ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

Bayesian SPAM Filtering with Optimization using SMO

Shashi Kant Rathore[1], Dr. Surendra Yadav[2]

Assistant Professor, Department of Computer Science & Engineering, Career Point University, Kota, India[1]

Associate Professor, Department of Computer Science & Engineering, Maharshi Arvind College of Eng. & Research Center, Sirsi, Jaipur, India[2]

ABSTRACT: This paper presents a hybrid Bayesian algorithm using reliable swarm intelligence to recognize and block SPAM emails. Due to low cost of communication & easily advertisement is possible with e-mails ,several people & companies use it to quickly distribute unsolicited bulk messages ,also called spam, to a large number of people in a very short time .Due to unnecessary traffic & security threats ,spam has become a major threat for not only business users to also network administrators & even to ordinary users. In addition to regulate, several technical solutions have been proposed and after a certain time it is found the content based on SPAM recognition is best suited in it. By including Bayesian rule (Bay's theorem) in content scanning, over all influence the popular e-mails for suitable recognition of SPAM mails can be increased worldwide.
But in this approach many limitations are proposed due to static values of probabilities for each token. For overcoming this limitation automated training is deployed for filtering purpose. for classifying and recognize the best tokens in SPAM can make more strong automated trained filter by including in it nature based optimization techniques SMO (spider monkey optimization)are best suited for this and have include in this paper.

KEYWORDS: SPAM, HAM, SMO, SMTP, SPF, ABC etc.

I. INTRODUCTION

The inventor which has influenced our lives & become a fast & popular means of communication in personal as well as in business communication is electronic mails.

In electronic communication emails are widely used in our social or professional life. Today most e mail systems are based on SMTP [1]. SPAMS: The emergence of internet has provided for every conceivable viewpoint where people can address with wide range of opinions on a single platform which is fast and easy to access. The internet provides the ability of emails system which are based on SMTP.SMTP which is the protocol in widespread use today,,

SMTP used for transporting emails over internet to various hosts in a short duration .The emails that are unsolicited and known as spam. Sometimes authentication of spam for internet users degrades the utility of emails.

As per spam and phishing statistics report 2014-16 , spam use is increased because it is cheapest source to send bulk messages to a large no. of people but the report also shows that the email traffic due to this in increased so far.

Spamming is the use of electronic messaging which usually varies from person to companies .email spam ,also known for unsolicited commercial email , junk mailer unsolicited commercial email frequently send commercially and distribute harmful contents as virus & there malware .this a common approach in social networking spam & a big threat for web economy & it causes great annoyance.

Bayesian spam filtering is the process introduced in this paper of using of newly naïve Bayesian classifier techniques

Naïve Baye's classifier [2] a simple probabilistic classifier that assumes the presence of a well-known class which is in common related to presence of any other feature which is in the Baye's theorem.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

Swarm intelligence [5] is the collective behavioring synthesis, natural or artificial systems in automation, natural or artificial systems in automation filters, and good result oriented performance for spam emails can be achieved by implementing it rather than Bayesian filtration for content based recognition approach.

The proposed research will also investigate the possibilities in this area, and the extent to which such `Recognition and Filtering' can and cannot be done. It is felt that while positive results of such research could be of great value to the network community, the attempt at uniting the methodologies of program verification and modeling/model checking is a worthwhile investigation in itself.

The main objectives of this Paper are as follows:

a) To design a proactive content based optimised Spam filter.

b) Proposed system will use the optimised Bayesian Filter using Spider Monkey optimisation.

c) Author will try to generate content based filtering that will also verify the email content based on Optimised Trained Data.

d) Optimisation of Trainee filter will be done by Spider monkey.

II. CONTENT BASED - BAYESIAN CLASSIFIERS

Rather than enforcing across-the-board policies for all messages from a particular email or IP address, content-based filters evaluate words or phrases found in each individual message to determine whether an email is Spam or legitimate.

In probability theory and statistics, Bayes' theorem (alternatively Bayes' law or Bayes' rule) describes the probability of an event, based on prior knowledge of conditions that might be related to the event.

Bayes' theorem is stated mathematically as the following equation:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Naive Bayes classifier [6] is a simple probabilistic classifier based on applying Baye's theorem with strong (naive) independence assumptions. In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature.

Bayesian Spam filtering (a form of e-mail filtering) is the process of using a Naive Bayesian classifier to identify Spam e-mail. Particular words have particular probabilities of occurring in Spam email and in legitimate email. For instance, most email users will frequently encounter the word "FREE" in Spam email, but will seldom see it in other email. The filter does not know these probabilities in advance, and must first be trained so it can build them up. To train the filter, the user must manually indicate whether a new email is Spam or not. For all words in each training email, the filter will adjust the probabilities that each word will appear in Spam or legitimate email in its database. Bayesian filtering uses a naive version of Bayes rule to compute the overall probability that an e-mail is a Spam assuming that the individual probabilities of each token appearing in a Spam are independent of one another. The overall probability that the new e-mail is a Spam is then computed using

$$p = \frac{ab}{ab-(1-a)(1-b)}$$

In above approach, we can improve the searching, if we use optimisation for searching probability of email features/Tokens.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

III. OPTIMIZATION

Optimization is an art of selecting the best alternative(s) amongst a given set of options. The process of finding the largest or the smallest possible value, which a given function can attain in its domain of definition, is known as optimization. Population-based optimization algorithms find near-optimal solutions to the difficult optimization problems by being motivated from nature.

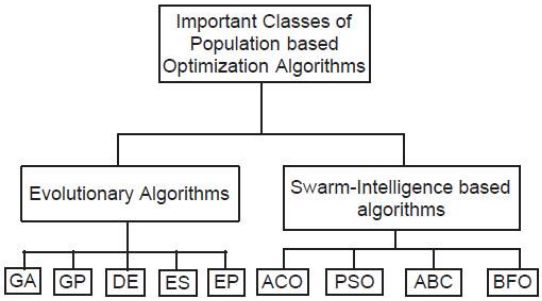

Figure: Population Based Optimization Algorithms

This section briefly presents a few examples of scientific and engineering swarm intelligence studies [11]

Clustering Behaviour of Ants

Ants build cemeteries by collecting dead bodies into a single place in the nest. They also organize the spatial disposition of larvae into clusters with the younger, smaller larvae in the cluster centre and the older ones at its periphery. This clustering behaviour has motivated a number of scientific studies. Scientists have built simple probabilistic models of these behaviours and have tested them in simulation (Bonabeau et al. 1999) [7].

Flocking and Schooling in Birds and Fish:

Flocking and schooling are examples of highly coordinated group behaviours exhibited by large groups of birds and fish. Scientists have shown that these elegant swarm-level behaviours can be understood as the result of a self-organized process where no leader is in charge and each individual bases its movement decisions solely on locally available information: the distance, perceived speed, and direction of movement of neighbours. These studies have inspired a number of computer simulations (of which Reynolds' Boids simulation program was the first one) that are now used in the computer graphics industry for the realistic reproduction of flocking in movies and computer games [6].

Particle Swarm Optimization

In particle swarm optimization (PSO), a set of software agents called particles search for good solutions to a given continuous optimization problem [7]. Each particle is a solution of the considered problem and uses its own experience and the experience of neighbor particles to choose how to move in the search space. PSO has been applied to many different problems and is another example of successful artificial / engineering swarm intelligence system.

Artificial Bee Colony Optimization

ABC is a new swarm intelligence algorithm proposed by Karaboga in 2005, which is inspired by the behavior of honey bees. Since the development of ABC, it has been applied to solve different kinds of problems. One of the swarms exists in nature is honey bee swarm which follows collective intelligent manner, while searching the food [8]. The honey bee swarm has many qualities like bees can communicate the information, can memorize the environment, can store and share the information and take decisions based on that. With reference to ABC, the potential solutions are food sources of honey bees. The fitness is determined in terms of the quality (nectar amount) of the food source. There are two fundamental processes which derive the evolution of an ABC population: the variation process, which enables exploring different areas of the search space and the selection process, which ensures the exploitation of the previous experiences.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

Spider Monkey Algorithm:
SMO is a subclass of swarm intelligence, proposed by Dr. Harish Sharma and Dr. Jagdish Chand Bansal., in the year 2014 [4]. a new approach for optimization is proposed by modeling the social behavior of spider monkeys. Spider monkeys have been categorized as fission-fusion social structure based animals. Initially work in a large group and based on need after some time, they divide themselves in smaller groups led by an adult female for foraging. There-fore, the proposed strategy broadly classified as inspiration from the intelligent foraging behavior of fission-fusion social structure based animals.

SMO technique is based on population repetitive methodology. It consists of seven steps [10]. Each step is written below:

I. Initialization of Population:
II. Local Leader Phase (LLP):
III. Global Leader Phase (GLP):
IV. Global Leader Learning (GLL) Phase:
V. Local Leader Learning (LLL) Phase:
VI. Local Leader Decision (LLD) Phase:
VII. Global Leader Decision (GLD) Phase:

IV. RESEARCH METHODOLOGY

Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature.[6] A database of words as shown in below table, will be used, which are mostly comes in Spams. This database will be created from survey of various Spam detailing sites and with consumers' consult. Then each word of the incoming email will be checked with the database, and the local probability of each word of the email will be calculated.

Probabilities Using Static Bayesian Filter
Following is the data base of the tokens/words with their static Probability[4], normally founds in Spam Emails. By which, Spam filters can be triggered:

S.No	Feature	P(f)	S.No	Feature	P(f)
1	4U	0.8	16	Father	0.3
2	Accept	0.5	17	Mother	0.3
3	Act	0.4	18	Cashcash	0.9
4	Additional	0.4	19	Casino	0.9
5	Addresses	0.3	20	Check	0.4
6	All	0.4	21	Claims	0.8
7	Amazing	0.8	22	Limited	0.7
8	Apply	0.7	23	Lower	0.6
9	Auto	0.9	24	Marketing	0.7
10	Avoid	0.4	25	Month	0.3
11	Billion	0.9	26	Name	0.3
12	Brand	0.6	27	Urgent	0.4
13	Cable	0.5	28	Congratulations	0.6
14	I	0.3	29	Dear	0.5
15	We	0.2	30	Different	0.3

Table: Sample Spam's Words with static Probabilities

Probabilities Using Automated Bayesian Filter
In this approach, Dynamic Probabilities have been used rather than static values of Probabilities for Spam database. Probabilities of features will be calculated automatically using trained filter, which is used by Bayesian Filter to calculate overall probability of email.

To compute the probability P for a feature f, the following formula is used [5]:

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

$$P_f = \frac{\frac{s}{t_s}}{\frac{s}{t_s} + \frac{kn}{t_n}}$$

s is the number of occurrences of feature f in Spam messages, t_s is the total number of Spam messages in the training set, n is the number of occurrences of feature f in non-Spam messages, and t_n is the total number of non-Spam messages in the training set. k is a number that can be tuned to reduce false positives by giving a higher weight to number of occurrences of non-Spam features.

S. No	Feature (f)	S	n	$\frac{s}{t_s}$	$\frac{kn}{t_n}$	$\frac{s}{t_s}+\frac{kn}{t_n}$	P_f
1	4U	300	2	0.6	0.02	0.62	0.97
2	Accept	200	123	0.4	0.82	1.22	0.33
3	Act	100	123	0.2	0.82	1.02	0.2
4	Additional	233	122	0.47	0.82	1.29	0.37
5	Addresses	233	123	0.47	0.82	1.29	0.37
6	All	433	156	0.87	1.04	1.91	0.46
7	Credit	445	13	0.89	0.09	0.98	0.91
8	Apply	432	133	0.87	0.89	1.76	0.5
9	Card	433	122	0.87	0.82	1.69	0.52
10	Avoid	212	211	0.43	1.41	1.84	0.24

Table: Sample Spam's Words with Dynamic Probabilities

Example for Spam

Consider an email having following sentence:

"Paying too much for Credit Card? Now, you have a chance to receive a FREE TRIAL!"

Tokenization

Features will be extracted from each sentence of email:

Paying	Credit	Card?	chance	receive	FREE	TRIAL!

Bayesian Classifier

Firstly filter tokenizes the whole email in small words. The individual probabilities of each word appearing in a Spam are independent of one another. These probabilities have been checked from database as mentioned above:

Paying	Credit	Card	chance	receive	FREE	TRIAL!
0.39	0.91	0.52	0.56	0.3	0.88	0.96

The overall probability that the new e-mail is a Spam is then computed as following method.

The combined probabilities for three tokens with probabilities a, b, c would be computed:

$$p = \frac{abc}{abc - (1-a)(1-b)(1-c)}$$

And so on [3]. Following table shows the features with their calculated probabilities:

Features	Notation	Probability
Paying	a	0.39
Credit	b	0.91
Card	c	0.52
chance	e	0.56
receive	f	0.3
FREE	g	0.88
TRIAL!	h	0.8

Table: Bayesian Notations for Spam Features

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

Bayesian classification is an approach of text classification, which searches the textual content of an email and uses algorithms to identify Spam email. If this probability is more than a certain level (pre defined) then that email is considered as a Spam.

For the above example overall probability for this email has been calculated and it is 0.88. that is more than our threshold value i.e 0.67. So the email containing above sentence will be recognized as Spam.

V. RESULT ANALYSIS

For the analysis purpose research has been applied to some group of 10 emails and as per above mentioned proposed method following results has been shown:

Confusion matrix of result using Static Values of probabilities:
The following table gives the confusion matrix containing True Positive, False Positive, True Negative, False negative and accuracy when we use the static probability of features in detection of SPAM emails.

n=50	Accuracy =78%		
PREDICTED	**Spam**	**Ham**	**Precision**
Spam	19	5	89%
Ham	6	20	82%
Recall	80%	90%	0.79

Confusion matrix of result using Dynamic Values of probabilities:
The following table gives the confusion matrix containing True Positive, False Positive, True Negative, False negative and accuracy when we use the dynamic probability of features using automated Bayesian trained filter in detection of SPAM emails.

n=50	Accuracy =90%		
PREDICTED	**Spam**	**Ham**	**Precision**
Spam	21	2	92%
Ham	4	23	85%
Recall	84%	92%	0.88

Confusion matrix of result using Optimized Values of probabilities:
The following table gives the confusion matrix containing True Positive, False Positive, True Negative, False negative and accuracy, when we use the static probability of features in detection of SPAM emails.

n=50	Accuracy =94%		
PREDICTED	**Spam**	**Ham**	**Precision**
Spam	23	1	96%
Ham	2	24	89%
Recall	90%	95%	0.93

Comparison:
The following table gives the comparison of accuracy in SPAM detection, when we use the static probabilities, Dynamic probabilities and optimised probabilities of features in detection of SPAM emails individually.

Email groups	**Result using static Values**			**Result using Auto mated**			**Result using optimization**		
	Spam Precision	**Spam Recall**	**Accuracy**	**Spam Precision**	**Spam Recall**	**Accuracy**	**Spam Precision**	**Spam Recall**	**Accuracy**
1st Group of 10 emails	0.5	0.5	0.6	1	0.8	0.9	0.5	0.8	0.9
2nd group of 10 emails	0.5	0.6	0.5	0.67	0.8	0.7	0.5	0.8	0.8
3rd Group of 10 emails	0.75	0.6	0.7	0.75	0.6	0.7	0.75	1	0.9
4th Group of 10 emails	0.6	0.6	0.6	1	0.6	0.8	0.6	0.8	0.9
5th Group of 10 emails	0.6	0.5	0.5	1	0.8	0.9	0.6	0.8	0.9

Table: Result Comparison

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

VI. CONCLUSION

This work provides the design of a new anti Spam system. By identifying existing algorithms, and developing new algorithms that will lead to more accurate initial solutions and robust outlier rejection (to reduce the probability of finding false minima). I will compare different approaches using simulated data to determine the best approach.

Result shows that Dynamic Bayesian Filter using automated trained filter more powerful than static Bayesian filter and by using optimization in choosing probabilities for the tokens, We can enhance the accuracy of detecting Spam emails.

REFERENCES

[1] Castillo G. (2006) Adaptive Learning Algorithms for Bayesian Network Classifiers. PhD Thesis. Chapter 3.3. Naïve Bayes Classifier
[2] Domingos, P. and Pazzani M. (1997) On the optimality of the simple Bayesian classifier under zero-one loss, Machine Learning, 29(2-3):103–130
[3] Duda, R.O. and Hart, P. E. (1973) Pattern Classification and Scene Analysis. John Wiley & Sons, Inc.
[4] John, G.H. and Langley P. (1995) Estimating Continuous Distributions in Bayesian Classifiers. Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. pp. 338-345. Morgan Kaufmann, San Mateo.
[5] Mitchell T. (1997) Machine Learning, Chapter 6. Bayesian Learning, Tom Mitchell, McGrau Hill
[6] Tan, P., Steinbach, M., Kumar, V. (2006) Introduction to Data Mining . Chapter 5: Classification- Alternative Techniques. Section 5.1 (pag 166-180) , Addison Wesley Higher Education
[7] Basu, B. and Mahanti, G.K. (2011) 'Artificial bees colony optimization for synthesis of thinned mutually coupled linear array using inverse fast Fourier transform', in Devices and Communications (ICDeCom), 2011 International Conference on, IEEE, pp.1–5.
[8] Baykasoglu, A., Ozbakir, L. and Tapkan, P. (2007) 'Artificial bee colony algorithm and its application to generalized assignment problem', Swarm Intelligence: Focus on Ant and Particle Swarm Optimization, pp.113–144.
[9] Benala, T.R., Jampala, S.D., Villa, S.H. and Konathala, B. (2009) 'A novel approach to image edge enhancement using artificial bee colony optimization algorithm for hybridized smoothening filters', in Nature & Biologically Inspired Computing, NaBIC 2009, World Congress on, IEEE, pp.1071–1076.
[10] Urvinder Singh, Rohit Salgotra, and Munish Rattan. A novel binary spider monkey optimization algorithm for thinning of concentric circular antenna arrays. IETE Journal of Research, pages 1–9, 2016.
[11] Xin-She Yang. Nature-inspired metaheuristic algorithms. Luniver press, 2010.