ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 12, December 2014

A Victimization Optical Back Propagation Technique in Content Based Mostly Spam Filtering

Dr.Kathir.Viswalingam[1], G.Ayyappan[2]

Dean (R&D), Bharath University, Chennai, India[1]

Assistant Professor, Department of Information Technology, Bharath University, Chennai, India[2]

ABSTRACT: Spam is email sent in bulk wherever there's no direct agreement in situ between the recipient and also the sender to receive email solicitation. to forestall the delivery of this spam, an automatic tool named a spam filter is employed. during this paper, (OBP) "Optical Back Propagation" technique is employed as an automatic tool to spot whether or not a message is spam or not supported the content of the message. Spam-based dataset-a dataset from UCI "University of California, Irvine" machine learning repository, is employed as coaching and testing dataset to coach the network so tested it. The samples of this dataset ought to be first of all preprocessed (normalization or feature choice before normalization) to be appropriate to the network. The results OBP spam filtering is affordable in term of accuracy, precision, recall, false Positive, false negative, and speed of net.

KEYWORDS: Neural Network, Optical Back Propagation, Spam, Spam Filter.

I. INTRODUCTION

Electronic mail (email) is associate degree economical variety of communication that has become wide adopted by each people and organizations. Today, a lot of and a lot of folks square measure relying on e-mail to connect them with their friends, family, colleagues, customers and business partners. sadly, as email usage has evolved, thus too has its threats, especially spam, that is additionally called unsought bulk email or junk, has become associate degree progressively troublesome threat to sight and is being delivered in implausibly high volumes [1].

Spam may be a major problem that doubtless threatens the existence of e-mail services. especially, it's currently a non-trivial task to seek out legitimate e-mails in associate degree e-mail inbox untidy with spam. Spam is additionally a chic drawback that prices service suppliers and organizations billions of bucks per annum in lost information measure. additional to the information measure value, it's additionally calculable that every piece of spam prices a corporation one greenback in lost worker productivity [2].

There square measure many approaches that try and stop or scale back the massive quantity of spam on people. These approaches embrace legislative measures like anti-spam laws over world-wide. different techniques square measure called Origin-Based filters that square measure supported victimization network data and information science "Internet Protocol" addresses so as to sight whether or not a message is spam or not. the foremost common techniques square measure the filtering techniques trying to spot whether or not a message is spam or not supported the content and different characteristics of the message [3]. This paper presents a spam filtering technique supported the content of the message to tell apart whether or not a message is spam or not. the remainder of this paper is structured as follows.2 presents the analysis publications that cowl spam filtering. Then, offers some background on (OBP) "Optical back propagation". Section four presents the planned methodology for determination the spam drawback. The results of the planned OBP spam filtering square measure evaluated in section five. Finally, Section six concludes the given work.

10.15680/ijircce.2014.0212037

II. RELATED WORK

Several makes an attempt within the literature are urged for determination the matter of the spam. These are: In [4], associate degree investigation the impact of applying a lot of sophistication to lower layers within the filtering method, specifically extracting data from e-mail is given. Many styles of obfuscation were mentioned that were turning into ever a lot of gift in spam so as to do confuse and circumvent this filtering processes. The results obtained by removing sure styles of obfuscation show to enhance the classification method. 2 classifiers were used (K-NN) "K- Nearest Neighbor" with 3 neighbor (k=3) thus it referred to as (3-NN), and “Bayesian”. In [5], a neural network (NN) approach is applied to the classification of spam. They found out that NN configuration can have the simplest performance and least error to desired output. They thought of that NN that was trained victimization fifty seven email parameters made rock bottom range of misclassifications.

In [6], associate degree reconciling metaphysics is employed to seek out associate degree economical spam email filtering methodology. Four classification methods: NN, (SVM) “Support Vector Machine” classifier, (NB) “Naïve Bayes” classifier, and J48 classifier were evaluated the results supported totally {different | completely different} datasets and different options. In [7], the techniques concerned within the style of the spam filters that embrace NB, SVM, NN, and (CBART) “Classifier supported Bayes Additive Regression Tree” square measure mentioned. They discuss the effectiveness and limitations of applied mathematics filters in filtering out varied styles of spam from legitimate e-mails.

In [3], a modification on (ANN) "Artificial Neural Network" within the input layers is applied which permit the input layers to be modified over time and to exchange useless layers with new promising layers that offer promising results. They referred to as their work (CLA_ANN) “Continuous Learning Approach Artificial Neural Network” and use a developed perception learning algorithmic program approach. Their modifications on CLA_ANN offer promising results that would be employed in the method of fighting against spam.

In [8], the foremost standard machine learning methods: theorem, K-NN, ANNs, SVMs, (AIS) "Artificial immune system" and (RS) "Rough Sets" classification square measure reviewed. They applied 2 procedures within the preprocessing stage. Stopping: is used to get rid of common word and Case-change: is used to alter the (Body) into little letters. The experiment is performed with the foremost frequent words in spam email; they choose one hundred of them as options.

In [9], a brand new technique for filtering spam is given. The technique consisted of one perception that was designed to be told and distinguish legitimate and illegitimate causing server parameter values and messages.

In [10] a spam filtering system victimization (HMMs) “Hidden mathematician Models” and ANN to separate out spam wherever word obfuscation on the keyword is conducted to evade detection. the utilization of hidden mathematician models is to capture the applied mathematics properties of spam variants happiness to identical category. the utilization of artificial neural network increased performance activity of the filtering system particularly on the flexibility of the system to be told a lot of from any new spam messages that entered the system. The spam filter classified email basing on spam keyword, thus an inventory of common spam keyword was gathered from this corpus for testing the dataset.

III. OPTICAL BACK PROPAGATION

OPB may be a variety of back propagation algorithmic program. This methodology has been applied to the supervised learning (a machine learning paradigm for feat the input-output relationship data of a system supported a given set of paired input-output coaching samples) for multi-layer (NN) neural networks (consist of input layer, hidden layer, and output layer). it's most frequently used as coaching algorithmic program. The OBP algorithmic program is intended to beat a number of the issues related to normal (BP) “Back-Propagation”. One among the vital properties of this algorithmic program is that it will break loose native minima with high speed of convergence throughout the coaching amount. The convergence speed of the educational method may be improved by adjusting the error, which is able to be transmitted backward from the output layer to every unit within the intermediate layer [11]. this type of algorithmic program used for coaching method that depends on a multilayer NN with a really little learning rate, particularly once employing a massive coaching dataset size [12]. In BP, the error at one output unit is outlined as in equation (1). Wherever is that the desired output, and is that the actual output of the sample. While the error at one output unit in adjusted OBP are going to be as in equation two.

IV. METHODOLOGY

In this paper, the OBP technique is employed to filter incoming email. OBP algorithmic program is applied on coaching dataset, the dataset is chosen willy-nilly from publically obtainable datasets within the (UCI) machine learning repository it referred to as spam- based mostly dataset. every sample within the dataset may be a fifty eight attributes (57of them square measure from the content of email and one attributes may be a binary label talk to the category of email, zero for legitimate email and one for spam).

V. STANDARDISATION AND OPTIONS CHOICE

Before applying the algorithmic program, the dataset ought to be pre processed to rework messages into an even format that may be understood by the neural networks, the pre processing is finished in 2 stages: standardization method and also the second is options choice method.

The standardization method is applied on values of the dataset to line them in uniform vary. The vary wont to set the dataset in it's(0, 1). The feature choice method is applied by use (PCA) “Principe element Analysis” technique to pick out specific options (attributes) from spam-based dataset as a result of the dataset contains lots of redundancy, wherever choosing specific attributes from those fifty seven eliminate the redundancy and keep solely vital attributes. PCA results a dataset with thirty one attributes. Algorithmic program one clarifies however PCA is applied on spam based mostly dataset.

VI. OBP FOR SPAM FILTERING

After pre processing method is finished, the OBP is employed to coach the neural network on set of samples referred to as coaching samples from spam-based dataset to get the corresponding adjustment weights required to provide the proper output. The input sample (A) is given to the input layer of the network. These inputs square measure propagated through the network till they reach the output units. This passing play produces the particular or foretold output sample. The particular output(y) is computed in step with equation four.

Because optical back propagation may be a supervised learning algorithmic program, the required outputs square measure given as a part of the coaching vector. associate degree OBP uses operate to calculate the error signal, and there square measure 2 variety of error signal as a result of the operate continuously returns or values as shown in equation (2).

This error signal is then the premise for the optical back propagation step, whereby the errors square measure passed back through the neural network by computing the contribution of every hidden process unit and etymologizing the corresponding adjustment required to provide the proper output. The affiliation weights square measure then adjusted and also the neural network has simply “learned” from associate degree expertise. The error signal terms of the output layer (5)

OBP Spam Filtering analysis 2 totally different structures for OBP square measure used counting on whether or not the PCA was applied or not. The primary structure named OBP Structure-1 and also the other is termed OBP Structure-2. The results square measure obtained once setting OBP parameters; To conduct the experiment, the spam based mostly dataset is split into 2 components coaching dataset and testing dataset. The coaching dataset is employed to regulate the weights whereas testing dataset is employed to the performance of the planned OBP spam filtering technique during this paper, 2500 samples square measure hand-picked willy-nilly as coaching dataset and five hundred random samples square measure hand-picked as testing dataset. Four experiments square measure conducted. the primary and second experiments (Exp1) and (Exp2) conducted on same coaching dataset victimization OPB Structure-1 parameters

10.15680/ijircce.2014.0212037

and OBP Structure-2 parameters severally. The third experiment Exp3) and fourth experiment (Exp4) deals with OBP Structure-1 and OBP Structure-2 severally once applied on testing dataset. Depicts the speed of the neural network in step with range of iteration and also the error signal price of the 2 structures of OBP. For OBP Structure-1, the amount of iterations is nine and also the error signal price was three.525764 at the ninth iteration, and also the OBP Structure-2 has sixty iterations and also the error was seven.913597 at the sixtieth iteration. The Performance analysis for OBP Spam Filtering Technique

Measure	Exp1	Exp2	Exp3	Exp4
Accuracy	0.99	0.949	0.948	0.916
P	0.988	0.948	0.996	0.912
R	0.992	0.949	0.928	0.92
FP	0.012	0.051	0.032	0.088
FN	0.007	0.05	0.072	0.08

OBP Structures

VII. CONCLUSION AND FUTURE WORK

In this paper I created a study regarding artificial system (AIS) and that i found completely different application areas of AIS. A number of them embrace classification and bunch, optimisation, learning, image process, AI etc. Among these classifications and bunch is most generally used, therefore I created a study regarding completely different papers, that square measure used in this space. In these papers classification and bunch build use of necessary options of AIS such as feature choice, pattern recognition and machine learning.

REFERENCES

1. S. Yuvaraj and M. Suguna, "An Effective Defense against Compromised Machines by SAS Worm Detection,""International Journal of engineering and Management analysis, 2013.

2. T. Choi, Transactional Behavior based mostly Spam Detection, MSC. Thesis, Department of Systems and laptop Engineering Carleton University Ottawa, Ontario, North American nation Gregorian calendar month 2007.

3. A. T. Sabri, A. H. Mohammad, B. Al-Shargabi, and M. A. Hamdeh, " Developing New Continuous Learning Approach for Spam Detection victimization Artificial Neural Network (CLA_ANN)," European Journal of research project ISSN 1450-216X Vol.42 No.3 (2010).

4. M. Davy, Feature Extraction for Spam Classification, MSC thesis, Department of engineering, University of Dublin, Trinity school Sep, 2004.

5. D. Puniškis, R. Laurutis, and R. Dirmeikis, "An Artificial Neural Nets for Spam e-mail Recognition," physics and engineering science, Nr. 5 (69), 2006

6. S. Youn and D. McLeod, "Spam Email Classification victimization associate degree reconciling metaphysics," Journal of computer code, Vol.2,No.3, Sep 2007.

7. M. T. Banday and T. R. Jan, "Effectiveness and Limitations of applied mathematics Spam Filters," International Conference on New Trends in Statistics and improvement, 2009.

8. W.A. Awad, And S.M. Elseuofi, "Machine Learning ways for Spam E-Mail Classification," International Journal of engineering &amp; data Technology (IJCSIT), Vol. 3, No. 1, Feb. 2011.

9. O. Kufandirimbwa, and R. Gotora, "Spam Detection victimization Artificial Neural Networks (Perception Learning Rule)," on-line Journal of Physical and biological science analysis, ISSN 2315-5027; Vol. 1, Issue 2, pp. 22-29. June 2012.

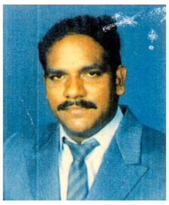
10. D. Ndumiyana, "Hidden mathematician Models and Artificial Neural Networks for Spam Detection," International Journal of Engineering analysis &amp; Technology (IJERT) Vol. 2 Issue 4, Gregorian calendar month - 2013 ISSN: 2278-0181.

11. P. Mehtani, and A. Priya, Pattern Classification victimization Artificial Neural Networks, BSC thesis, Department of Computer Science and Engineering National Institute of Technology Rourkela, Orissa, India, 2011.

12. N. A. Hamid, N. M. Nawi, R. Ghazali, and M. N. M. Salleh, "Improvements of Back Propagation algorithmic program Performance by Adaptively dynamic Gain, Momentum and Learning Rate," International Journal on New laptop Architectures and Their Applications (IJNCAA) 1(4): 866-878 The Society of Digital data and Wireless Communications.

10.15680/ijircce.2014.0212037

BIOGRAPHY



Dr.Kathir.Viswalingam is working as Dean (R&D) at Bharath University. He got his Ph.D in chemical Engineering in the year 1982 from university of madras. He is having an experience of total 32 years in Teaching, Research and Industry. He has published more than 50 research Articles, 5 Text Books and More Manuals. He is the Executive Editor of Scientific refereed journal namely “Indian Journal of applied Sciences and Innovative Technology” with ISSN: 2321-7790. His areas of interest are Information Technology, Environmental Engineering, Power Plant Engineering, Bio Technology and Reaction Engineering and Innovative Research Projects.

G.Ayyappan is an Assistant professor of the department of Information Technology in Bharath Institute of Science and Technology, Deemed University, Chennai (TN) India. He is having an experience of more than 6yearsin Teaching. His main areas of interest are Social Networks and data mining and their application.

10.15680/ijircce.2014.0212037