection. Feature extraction is a process, which transforms the large input data into a form that reduces redundancy. Research in recent years has shown that multiple features can increase the performance of algorithms in pattern recognition and analysis (Srisuwan and Ruchanurucks, 2014). Continuing in this line of research, this work also extracts multiple features, categorized into four main categories, as explained in the following subsections.

- A. Metadata Feature Set (MFS) Metadata features can be quickly extracted and have very low computation cost. The five basic metadata features extracted are image width, image height, bit depth, image file type and image size in KB. From these basic features, three more features, namely, aspect ratio, image area and compression are estimated. The image file type feature is constructed as binary feature, whose value is 1 if the file is of a specified type and 0, otherwise. This work considers four image file types, namely, JPEG, GIF, PNG and BMP. Thus, the Metadata Feature Set (MFS) is constructed as an 11-dimensional array, whose elements are presented in Table 1.
- B. Color Feature Set (CFS) As RGB (Red-Blue-Green) color space is the most basic and well-known model, the Color Feature Set (CFS) is constructed using the same model. Seven color features are extracted, namely, average RGB, skew, kurtosis, color histogram, color moments and color coherence vector. Thus, the CFS constructed as presented in Table 1.
- C. Texture Feature Set (TFS) Texture of images is defined as a function of the spatial variation in pixel intensities and this feature category is selected because of the reason that the texture of natural images and computer generated images (as the case

with spam) is very distinct. Five texture features, namely, autocorrelation, edge frequency, primitive length, co-occurrence matrix and wavelet coefficients are extracted from the images. The features extracted from co-occurrence matrix are contrast, energy, entropy, homogeneity and variance. Wavelet features are extracted using level 2 Haar wavelet transformation, as this is the simplest with low computational complexity. The Texture Feature Set (TFS) used in this work are summarized in Table 1.

- D. Shape Feature Set (SFS) Shape is another set of features that has been found to be useful during the discrimination of ham and spam images. In this work, the Shape Feature Set (SFS) (Table 1) is constructed using geometric moments, eccentricity, elongation, compactness, moments of inertia, orientation and a set of edge features. The edge features extracted are number of edges, average edge length.

2.2 Feature Fusion and Subset Selection

issue related to feature set size otherwise termed as “curse of dimensionality”. This work proposes an amalgamation of algorithms, applied in three stages, to perform fusion and subset selection. Stage 1 performs preliminary feature integration to combine the feature vectors from Step 1. Stage 2 uses a multiple filtering algorithm to select optimal features. In order to further improve the feature subsets, Stage 3 also uses a wrapper-based feature selection using Genetic Algorithm (GA) and Support Vector Machine (SVM). The resultant optimal feature subset occupies reduced feature space and at the same time, will help to increase image spam detection accuracy. This algorithm is referred to as “Multi-stage Feature Selection and Fusion Algorithm (MFSFA)” in this work.

Table 1 : Features Extracted

Feature Category	Feature Name
1. Metadata Features	MF1 : Image width (number of columns) MF7 : JPEG (0 or 1)
	MF2 : Image Height (number of rows) MF8 : GIF (0 or 1)
	MF3 : Aspect Ratio : MF1/MF2 MF9 : PNG (0 or 1)
	MF4 : File Size (KB) MF10 : BMP (0 or 1)
	MF5 : Image Area : MF1 * MF2 MF11 : Bit Depth (bits / pixel)
	MF6 : Compression (MF4/MF5)
2. Color Features	CF1 : Average RGB : The normalized average values in R, G and B channel of each pixels in an image.
	CF2 : Color Histogram : Defined as a compact summary of the image, it is represented as a vector, $H = (h_1, \dots, h_m)$ where h is a bucket and each bucket h_j contains the number of pixels of color j in the image. This work used a 6-bit color space thus extracting 64 feature vectors.
	CF3 : Color Moments : The three central moments extracted from the image are mean, standard deviation and skewness. This work, thus extracts 9 colour moment features, from 3 color bands of RGB.
	CF4 : Color Coherence Vector : Defined as the degree to which color pixels are members of large similarly colored regions and is estimated using the procedure described by Mehta <i>et al.</i> (2008). This feature Vector is constructed with 128 features.
3. Texture Features	TF1 : Autocorrelation
	TF2 : Edge Frequency
	TF3 : Primitive length
	TF4 : Co-occurrence Matrix (Contrast, energy, entropy, homogeneity and variance)
	TF5 : Wavelet Coefficients using 2D Haar Wavelet
4. Shape Features	SF1 : Geometric Moments SF5 : Moments Of Inertia
	SF2 : Eccentricity SF6 : Orientation
	SF3 : Elongation SF7 : Edge Features (No. of Edges, Edge Length and Average Edge Length)
	SF4 : Compactness

A. Stage 1 : Feature Integration

The first stage of the MFSFA is to perform a preprocessing step that combines the four feature vectors using a simple integration algorithm. Let the dimensions of the four feature vectors, MFS, CFS, TFS and SFS, be d_1 , d_2 , d_3 and d_4 . The dimensions of each of the four vectors are different, thus will result in imbalance during integration. This problem is solved by using a weighted fusion coefficient, w , during integration. The method of estimating w for two feature vectors is shown in Figure 2, which is repeatedly applied between two pairs of features to obtain the final Integrated Feature Set (IFS). Thus, the dimension of the final IFS (Integrated Feature Set) will be $(d_1 +$

$d2 + d3 + d4$)-dimension. Finally, the IFS is split into training (with 70% features) and testing data (with 30% features).

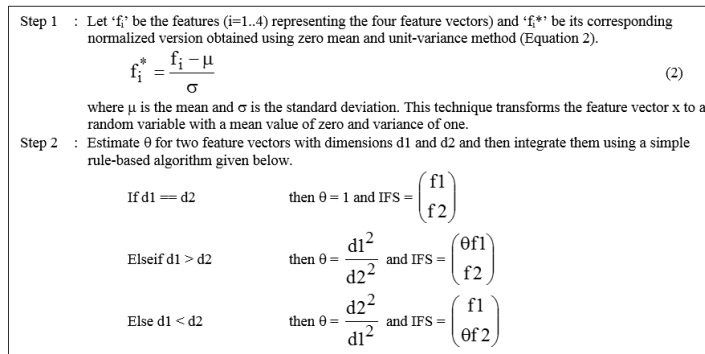


Figure 2 : Integration of Feature Vectors (Level 1) Algorithm

B. Stage 2 : Search Space Reduction

In the second of MFSFA, filtering algorithms are used to pre-select optimal features from the training feature set from Step 1, so that the feature space size is reduced. For this purpose, a multiple filtering algorithm is designed with Markov Blanket Filter. The four filtering algorithms are Mutual Information based Feature Subset Selection (MIFS), Pearson Correlation based Feature Subset Selection (PCFS), Chi-Squared test based Feature Subset Selection (CSFS) and Fisher criterion Score based Feature Subset Selection (FSFS) algorithms. A Markov Blanket Filter (Koller and Sahami, 1996) is then used to select only the non-redundant and relevant features that have maximum predictive power. This step results in four subsets (FSS1, FSS2, FSS3, FSS4), obtained from each of the selected filtering algorithm. This algorithm is referred to as "Multiple Filter-based Subset Selection Algorithm (MFA)"

C. Stage 3 : Wrapper-Based Feature Selection

It has been shown that a simple combination of the best individual features selected by the filter selection techniques does not necessarily lead to a good classification performance (Kapoor and Singh, 2016). So, in order to further refine the features selected, Stage 3 algorithm is used. This algorithm is referred to

as “Wrapper Feature Subset Selection Algorithm based on GA and SVM (WGS)”, in this work.

Initially, the four subsets obtained from the previous stage are used to construct a Feature Pool (FP), constructed through the use of union operation. The conventional GA-based feature selection algorithm is enhanced in two manners. The first is to accept FP as initial population and second is to convert the conventional algorithm as an extension to accept the current population. Both these changes will help in creating an optimal subset that can increase classification accuracy while reducing the feature space size. The general steps are shown in Figure 3.

The WGS algorithm uses GA to stage the search for optimal solutions from a population containing initial hypothesis that is generated using the feature subsets generated in the previous step. Here, each hypothesis (called as individual or chromosome) represents a potential solution and is encoded as a binary string, where a value '1' indicates a selected feature and a value '0' indicate the opposite. The initial population is then evolved in generations, where each chromosome of the current population is evaluated using the fitness function F (Equation 1).

$$F = w * c(FP) + (1 - w) * (1/s(FP)) \quad (1)$$

where 'w' is a parameter between 0 and 1 and is set to 0.8 after several empirical evaluations. The function is composed of two parts. The first part is a weighted classification accuracy $c(FP)$ from the classifier and the second part is weighted size $s(FP)$ of the feature subset represented by FP. Here, the parameter 'w' has to be chosen carefully so that increases the classification accuracy and at the same, reduces the size of FP. The induction algorithm used is the SVM classifier (Section 2.3). Three genetic operators were used during feature selection. They are selection, crossover and mutation. Roulette wheel selection is one of the most popular selection methods for genetic algorithm and is used by the study. The single-point crossover operator, where a cross-over point i is chosen at random so that the first i bits are contributed by one parent and the remaining bits by the second parent. Each individual has a probability pm to

mutate, in this work, a number of n bits to be flipped in every mutation stage are randomly chosen.

Careful analysis of the above method showed that the performance of WGS depends heavily on the initial population. The initial population should have relatively well-adjusted chromosomes so as to not result in premature convergence. Premature convergence results in a feature subset that may lose significant features as it may have been considered as bad chromosomes. The extended model dynamically changes the conditions for the next iteration by modifying the selected solution set to the initial population. For the new population, the selection scheme, the crossover type and the mutation rate are re-estimated and the whole process is repeated until an optimal solution is reached.

2.3 Step 3 : SVM Classifier

In order to detect spam and ham messages efficiently, this work uses the frequently used SVM classifier (Vapnik, 1995). An SVM classifies image emails by finding the perfect hyperplane that separates all feature points of one class (ham) from those of the other class (spam). The hyperplane for an SVM means the one with the largest margin between the two classes. Margin means the maximal width of the portion parallel to the hyperplane that has no interior data points.

2.4 AF2S

The AF2S is designed to combine all the algorithms described in Sections 2.1, 2.2 and 2.3 in the order presented in Figure 4 to detect image spam and ham mail messages. These steps cumulatively enhance the spam detection process as proved by the experimental results discussed in the following section.

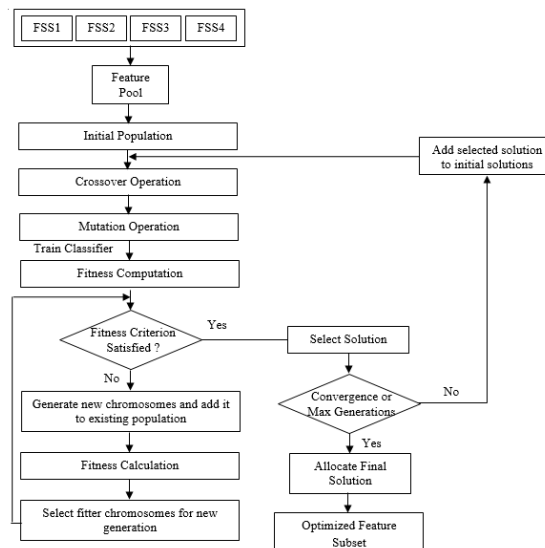


Figure 3 : Steps in WGS

3 EXPERIMENTAL RESULTS

Several experiments were conducted to evaluate the algorithms presented in this paper and to study its effect on spam detection. All the experiments were conducted using images from four benchmark datasets, namely, TREC 2005-2007 (<http://plg.uwaterloo.ca/~gvcormac/spam>), Dredze (https://www.cs.jhu.edu/~mdredze/datasets/image_spam), Princeton Spam Image Benchmark (<http://www.cs.princeton.edu/cass/spam>) and Image Spam Hunter (<http://www.cs.northwestern.edu/~yga751/ML/ISH.htm>). One single dataset was constructed using all the ham/spam image emails from the four dataset. The final dataset, thus contains, 2949 ham image emails and 11906 spam image emails. Details regarding the final dataset constructed in shown in Table 2. A 10-fold cross validation method was used with 70% of data used for training and rest for testing. A baseline system was first constructed for comparison purpose, which used the CFS and conventional SVM classifier for spam/ham classification. This baseline system is referred to as “Baseline system

for Spam/ham classification using Color features and SVM classifier (BSCS)” and do not use any feature selection algorithms and optimization operations.

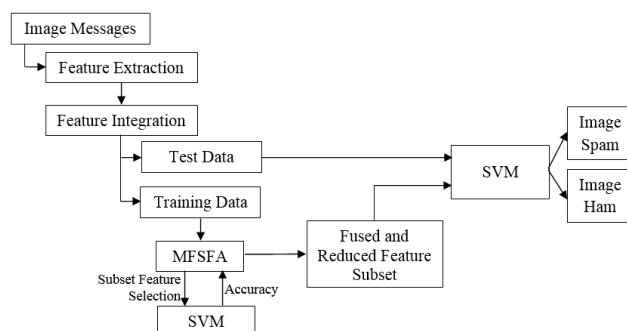


Figure 4 : Steps in AF2S

Table 2 : Details of Dataset Used

Dataset	No. of Ham Image Emails	No. of Spam Image Email
TREC	-	6728
<u>Dredze</u>	2021	3297
Princeton	-	1071
Hunter	928	810
Final Dataset	2949	11906

Six metrics, namely, Spam Recall (SR), Ham Recall (HR), Spam Precision (SP), Ham Precision (HP), Spam F-Measure (SFM), Ham F-Measure (HFM) and Accuracy (all measured in %) were used during performance evaluation. The experiments were designed in two stages. The first stage analyzes the effect of feature selection and fusion algorithms on spam detection. Stage 2 experiments evaluate the performance of the proposed AF2S anti spam filter.

3.1 Analysis of Feature Selection and Fusion Algorithms

Table 3 show the spam/ham precision, recall and F-measure of the MFA, WGS, MFSFA and BSCS algorithm along with accuracy

of ham/spam classification obtained while using different feature selection and fusion algorithms.

Table 3 : Analysis of Feature Selection and Fusion Algorithm

	SP	SR	SFM	HP	HR	HFM	Accuracy
BSCS	78.15	79.87	79.00	80.72	81.67	81.19	87.94
MFA	82.08	84.47	83.26	84.46	85.06	84.76	91.03
WGS	83.75	86.80	85.25	86.55	87.44	86.99	93.48
MFSFA	87.04	89.72	88.36	89.07	91.02	90.04	95.98

From the table, it can be seen that the inclusion of feature subset algorithm has increased the efficiency of anti-spam filter with all the selected performance metrics. The proposed MFSFA produced the maximum efficiency when compared to MFA and WGS, indicating that the combination of filter and wrapper based algorithms with GA as a success towards spam image email detection. The MFSFA has improved accuracy of spam, detection by 8.38% when compared to BSCS.

3.2 Analysis of AF2S

Table 4 shows the result obtained by the proposed spam filter, AF2S, introducing MFSFA in the baseline system, to classify ham and spam image emails.

Table 4 : Analysis of Anti-Spam filter

	SP	SR	SFM	HP	HR	HFM	Accuracy
BSCS	78.15	79.87	79.00	80.72	81.67	81.19	87.94
AF2S	88.70	91.40	90.03	90.50	92.00	91.24	97.23

From the table, it could be seen that both spam recall and spam precision has increased while using AF2S, with a gain of 11.89% and 12.61% over BSCS respectively. A similar trend can be envisaged

with ham recall and precision also. With respect to these two metrics, the AF2S showed an efficiency gain of 10.81% and 11.23% respectively over the baseline system. A maximum accuracy of 97.23% was obtained by AF2S, indicating that the proposed filter can detect spam image emails in an effective manner, thus, making it an ideal solution to web administrators and users for protecting mail boxes from spammers.

4 CONCLUSION

Increase in spam mails has necessitated the need for an efficient spam/ham identifier that can help Internet Service Providers, administrators and users to efficiently block image spam mail entering inbox. This paper presented an image spam/ham algorithm that used a feature vector that is an integration of multiple features and which was optimized using various techniques like filtering, genetic algorithm and wrapper method. The classifier for identifying spam and ham was support vector machine. Experimental results proved that the proposed method improves image spam identification when compared to the conventional algorithms. Future work is planned in the usage of text-based features that combines the OCR (Optical Character Reader) technology with machine learning classifiers to detect spam mails.

References

- [1] McAfee Labs Threats Report (2016)
<http://www.mcafee.com/in/resources/reports/rp-quarterly-threats-mar-2016.pdf> (Last Accessed During August, 2016), Pp. 1-46.
- [2] Shcherbakova, T., Vergelis, M. and Demidova, N. (2015) *Spam and phishing in Q3 2015, Quarterly Spam Reports, Securelist*, https://cdn.securelist.com/files/2015/11/Q3-2015_Spam-report_final_EN.pdf (Last Accessed During August, 2016), Pp. 1-21.

- [3] Wang, Z., Josephson, W., Lv, Q., Charikar, M. and Li, K. (2007) *Filtering Image Spam with Near-Duplicate Detection*, Fourth Conference on Email and Anti-Spam (CEAS), Pp. 1-10.
- [4] Dinh, S., Azeb, T., Fortin, F., Mouheb, d. and Debbabi, M. (2015) *Spam campaign detection, analysis and investigation*, Digital Investigation, Proceedings of the Second Annual DFRWS Europe, Elsevier, Vol. 12, Supplement, Pp. S12-S21.
- [5] Alsmadi, I. and Alhami, I. (2015) *Clustering and classification of email contents*, Journal of King Saud University-Computer and Information Sciences, Vol. 27, Issue 1, Pp. 46-57.
- [6] S. Dhanaraj and V. Karthikeyani, *A study on e-mail image spam filtering techniques*, Pattern Recognition, Informatics and Mobile Engineering (PRIME), 2013 International Conference on, Salem, 2013, pp. 49-55.
- [7] Srisuwan, T. and Ruchanurucks, M. (2014) *Smoke detection using GLCM, wavelet, and motion*, Proceedings of SPIE 9069, Fifth International Conference on Graphic and Image Processing (ICGIP 2013), 90691H, Pp. 1-6.
- [8] Yang, J., Jang, R., Zhang, Y. and Shen, H.B. (2013) *High-accuracy prediction of transmembrane inter-helix contacts and application to GPCR 3D structure modeling*, Bioinformatics, Pp. 1-6.
- [9] Mehta, B., Nangia, S. and Gupta, M. (2008) *Detecting image spam using visual features and near duplicate detection*, Proceedings of the 17th international conference on World Wide Web (WWW '08). ACM, New York, NY, USA, 497-506.
- [10] Koller, D. and Sahami, M. (1996) *Toward optimal feature selection*, Proceedings of the Thirteenth International Conference on Machine Learning, Pp. 284-292.
- [11] Kapoor, P. and Singh, M. (2016) *Minimum Redundancy Maximum Relevance (mRMR) based feature selection technique for pattern classification system*, International Journal on Recent and Innovation Trends in Computing and Communication, Vol. 4, Issue 4, Pp. 577-580.

- [12] Vapnik, V. (1995) *The Nature of Statistical Learning Theory*. Springer, New York.

