**Image Spam Filtering based on Maximum Entropy Segmentation method**

Quang Minh Ha

Faculty of Information Technology

Hanoi University

Hanoi, Vietnam

e-mail: minhhq\_fit@hanu.edu.vn

Van Dong Phung

Faculty of Information Technology

Hanoi University

Hanoi, Vietnam

e-mail: dongpv@hanu.edu.vn

Frank Jiang

School of Engineering and IT

University of New South Wales

Canberra, Australia

e-mail: F.Jiang@adfa.edu.au

Quang Linh Nguyen

Faculty of Information Technology

Hanoi University

Hanoi, Vietnam

e-mail: linh4nq@gmail.com

*Abstract*— Spam email has evolved from a simple text-based message to a complicated form including images embedded with spam message texts. To tackle this problem, the machine learning community has conducted many researches on different approach, from the simple image processing to a fast Optical Character Recognization (OCR) method. However, spam images are not only just advertisements but also unwanted images from disturbing senders. It is difficult for OCR to detect non-content images. Therefore, in this paper, we propose a new framework using a classification method (Naïve Bayesian Classifier) and Shannon entropy with multiple levels extensions to filter spam images. The experiment shows a promising result with acceptable prediction accuracy. It helps to create a new way to detect the spam emails by classifying images into meaningful categories (e.g., spam or ham, animals or human).

Keywords—Image spam filtering; Bayesian classifier; Maximum entropy segmentation method

# Introduction

Email Spams are unsolicited emails which are sent massively to many other recipients from one user [1]. According to the 2010 Annual Security Report from Symantec Message Labs Intelligence, approximately 89% of all emails are spam; resulting in an estimated 260 billion spam emails every single day [2]. Spam e-mails may contain text or both text and images. Current spam filters rely on recognizing different components of e-mail: header, body, attachments, and sender’s address to clarify its characteristics [3]. In order to tackle this phenomenon, the machine learning community has conducted many approaches mostly based on the text classification methods such as Bayesian Probabilistic approaches, Nearest neighbor classification, Decision tree, Neural networks, Support Vector Machines [4]. These approaches achieved an outstanding result with very high accuracy detection rate (for instance, more than 90% achieved by use of Naïve Bayesian Classifier [5]) when dealing with text-only e-mails.

However, to avoid the detection from spam filtering machines, spammers started to add images which usually contain the advertising content (text embedded in the picture). As spammers tend to use the program called CAPTCHA to transform normal words into distorted form [6]. A filtering system is required to have the capability to extract texts from images. Since embedded text extraction and analysis is a difficult problem, current filters cannot properly analyze these emails, Some approaches in this area have been employed to filter spam images.

There is the simple approach that applies simple image processing [7,8], while more advanced method adapts the OCR [9] (optical character recognition) to detect the embedded text in spam images. These approaches can effectively filter spam images, however ineffective to the performance, this makes it difficult to apply in the real world environment. To solve this performance peak, a research has introduced a “fast” classifier to improve the performance during Just-In-Time feature extraction [10]. Overall, current image spam filters mainly focus on detecting embedded text in e-mail pictures using simple image processing and OCR methods. These approaches rely on at least two steps to extract features: 1) processing an image into a pattern that can be recognized and 2) applying text extraction. More challenges comes from the real-life image spam problems where unwanted images may not contain any text.

In this paper, we propose a new approach to filter spam images. Instead of extracting texts from images, we extract image objects and treat them as features in the classification problem. Firstly images are converted into a gray-scale image, and then use Genetic Algorithms [11] to select its multilevel thresholds with Maximum entropy segmentation method as the fitness function. Each threshold (t) then will be treated as a feature. Using Naïve Bayesian Classifier, we train and test the Spam and Ham e-mails based on extracted features. It is author’s belief that the new method can be used to classify not only spam and ham but also image e-mails into different categories.

This paper is structured as follows. First, we discuss the preliminaries that relate to this paper including Naïve Bayesian Classifier, Maximum entropy segmentation method and dataset initializations. Next we introduce our new framework to filter image spam and the feature selection process. After that, an experiment will be carried out to test the proposed framework. Performance benchmarking and critical discussions are also included in this section. Finally, we conclude the paper and look to the future work.

# Preliminaries

## Naïve Bayesian Classifier

Naive Bayesian Classifier is based on Bayes’ Theorem, which is a simple formula used for calculating conditional probabilities. In the simplest form, the Bayes’ Theorem can be expressed by the following formula [14]:



One highly practical Bayesian method is the naive Bayes learner, often called the Naive Bayes classifier [15]. Its performance has been shown to be comparable to that of neural network and decision tree learning, especially when the dataset is small, since the Naive Bayes Classifier does not require a large amount of training data. This makes it very suitable for the problem in this paper.

The Naive Bayes classifier applies to learning tasks where each instance x is described by a conjunction of attribute values and where the target function f (x) can take on any value from some finite set V. A set of training examples of the target function is provided, and a new instance is presented, described by the tuple of attribute values (al, a2,…, an). The learner is asked to predict the target value, or classification, for this new instance. The Bayesian approach of classifying the new instance is to assign the most probable target value, vMAP (MAP-Maximum a posterior), given the attribute values (al, a2,…, an) that describe the instance.

The Naïve Bayesian Classifier is described as follow:



## Maximum Entropy Segmentation Method

Maximum entropy method was first introduced by Kapur [12], which applied the concept of Shannon entropy to maximize the amount of information of object and the background in the image. The optimal threshold (t) can be retrieved by measuring the entropy of gray level histogram [13]. The gray level ranges from 0 to 255 (with 256 colors).

In details, an image can be divided into objects (O) and background (B) by a threshold *t* (*t* is a gray-scale level). H(*t*) denotes the entropy of an image and be calculated as follows:



= 

Where is the frequency of the gray level i (0,1,2…L-1). N is the total number of pixels in the image. The maximum entropy of the image is the maximum value of H(*t*) with .

The total entropy of the image with multilevel thresholds is calculated as follows:







where. The maximum entropy of the image is the maximum value of 

As can be seen, the maximum entropy of the image in multilevel thresholds problem can be treated as selected features of the image and can be used by different classifiers such as Naïve Bayesian Classifier (which is applied in this paper).

## Dataset

We get the dataset for this paper from Northwestern University [18] and Johns Hopkins University [16], which was used in their researches. 930 spam images and 320 ham images are selected from the original set of 3143 images for the experiments. Most of spam images are advertisements that easily can be seen from spam emails everyday. Ham images are common images in many categories: daily life, landscape photos, computer graphics, etc.

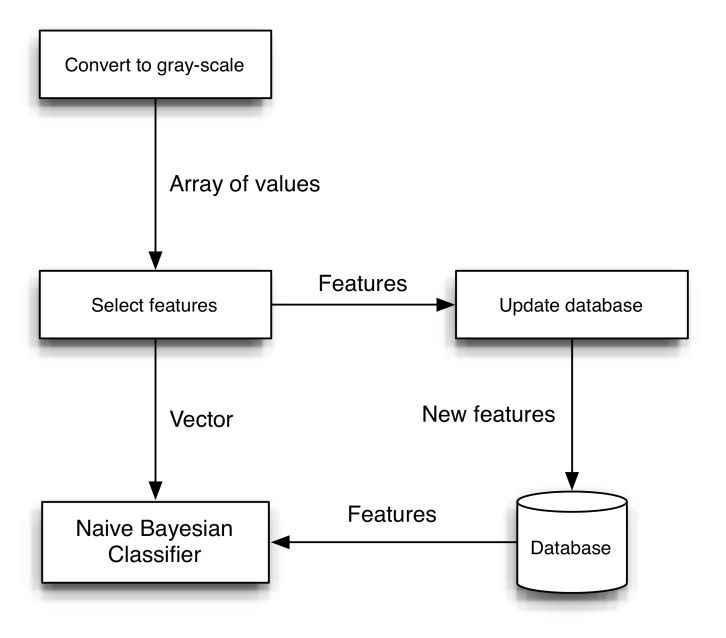
Since this paper is focused on image classification, we do not include emails associated with the images.

# Theoretical Framework

## Descriptions of the Framework

The proposed framework is depicted in Figure 1. The detailed processing steps are described as follows:

* First, spam image is read and converted into gray-scale image. Each pixel of the image will contain a gray level (from 0 to 255), forming a two-dimensional array of values. The array is stored into the memory for the next step.
* Next, using the array of gray level values, we run the Genetic Algorithms to select the best thresholds. The number of thresholds is pre-defined (see Section IV).
* After that, the thresholds are used as features of a vector to be learned (score calculated) or to be classified. Each image is treated as a vector.



1. Image spam filtering framework based on maximum entropy segmentation method.

## Feature selection using GA

Using Genetic Algorithms, we treat each set of thresholds as a Chromosome, using the equation to calculate the entropy of image as Fitness function. In the selection process, Roulette Wheel method was used [17]. The following pseudo-code describes the feature selection procedure:

|  |
| --- |
| **Procedure** Find-entropy (f) **returns** void  **input**: f, the pgm file path  **ouput**: a set of thresholds will be write to a file.  **global**: BEST\_CHROMOSOME to store the best result  save\_best\_obj threshold for BEST\_OBJECTIVE to save entropy list  read data from f;  **while** time out **or** BEST\_OBJECTIVE smaller than optimal value  **do**  calculate entropy for image and add to entropy list  selection, find out BEST\_CHROMOSOME and BEST\_OBJECTIVE  mutation  crossover  **if** BEST\_OBECJTIVE < save\_best\_obj **then**  write entropy list to file  **end if**  **end while** |

## Adapting Naïve Bayesian Classifier

We decide to use Naïve Bayesian Classifier in this paper because of its simplicity in implementation which makes it easier to adapt to the image processing process. Moreover, the amount of dataset is quite small and this classifier fits best in the given condition.

In this framework, Naïve Bayesian Classifier was used to filter spam images. In order to adapt this statistical classification method into this problem, we treat each gray-scale color as a “token” in text classification problems. Then, using *m-estimate of probability* [15], we calculate the score of each gray-scale color (there are 256 colors) in each class (spam or ham).The equation is described as follows:

P[tk | vj] = 

# Experiments

## Experiment Settings

In this research, different numbers of thresholds were used to evaluate the performance of the classifier. The experiment sets are described as follows:

1. Experiment settings with different thresholds set.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Set 1** | **Set 2** | **Set 3** | **Set 4** |
| **Number of thresholds** | 5 | 6 | 10 | 20 |

As shown in Table I, Set 1 and Set 2 have a slightly change in the number of thresholds (5 and 6) while Set 3 and 4 own a large number of differences in thresholds count (10 compares to 20). These two groups test the relationship between number of thresholds, performance and precision rate.

## Performance of Naïve Bayesian Classifier

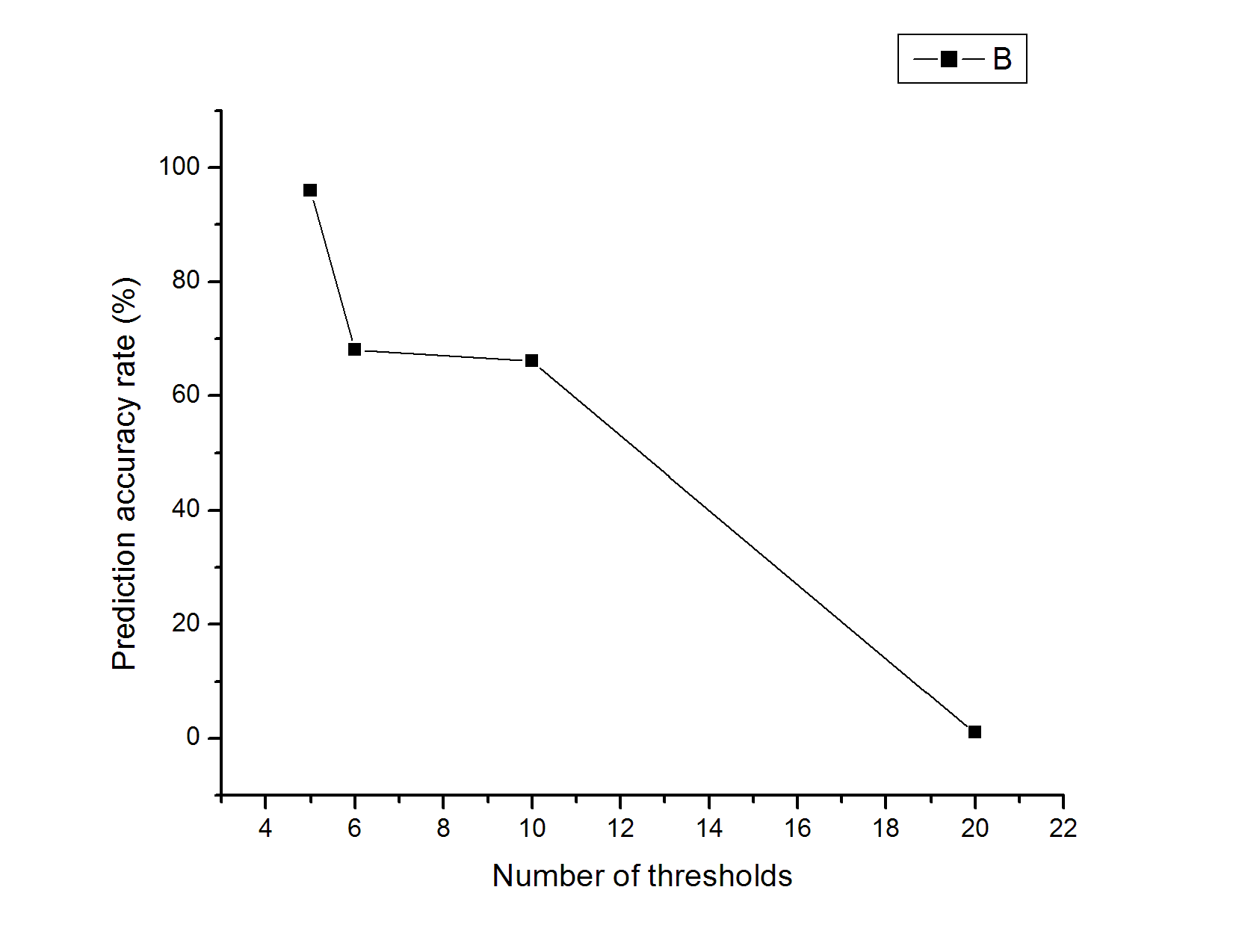
We determine the following criteria to evaluate the performance: training time (in seconds), filtering time (in seconds) which is the time taken to classify all spam images, the performance of each image (per image in seconds) and the prediction accuracy to filter spam images in each set (1 to 4). Table II shows the performance result.

1. Performance result using Naïve Bayesian Classifier with spam images

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Set 1** | **Set 2** | **Set 3** | **Set 4** |
| **Training time(s)** | 15 | 15 | 24 | 15 |
| **Filtering time(s)** | 3 | 2 | 2 | 2 |
| **Per image time (s)** | 0.003 | 0.002 | 0.002 | 0.002 |
| **Prediction accuracy (%)** | 96.078 | 68.627 | 66.667 | 1.960 |

## Discussions.

First, Table II shows a relationship between the number of thresholds and prediction accuracy. The larger number of thresholds, the lower prediction accuracy. With low number of thresholds (such as Set 1 that has 5 thresholds), we have images with more distinctive features with others (threshold defines the separation between objects and background). With high number of thresholds, the image is extracted into a larger number of smaller objects; therefore making it less distinctive with ham images. Figure 2 shows the relationship between number of thresholds and prediction accuracy as follows:



1. The relationship between prediction accuracy rate (%) and number of thresholds.

Second, Table II also shows an acceptable performance of the framework. It took the classifier less than 1 second to filter an image with high prediction accuracy (more than 96% with Set 1). This is promising since it is a new approach.

Finally, comparing with a research result from Mark Dredze [10], which also uses Naïve Bayesian and a new Just-In-Time feature extraction (JIT) in the experiment, this new framework shows a better performance (see Table III).

1. Performance Comparison with other approach

|  |  |  |
| --- | --- | --- |
| Approach | Our framework | JIT with NB |
| Prediction accuracy (%) | 96.078 | 80 |

# CONCLUSION

In this paper, we propose a new approach to filter spam images. It is the combination of Maximum entropy segmentation, genetic algorithms (for feature extraction) and Naïve Bayesian as the classifier. The experiment results show a good prediction accuracy rate with very high performance. We also compare the experiment result with another approach. In the future, we would extend the research to an image classification problem, to recommend users and prevent the access to malicious websites with large amount of unsolicited images. -

##### Acknowledgement

This research was supported by the Vietnam National Foundation for Science and Technology Development (NAFOSTED) under project number 102.01-2010.09; Rector-Funded Visiting Fellowship (2012) at the UNSW, Canberra, Australia, and the scheme of UNSW Vice-Chancellor's research initiatives (2011-2014).

##### References

1. Yan, G. (2008). Image spam hunter. *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008*. p1765 – 1768.
2. Pingdom. (2011). Diving deep into email spam statistics. *Available: http://royal.pingdom.com/2011/01/19/email-spam-statistics/.* [Accessed 10th Aug 2012].
3. Battista, B. (2007). *Image Spam Filtering by Content Obscuring Detection. CEAS 2007*. p40.
4. Yiming, Y. (1999). A re-examination of text categorization methods. *SIGIR '99 Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval.* p42 - 49.
5. Androutsopoulos, I. (2000). An Evaluation of Naive Bayesian Anti-Spam Filtering. *Proc. of the workshop on Machine Learning in the New Information Age*.
6. Luis von, A. (2003). CAPTCHA: Using Hard AI Problems for Security. *Advances in Cryptology - EUROCRYPT 2003, International Conference on the Theory and Applications of Cryptographic Techniques, Warsaw, Poland, May 4-8, 2003, Proceedings*. p294-311.
7. Aradhye, H. B., Myers, G. K., and Herson, J. A. (2005). Image analysis for efficient categorization of image-based spam e-mail. *Proceedings of the 2005 Eighth International Conference on Document Anal- ysis and Recongition (ICDAR ’05)*.
8. Wu, C.T., Cheng, K.T., Zhu, Q., & Wu, Y.L. (2005). Using visual features for anti-spam filtering. *IEEE International Conference on Image Processing*.
9. Fumera, G., Pillai, I., & Roli, F. (2006). Spam filtering based on the analysis of text information embedded into images. *Journal of Machine Learning Research,* 7, p2699–2720.
10. Mark, D. (2007). Learning Fast Classifiers for Image Spam. *CEAS 2007*. 1 (1), p6.
11. Stuart, R. (2002). Informed Search and Exploration. In: *Artificial Intelligence: A Modern Approach*. 2nd ed. United States: Prentice Hall. p116-122.
12. Kapur, J.N. (1985). Anewmethod for gray-levelpicturethresholding using the entropy of the histogram. *Computer Vision, Graphics, and Image Processing*. 29 (3), p273–285.
13. Tang, H. (2010). Image segmentation based on improved PSO . *2010 International Conference On Computer and Communication Technologies in Agriculture Engineering (CCTAE)*. 3 (1), p191 - 194.
14. Joyce, J. (2008). Bayes' Theorem. The Stanford Encyclopedia of Philosophy, Edward N. Zalta (ed.)
15. Mitchell, Tom M. *Machine Learning.* s.l. : McGraw-Hill, 1997.
16. Johns Hopkins University. 2007. Image Spam Dataset. [ONLINE] Available at: [http://www.cs.jhu.edu/~mdredze/datasets/image\_spam/](http://www.cs.jhu.edu/%7Emdredze/datasets/image_spam/). [Accessed 02 August 12].
17. Darrell, W. (1994). A genetic algorithm tutorial. *Evolutionary Programming*. 4 (2), p65-85.
18. Northwestern University. 2006. Image Spam Hunter [ONLINE] Available at: *http://www.cs.northwestern.edu/~yga751/ML/ISH.htm*. [Accessed 02 August 12].