Indian Journal of Science and Technology, Vol 8(13), DOI: 10.17485/ijst/2015/v8i13/65753, June 2015
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

A New Feature Selection Algorithm for Efficient Spam Filtering using Adaboost and Hashing Techniques

Khongbantabam Susila Devi[1*] and R. Ravi[2]

[1]Department of Information and Communication Engineering, Anna University, Chennai, 600 025, India; susilaphd14@gmail.com

[2]Department of Computer Science and Engineering, Francis Xavier Engineering College, Tirunelveli, 627003, India

Abstract

Email spam is one of the significant issues of the today's Internet. A steady measure of the spammer attack, which brings harm to corporate and bothering individual clients. There are lots of strategies to battle against the spam mail. Being as an effective procedure, it is appreciative to experience and channel the email spam. In this paper, we proposed a new approach by utilizing, the hashing algorithm with AdaBoost technique, an AdaBoost Technique classifies the text and image values. The proposed approach essentially accelerates the procedure of Adaptive Boosting (Adaboost) by lessening the amount of information focuses. This guarantees that Adaboost can prepare productive and insignificant misfortune of precision. The result of the proposed system is a decreased set of delegate preparing focuses, hence diminishing the general computational complexity of preparing and expanding the speed of the training process. This is will be used for large scale application.

Keywords: Adaboost, Classification, Hashing, Spam Filtering

1. Introduction

Presently, spam is considered as a most vital issue for web clients. The familiar of electronic mail (or E-mail), several people and companies discovered it an easy approach to distribute a massive measure of unsolicited messages to a tremendous number of users easily. These unwanted mass messages or junk emails are called spam messages[1]. E-mail spam has become an epidemic problem that can negatively affect the convenience of electronic mail as a communication means. Also wasting the user's time and effort to scan and delete the massive amount of received junk E-mails. It affects the network bandwidth, storage space and slows down the email server[2]. Nowadays, most of the spam messages have both text and image contents, for example "advertisements", the major challenge of spam detection problem is the "spammers will always find new ways to attack spam filters".

The issue of text filtering has been concentrated on two distinctive groups such as the Machine Learning (ML) and Information Retrieval (IR)[3]. Various algorithms have been proposed and assessed for text filtering previously, for example different frameworks have been produced to naturally classify the messages, including frameworks focused around decision-making rules, Bayesian classifiers support vector machines, neural networks and specimen based techniques some of which incorporate significant results. Numerous studies have been directed at individual parts of unique spam on which our proposed system depends[4,5].

A main reason for an increase in the spam, is that they

*Author for correspondence

are presented as a spam message with images. Volume of image spam communication has been increased today, which frustrates end-users. Generally, the Anti-Spam system is designed to deduct the image spams. In order to reduce the image spam, it is important to know that how to effectively filter out the image spam messages. A simple spam filter methods such as paralleling the hashes of image data and executing Optical Character Recognition (OCR) on images.

Current study endeavors to concentrate content characteristics, in order to classify the messages and distinguish words which are most bottomless in the Hash table by AdaBoosting approach and finally do filtering. An index quality model is intended for Adaboost to conform the weight of the data preparation. We have likewise shown the parallelized adaptation of the index for large scale applications. An Indexing technique based on Hash table function. Picture recognition analyses are completed to show the viability and effectiveness of the methodology.

In this paper, a new approach utilizing AdaBoost is proposed for the Email Spam filtering. Demonstration of the Adaboost algorithm appraised the forecasts and it is an extremely appropriate algorithm for tending to the spam filtering problem. We have extremely acquired the precise classifiers on the hashing and the algorithm is exceptionally hearty in further bolstering and over fits another good fortune of utilizing Adaboost, where it doesn't require the earlier machine filtering. Since, it can efficiently oversee huge capabilities of several methods. The paper is organized as follows. Previous related works are explained in section 2. Section 3, gives a brief introduction to proposing system, then section 4 is devoted to the AdaBoosting, section 5. Parallelization, section 6 devoted to performance evaluation and section 7. Contains the conclusion of this study.

2. Related Work

Many data mining research has addressed the biggest problem is detecting spam, it as a text classification problem this will be reduced by using the VIVO spam filtering technique. But in real world data set characteristics that make it a rich and challenging domain for data mining[6].

Past examinations of dynamic adapting in spam filtering have concentrated on the pool-based situation, where there is expected to be an expensive, unlabeled information set and the objective is to iteratively distinguish the best subset of samples for which to appeal marks. On the other hand, the improvement is an excessive methodology[7]. It can significantly lessen the naming and preparing expenses along with an immaterial extra over head in keeping up abnormal amounts of arrangement execution.

Collaborative spam filtering considers the discovery of not-awhile ago seen spam content, by abusing its massiveness. Our framework utilizes two novels and perhaps favorable methods for synergistic spam filtering[8,9]. Because of the framework's unpredictability, numerous parameters may influence the results, the demonstrated decisions about the key things which clarifies whether and why the framework is truly working admirably or not would oblige a great deal of extra testing (reenactments) and the utilization of factorial examination strategies.

A novel methodology to spam filtering focused on the adaptive statistical data compression models. The way of these models permits them to be utilized as probabilistic text classifiers focused around the character-level alternating binary sequences. By determining the messages as sequences, tokenization and another blunder inclined preprocessing steps are overlooked through the a system that is very robust[10,11]. The extensive memory necessities of these models are a real difficult of this methodology.

In content based spam filtering, the significant center arranges the email as spam or as ham, in light of the data that is available in the body or the content of the mail. So the header segment is disregarded if there should be an occurrence of content based spam filtering[12,13].

Artificial Immune System (AIS) is an order between the research regions that intends to manufacture computational brain power models by taking persuasion from Biological Immune System (BIS). BIS is a versatile regular framework, which has a few fascinating properties, for example, conveyed discovery, clamor tolerance, and support taking. It can discover and respond to attacking pathogens focused around signs and cooperation among immune cells[14,15].

A machine-learning method called boosting to the issue of content arrangement. The fundamental thought of boosting is to consolidate numerous basic also reasonably erroneous order rules into a single, highly faultless classification guideline. The basic rules are prepared sequentially; conceptually, each one scheme is prepared on the cases which were most troublesome to classify by the preceding rules[16,17].

3. Proposed System

3.1 Feature Selection for Text

A few systems exist in the feature selection of text documents. Among those divided in this search are word stemming, stop positions, common data typical choice, selecting the ideal number of features, and three sorts of peculiarity vectors: Boolean, Term Frequency (TF), Term Frequency - Inverse Document Frequency (TFIDF)[18].

So as to classify new messages, a spam filter can examine them either as an independent instance, simply (for checking the vicinity of specific words if there arises an occurrence of keyword filtering) or in gatherings. Notwithstanding this, an AdaBoost filter investigates an accumulation of marked preparing information, and a filter which includes user cooperation gets additionally numerous userresults about a portion of the new messages for the analysis.

Figure 1 and Figure 2 represents the email message comprises of two parts, to be specific body and header. The message body is normally content in a common dialect, conceivably with HTML markup and graphical components. The header is organized situated fields, each one having a name, worth, and particular importance. Some of these fields, as from, to, or Subject, are standard, and others may rely on upon the product included in message transmission, for example, spam filters are introduced on mail servers. Subject field holds what the client sees as the subject of the message and is frequently treated as a part of the message body.

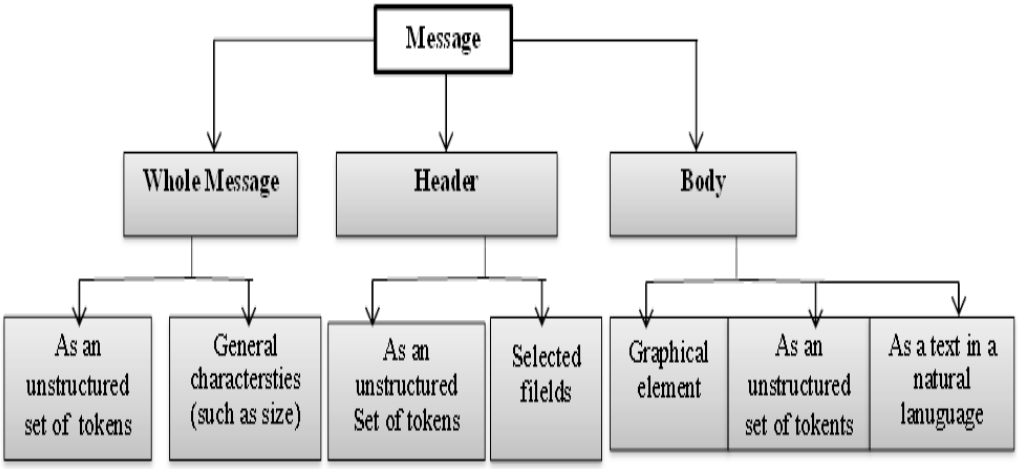

Figure 1. Email message structure.

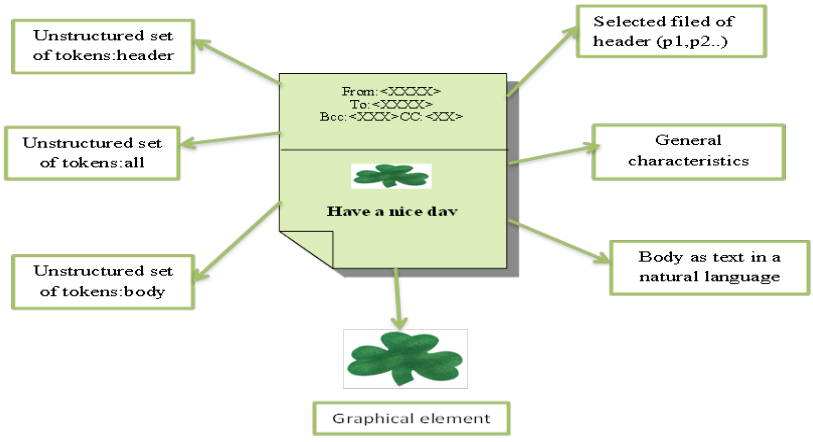

Figure 2. Feature selection from Message structure.

3.2 Feature Selection for Image

The Optical Character Recognition (OCR) technique is used for spam image filtering. Which divides the message content from the graphic pictures. After the dissemination of Adaboost filtering methods, the spammers received the utilization of picture spam. The content of an advertisement is set in a picture, with the goal that it is difficult to examine the message content with plain content based filter. This method prompts the filter requirements and focused around the picture dissection. In, picture based filtering the primary issue is to discover characteristics of both important and simple messages in order to concentrate, while grouping. So, typically most of the improved models are proposed to perceive the spam in pictures. Specifically, remove five features from the pictures, in a particular the division of the picture possessed by area is recognized as a content, color immersion and shade heterogeneity ascertains the content independent and non-message content.

3.3 System Architecture

The above Figure 3 describes the overall system of Email Spam filters. Which includes the spams based on text and images, Classification of text using text Classifier and the image classification using OCR framework, where the image values are filtered efficiently. These two filtering systems use the AdaBoosting classification and Hashing method to improve the identification of spam messages.

4. AdaBoosting

The Adaboost algorithm takes in a solid classifier by straightly joining (more straight forward) weak classifiers as per,

$$H(x) = \sum_{t} a_t h_t(x) \qquad (1)$$

Where $h_t(x)$ refers to a weak classifier. These weights are upgraded adaptively at every emphasis of Adaboost and assume an imperative part in deciding the mixture component for every weak classifier, i.e. $\{a_t\}$in equation 1.

At every cycle, Adaboost chooses the weak classifier that minimizes the weighted error,

$$\epsilon_t = \sum_{t} \omega_t \left[h_t x_t \neq y_y \right] \qquad (2)$$

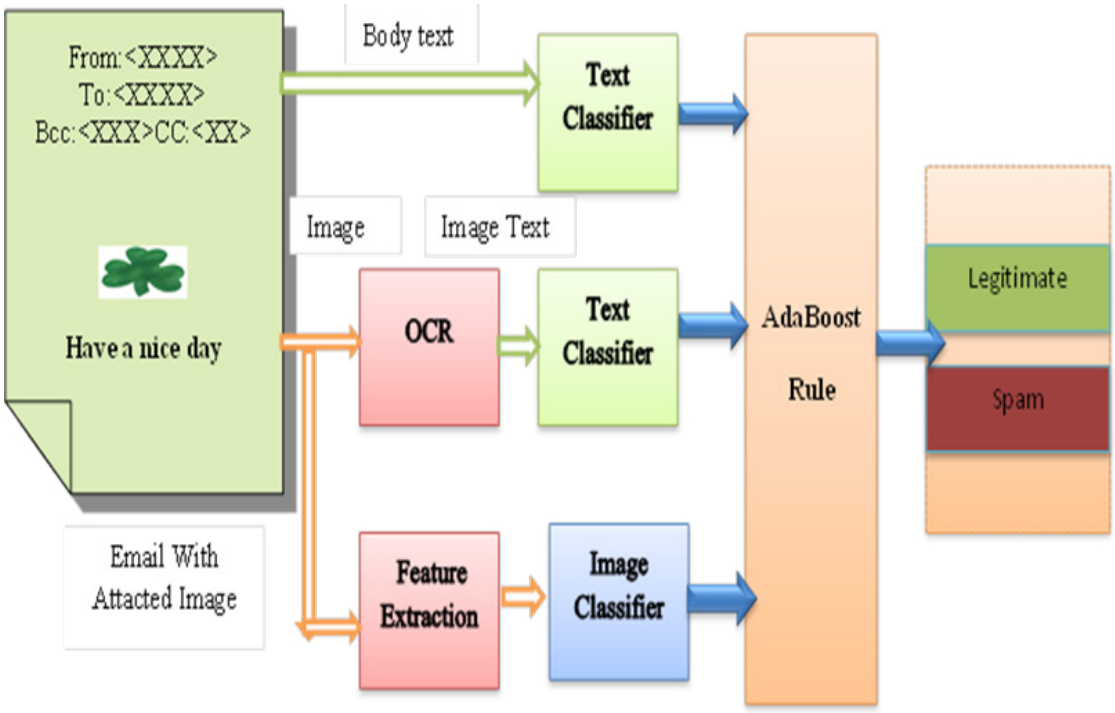

Figure 3. System architecture.

w_i is the weight and refers the output for input x_i . This error is calculated by the weight w_i on which the weak classifier is trained. The weight of each classifier is computed using this error

$$a_t = \frac{1}{2} In\left(\frac{1-\epsilon_t}{\epsilon_t}\right) \quad (3)$$

Therefore, the weights are updated with

$$w_i^{t+1} = \frac{w_i^t \exp(-a_t y_i h_t(x_i))}{Z_t} \quad (4)$$

Where w_i^t refers the weight of training sample x_i iteration t and Z_t is the normalization factor which is selected to the w^{t+1} will be a probability distribution.

Adaboost was the first adaptive boosting algorithm as it consequently conforms its parameters to the data based focused around the real execution in the current cycle.

4.1 Overview of the AdaBoost Algorithm

4.1.1 Initialization

Set the iteration counter m=0 and individual weights w_i for observations

$i = 1, \ldots, n$ to $w_i^{[0]} = \frac{1}{n}$.

4.1.2 Base-learner

Set m:=m+1 and compute the base-learner for the weighted data set:
Re-weight observations with

$$w_1^{[m-1]}, \ldots, w_n^{[m-1]} \xrightarrow{base-learner} \hat{h}^{[m]}(.)$$

4.1.3 Update Weights

Compute error rate and update the iteration-specific coefficient $a_m \rightarrow$ high values
For small error rates. Update individual weights

$w_i^{[m]} \rightarrow$ higher values if
Observation was misclassified.

4.1.4 Iterate

Iteration step 2 and 3 until m= m_{stop}

4.1.5 Final Aggregation

Compute the final classifier for a new observation x_{new} :

$$\hat{f}_{Adaboost}(x_{new}) = sign\left(\sum_{m=1}^{m_{stop}} a_m \hat{h}^{[m]}(x_{new})\right)$$

5. Parallelization

Parallelization is a simple method in order to improve the image classification using OCR, which classifies the image values, that the classification results are determined using the AdaBoosting approach and finally do filtering it is known as a hash-based detection mechanisms.

Hashing might be parallelized regularly for the tables work in parallel. The parallel adaptation of the file is outlined Figure 4. The inquiry is sent to all the table and further it is forwarded to the related bucket. Focuses in those buckets are then returned. Additionally, each one table can hold more focuses and consequently the file can manage bigger datasets. Furthermore, a single table could be further partitioned into two or more tables which still work in parallel and consequently might be conveyed on machines without huge memory.

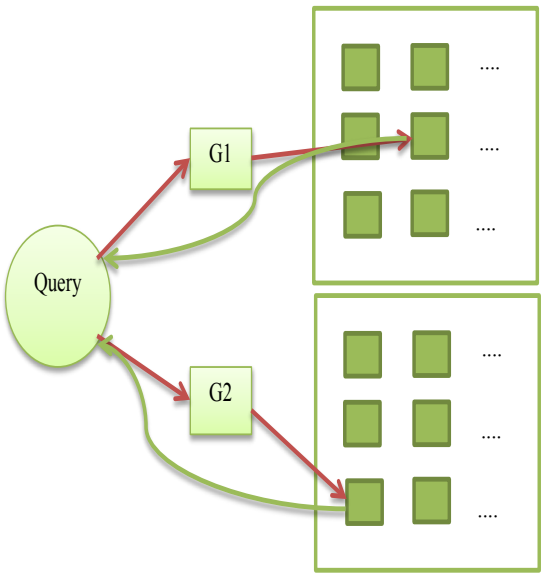

Figure 4. Parallelized hashing.

Boosting comparatively young, yet exceedingly powerful machine learning technique. The main idea behind the boosting algorithm is to association multiple weak learners. These finally have the form

$$h(x) = \begin{cases} c, & if\ x^i = 1 \\ -c, & else \end{cases} \quad (5)$$

Therefore $c \in \{1, -1\}$. These conclusion base takes binary features as an input, moreover support a positive classification select c = 1 or negative classification selects c = -1.

6. Performance Evaluation

6.1 Feature Selection

Table 1 shows the consequences of the fundamental feature selection tests, which tried the classifiers versus the amount of characteristics. These investigations utilized the baseline configurations, aside from the Adaboost classifier utilized 100 rounds. These results indicate no significant justification for utilizing more than 1000 attributes, so that was picked for the baseline configuration for every classifier in subsequent experiments.

Table 1. Spam recall at 1 percent false positive rate versus the number of attributes

Number of Attributes	Spam Recall (%) @ 1% FPR		
	Naive Bayes	SVM	AdaBoosting
50	75.3	74.1	82.9
100	86.2	93.2	93.5
300	93.4	93.3	95.2
500	96.6	95.2	96.9
1000	97.8	96.1	97.0
3000	97.2	95.3	96.8
5000	97.1	98.1	97.1
10000	98.7	92.7	96.8

Additionally looked at the baseline configurations against the firms that are utilized a case unfeeling feature selector. The results in Table 2 show no huge profit to changing over all tokens to lower case before building the word reference.

Table 2. Spam recall at 1% false positive rate versus case sensitivity

Configuration	Spam Recall(%) @ 1% FPR		
	Naive Bayes	SVM	AdaBoost
Baseline	95.5	94.2	95.7
Case-Insensitive	95.3	92.7	95.6

6.2 Feature Extraction

The baseline configuration of each classifier is compared with several parsing strategies. These results are shown in Table 3, Table 4, Table 5.

Table 3. Spam recall at 1 percent false positive rate versus inclusion of header names

Configuration	Spam Recall(%) @ 1% FPR		
	Naïve Bayes	SVM	AdaBoost
Baseline	93.5	92.6	94.2
Header Names	92.7	91.1	95.0

Table 4. Spam recall at 1% false positive rate versus use of "header tags"

Configuration	Spam Recall(%) @ 1% FPR		
	Naïve Bayes	SVM	AdaBoost
Baseline	93.7	95.3	97.1
Use Header Tags	92.5	92.9	95.0

Table 5. Spam recall at 1 % false positive rate versus inclusion of header values

Configuration	Spam Recall(%) @ 1% FPR		
	Naïve Bayes	SVM	AdaBoost
Baseline	93.6	91.3	95.9
Omit Headers	92.1	91.2	93.1

The three algorithms were additionally looked at on the premise of spam recall at diverse false positive rates. The result is shown in Table 6 and Figure 5.

Table 6. Spam recall at 1 % false positive rate versus corpus size

Number of messages	Spam Recall(%) @ 1% FPR		
	Naive Bayes	SVM	AdaBoost
50	91.7	92.2	38.5
100	94.3	93.1	63.9
200	93.5	93.3	83.4
400	96.1	96.0	93.5
800	95.5	96.1	95.1

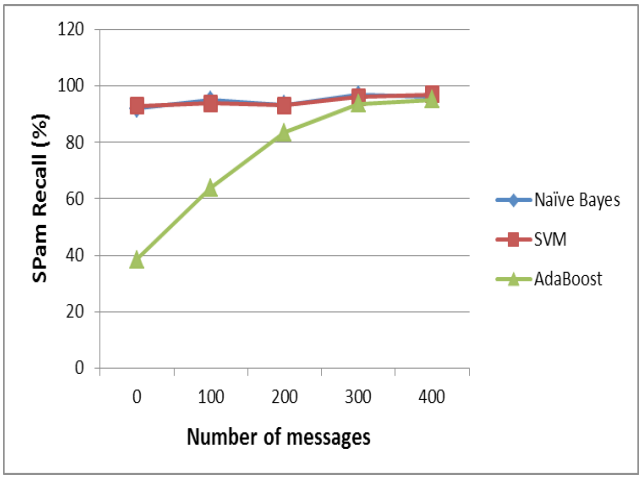

Figure 5. Spam recall at 1% false positive rate versus corpus size.

This Table 7 Figure 6 shows that while the contrasts between algorithm training and classification times are amazing, general, they are less contrasted with the message preprocessing undertakings that are the same regardless of the algorithm.

Table 7. Algorithm task execution times (per message)

Task	Execution Time (ms/message)		
	Naive Bayes	SVM	AdaBoost
Feature Selection	197	197	197
Parsing and Instance Mapping	35.3	35.3	35.3
Training	0.20	62.3	20.7
Classification	0.34	0.03	0.02

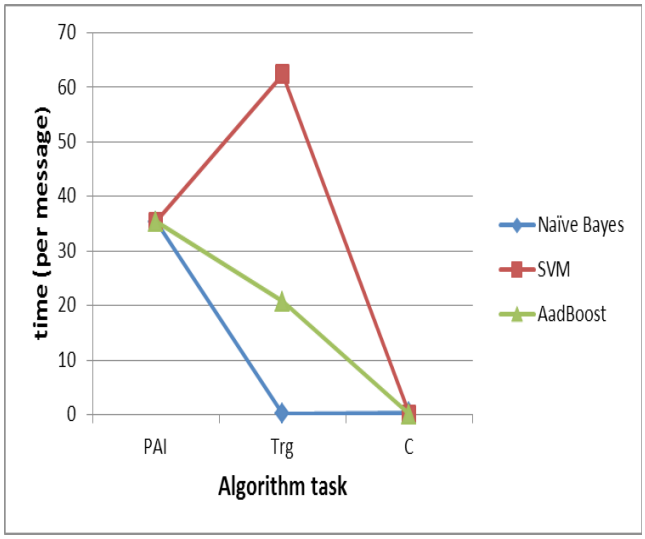

Figure 6. Algorithm task execution times (per message).

7. Conclusion

In this paper, the proposed methodology is implemented to consolidate the ideas from foundation learning into document representations for text document classification. An extremely fruitful troupe learning algorithm-Adaboost,

was proposed to perform the last classifications that are focused around the word vector representations and the reasonable peculiarities. Boosting Algorithms, when utilized with binary feature representations, scale well with an expansive number of measurements that normally happen when super ideas are also utilized. In the meantime, Adaboost is equipped for coordinating heterogenous features that are focused around distinctive ideal models without needing to change any parameters in the feature space representation. Analysis results, the demonstration in effectiveness, which makes it more practical for extensive scale requisitions. In addition the portrayed the parallelized kind of the index scheme. An aggregate Hash table and classifies the emails utilizing anabundant element of iterated words, the system can remove the attributes whether in header or the body. It is suggested to utilize the better and more proficient classifiers.

8. References

1. ShafighAski A. A proposed algorithm for spam filtering emails by hash table approach. International Research Journal of Applied and Basic Sciences. 2013; 4(9):2436–41.
2. Ikonomakis M, Kotsiantis S, Tampakas V. Text classification using machine learning techniques. Wseas Transactions on Computers. 2005 Aug; 4(8):966–74.
3. Mayr A, Binder H, Gefeller O, Schmid M. The evolution of boosting algorithms - from machine learning to statistical modeling. Methods of Information in Medicine. 28 Apr 2014; 53(6).
4. Beiranvand A, Osareh A, Shadgar B. Spam filtering by using a compound method of feature selection. Journal of Academic and Applied Studies. 2012 Mar; 2(3):25–31.
5. Khoshelham K, Nardinocchi C, Frontoni E, Mancini A, Zingaretti P. Performance evaluation of automated approaches to building detection in multi-source aerial data. ISPRS Journal of Photogrammetry and Remote Sensing. 2010; 65:123–33. doi:10.1016/j.isprsjprs.2009.09.005.
6. Fawcett T, Palo Alto CA. In vivo spam filtering: a challenge problem for KDD. ACM SIGKDD Explorations Newsletter. 2003 Dec; 5(2).
7. Sculley D. Online active learning methods for fast label efficient spam filtering. CEAS 2007 Fourth Conference on Email and AntiSpam; 2007 Aug 23.
8. Sarafijanovic S, Le Boudec J-Y. Artificial immune system for collaborative spam filtering. NICSO 2007. The Second Workshop on Nature Inspired Cooperative Strategies for Optimization; 2007 Nov 8-10; Acireale, Italy.
9. Rokach L. Ensemble-based classifiers. ArtifIntell Rev. 2010; 33:1–39. doi:10.1007/s10462-009-9124-7.
10. Andrej B, Cormack GV, Filipic B, Lynam TR, Zupan B. Spam filtering using statistical data compression models. Journal of Machine Learning Research. 2006; (7):2673–98.
11. Patil DR, Pattewar TM. A comparative performance evaluation of machine learning-based NIDS on benchmark datasets. International Journal of Research in Advent Technology. 2014 Apr; 2(2):2321–963.
12. Puri S, Gosain D, Ahuja M, Kathuria I, Jatana N. Comparison and analysis of spam detection algorithms. International Journal of Application or Innovation in Engineering & Management (IJAIEM). 2013 Apr; 2(4).
13. Uddin M, Alsaqour R, Maha A. Intrusion detection system to detect DDoS attack in gnutella hybrid P2P network. Indian Journal of Science and Technology. 2013 Feb; 6(2):71–83.
14. Tan Y, Mi G, Zhu Y, Deng C. Artificial immune system based methods for spam filtering. IEEE; 2013.
15. Anbazhagu1 UV, Praveen JS, Soundarapandian R, Manoharan N. Efficacious spam filtering and detection in social networks. Indian Journal of Science and Technology. 2014 Nov; 7:180–4.
16. Schapire RE, Singer Y. BoosTexter: a boosting-based system for text categorization. Machine Learning. 2000; 39:135–68.
17. Lee Y, Han DK, Ko H. Reinforced AdaBoost learning for object detection with local pattern representations. The Scientific World Journal. 2013: 14. Available from: http://dx.doi.org/10.1155/2013/153465
18. Tang H, Wu Jun, Lin Z, Lu M. An enhanced adaboost algorithm with naive Bayesian text categorization based on a novel re-weighting strategy. International Journal of Innovative Computing, Information and Control. 2010 Nov; 6(11).