Spam Filtering using K mean Clustering with Local Feature Selection Classifier

Anand Sharma
M-Tech Scholar
Dept. of Info. Tech.
IET, Alwar(Raj)
Alwar(India)

Vedant Rastogi
Associate Professor
Dept. of Info. Tech.
IET, Alwar(Raj)
Alwar(India)

ABSTRACT

In this paper, we present a comprehensive review of recent developments in the application of machine learning algorithms to Spam filtering, focusing on textual approaches. We are trying to introduce various spam filtering methods from Naïve Bias to Hybrid methods for spam filtering, we are also introducing types of filters recently used for spam filtering along with architecture of spam filter and its types .In this paper we are proposing a technique using Local feature classification methods with K mean clustering algorithm in classifier, for spam filtering term selection we are using Document frequency method, for feature extraction we are using bag of words model for classification we are using k-mean clustering method along with local concentration based extraction of content. This method gives good results along with all parameters.

Keywords

Spam filtering, K mean.

1. INTRODUCTION

In last few years, the increasing use of e-mail has led to the appearance and further acceleration of problems due to unsolicited bulk e-mail messages, commonly called to as Spam. Evolving from a small irritation to a major concern, given the high circulating volume and not proper content of some of these messages, Spam is start to reduce the reliability of e-mail. Personal users and organizations are affected by Spam with respect to the use of network bandwidth utilized receiving these messages and the time wasted by users classifing between Spam and normal (legitimate or ham) messages. A business model depending on Spam marketing is typically advantageous because the costs for the sender are less, it tends large number of messages can be sent, this aggressive behavior being one of the major characteristics of Spammers. An additional approach acquired is the use of Spam filters, which is based on study of the message contents and additional information, effort to classify Spam messages. The action to be in use once they are identified usually depends on the setting in which. If used as a client side filter it is embed only on client system and classify mails labbled as spam of legimate. Where as another is server side filer it is present on mail serves provider server handling various messeges as spam and send to respective user.

However Spammers are always trying to crack the filters to send there messages as spam filters based on user defined rules, based on observed content regularly used by spammers in messages in such messages Spammers then began sending the messages by changing the certain terms that are very regularly used in spam messages (eg. by writing congratulations as Congragulation, Gift as G_I_F_T). on an attempt spam filter cannot correctly find such words.

If we learn statics from commtouch Q1 internet threats tread reports the number of daily spam emails extensively exceeded the 100 billion mark (117.8 billion) in March 2013, and also Spam levels doubled in between time from December 2012 to March 2013, a 98% increase. Phishing also increased dramatically, with the number of phishing emails growing to more than 74% in March, compared to the previous December. To overcome the problem of spam filtering many machine learning techniques have been proposed from naïve Bayes (NB) ,k-nearest neighbour method (KNN) support vector machine (SVM), Artificial immune system (AIS) and Hybrid methods of spam filtering etc. All methods are working as classification method in which email is classified as spam or legitimate mail. Email Classification task involve following steps term selection, feature extraction and last classification.

2. ARCHITECTURE OF SPAM FILTER

Generally email is divided into three parts Header, Subject line, Body where as Header contains the information of sender, receiver etc. Subject Line. In which subject the email is and last field is content or body of email this is actual content of email.

Before the available information can be used by a classifier in a filter, appropriate pre-processing steps are required.

The steps involved in the extraction of data from a message are illustrated in figure below and can be grouped into:

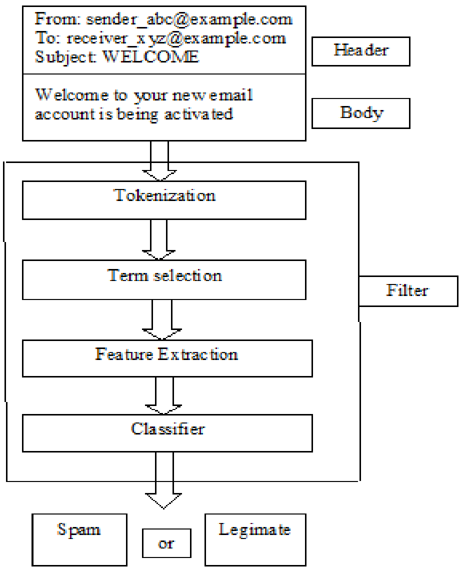
From: sender_abc@example.com
To: receiver_xyz@example.com
Subject: WELCOME
Header
Welcome to your new email account is being activated
Body
Tokenization
Term selection
Filter
Feature Extraction
Classifier
Spam
or
Legimate

Fig: Architecture of spam filter

(1) Tokenization, in which the words get extracted from the message body.

(2) Term Selection, which rank each term according to there usefulness.

(3) Feature Extraction, which gives reduced set of data

In this paper we are using document frequency as our tern selection method where as Bag of words is our feature extraction based on local feature extraction and k mean clustering to make two clusters of spam and legitmate respectively

3. LITERATURE REVIEW

3.1 Term Selection Method

As we have discussed the main role of term selection is to rank the terms according way they classify terms in different categories

3.1.1 Information Gain (IG)

Information gain is used for ranking of features based 0n type they classify the messages in to classes as spam,legimate etc. Information gain work on principle of entropy. The entropy I of a given dataset S is given as

$$I(S) = -\sum_{i=1}^{x} pi\ log2\ (pi)$$

Where x denotes the total number of classes and pi the feature that fit in to class i. The decrease in entropy or the information gain is computed for every attribute A constant with

$$IG(S,A) = I(S) - \sum_{z \in A} \frac{|SA,z)|}{|S|} I(S_{A,}z)$$

Where z is a value of A and SA, z is the set of instances where A has value z.

3.1.2 Document Frequency (DF)

Document frequency is another very simple method for feature selection. It is calculated by calculating number of documents in which a term/feature occurs .In line with the DF method, the terms which are having their DF below from a predefined threshold such terms are removed from the set of terms. The DF of a term can be calculated as follows:

$$DF(t_i) = |\{m_j \,|m_j \in M, and\ t_i \in m_j\}|$$

Where M is complete set of training messages and m_j is message in M.

3.2 Feature Extraction Methods

This is second stepping of spam filter in this the set of data or terms which is extracted in term selection this set in minimized. For feature extraction following techniques has been used.

3.2.1 Bag of word model

Bag of word model is one in all the widely used feature extraction method in spam filtering applications. It converts a message to a d-dimensional vector [x1, x2…... xd] by considering occurrence of already selected terms, that are chosen by utilizing term selection method. Within the vector, xi will be viewed as a function of the term ti's occurrence within the message.

3.3 Spam Filtering Techniques

3.3.1 Naive Bias Classifier

Naive Bayes classifiers are a mostly used technique of e-mail filtering. This method involve bag of words features to classify spam e-mail, this technique commonly used in text classification. Naive Bayes classifiers use Bayesian assumption to calculate a probability that an email is or is not spam. In Baysian framework the messages get divided in to different representations then the probability that given representation of message, denoted as representation of a message, denoted as x'=(x1,x2,x3,….,xn),belongs to class c is given by:

$$P(c/x) = [P(x/c)P(c)]/(P(x))$$

Where,

P(c/x) is probability of term to be occur in message

P(x) is probability of term to be occur in message

This probability can be used to make the decision about if a message should belong to the category. In order to compute this probability, estimations about the message in the training set are made. When computing probabilities for spam classification we can prevent the denominator because we only have 2 classes (spam and legitimate), and one message cannot be classified into only one of them, so denominator is the same for every set of words can be estimated as the number of message in the training set belonging to the category, divided by the total number of documents.

Calculating P(x/c) is a bit more complicated because we need in the training set some messages identical to the one we want to classify. When using Bayesian algorithm it is very frequent to find the assumption that terms in a message are independent and the order they appear in the message is irrelevant.This way probability can be calculated as

$$P\left(\frac{x}{c}\right) = \prod_{i=1}^{y} P(ti/c)$$

Where, y the number of terms considered in the message representation and ti is i th word in message[1][6]

3.3.2 Support Vector Machine

Support Vector Machines (SVM) is very accepted in spam classification taking into consideration the accuracy with these algorithms. SVMs are defined as an algebraic equation generating maximum margin hyper-plane to different training instances, with polynomial kernels. Training instances not required to be linearly separable. The main objective is to construct a hyper-plane for separation. The basic form of hyper-plane can be generated as a linear function of the attributes. SVMs are fast to learn and highly effective in learning-based spam filters.[1][4][7]

3.3.3 K Nearest Neighbors

Another technique in learning algorithms involves storing training data sets after being pre-processed and represented. therefore, when a new datasets needs to be classified, it is compared with saved documents and assigned the more suitable category with respect to its similarity to those stored in each class. The most popular one in this category is k-Nearest Neighbors (kNN). The value of k designates the number of neighbours used for classification. A important step of this method is the choice of similarity function between messages. The method frequently used to calculate the similarity measure between messages is the "cosine distance", where cosine is defined as the angle between the vectors representing the compared messages. This distance function normalizes the length of the messages, and hence considered efficient.[4]

3.3.4 Artificial Neural Network

It is based on pattern recougnisation each message can be classified according to pattern among these patterns the messages is selected as spam or legimate. First neural network should be trained. This training contains analysis of message content of large samples of both spam and legimate messages.[1]

To create training sets of spam and legimate emails, email are carefully reviewed according to simple, definition of spam. Although the average user normally considers all unwanted emails as "spam". Similarly, non-spam email should be avoided for personal email communications between users, and limit any forms of bulk mailings, regardless of whether they were solicited or not. Once these sets have been collected and approved, the neural network is ready for training .By using various statistical methodologies the unique words for both spam and legimate class are identified e.g. congratulations, prize, luckey,winner these words for spam class respective words for legimate mails also identified.

Now Artificial neural network training sets are ready to use first words from messages is identified ignoring the phrases like “to,the,for” which is present in all emails, then the identified words checked whether they present in spam trained sets or legimate trained set. If Artificial neural network seems message to report as spam then that mail classified as spam otherwise in legimate mail box. [5][9]

3.3.5 Artificial Immune system

Artificial Immune System It assigns a weight to each detector, which is incremented or decremented when it recognizes an expression in a Spam legitimate or message, with the thresholded sum of the weights of the matching detectors being used to determine the classification of a message. The system can be corrected by either incrementing or decrementing the weights of all the matching detectors. Using a personal corpus and 100 detectors,generated mostly from SpamAssassin heuristics, true positive and negative rates of 90% and 99%, respectively, were obtained. It was concluded that the proposal achieved acceptable results considering the small number of detectors used.[8][15]

3.3.6 Hybrid Method

One of the latest approaches in spam filtering is hybrid filtration system which is a combination of different algorithms, especially if they use unrelated features to produce a solution. In this case it can be applied various filtering techniques and get the advantages of the traditional and learning-based methods

4. PROPOSED METHOD

4.1 Background

As we know process of spam filtering contains three steps. So here for term selection we term frequency method. Basically information gain method gives usefulness of particular term or feature. In our experiment first we have to create two local dictionary one for spam and other for legitimate email. In second step we apply Local Concentration (LC) based feature extraction method, in which we divide our message into different local area. So in feature extraction we have a new feature plane also known as local area feature vector. Lastly we apply our classifier to this feature vector. In this way we classify our message as spam or legitimate one.[13]

4.2 Term Selection

To use the LC feature extraction method in process of spam filtering. In structure of LC based structure the tokenization is well used step where message is get divided into terms by finding blank spaces during term selection. In tokenization step the message is get divided in huge number of terms where as this is training phase in our model we are using term frequency model for term selection.

The algorithm for information gain works in five steps

```
1. Initialize already existing set as null set
2. For terms selected by tokenization step
For
Calculate importance of the term according to a certain term selection methods;
End for
3. Arrange the terms according to descending order of the importance;
4. Add the front % terms to the preselected set;
5. For every term is selected set do
if || P(tk|cl) –P (tk|cs) || >α, α >=0 then
if || P(tk|cl) –P (tk|cs) || > α, α >=0 then
Add the term to DSs;
else
Add the term to DSl;
endif
else
Discard the term;
endif
endfor
```

Considering already selected term as source calculate tendency from equation

Tendency (tk) = P (tk|cl) - P (tk|cs)

Where P (tk|cl) is the probability of tk’s occurrence, given mail as legitimate e-mail, and P (tk|cs) is the probability of tk’s occurrence predicted in spam. The tendency tk give the difference between the frequencies as spam or legimate. From the above algorithm the terms frequently occurred in legimate mail are added in Document set DS1, and terms frequently occurred in spam messages as Document set DS2.

4.3 Feature Extraction

This step is next step in email classification here we are going to Local concentration model for feature extraction.To create an LC-based feature vector for every message, a sliding window of wn-term length is used to slide over the message with a step of wn-term, by which we sure that there is no gap nor overlap between adjacent windows. At each movement of the window, a spam genes concentration SCj and a legitimate genes concentration LCj are calculated according to the two DS and the terms in the sliding windows as follows[13]

$$SC_j = N_s/N_t$$

$$LC_j = N_l/N_t$$

Where

Nt number of different terms in window.

Ns number of different terms in window matched with Document set of spam DS1.

Ni is number of different terms in window matched with document set of legitmate DS2.[13]

It can be given by algorithm below

1. Move a sliding window of wn-term length over a given message with a step of wn -term;

2. for every position of the sliding window

do

2.1 Calculate the spam genes concentration of the window according to equation for calculating number of spam by (SCj = Ns/Nt);

2.2. Calculate the legitimate genes concentration of the window as per equation (LCj = Nl/Nt);

end for.

Construct the feature vector likes:

$(<SC_1, LC_1>, <SC_2, LC_2> \ldots <SC_n, LC_n>)$

4.4 Classification

As per the process of spam filtering first terms get extracted by means of tokenization then through LC based extraction the frequency of words in datasets is calculated then this results feed as input to classifier. In our approach we are going to use K mean clustering algorithm for classification of message either in spam or in legimate.

K-Means clustering algorithm, collect the extracted terms according to their feature values into K number clusters. Objects exists in a same cluster have similar feature values. K is any positive value that determines the number of clusters and is identified at the inception of the execution of the algorithm; centroid can be calculated by taking mean of rank of all the terms in cluster

This algorithm for k means clustering works in 5 steps as

1) Label the number of clusters.

2) Describe K different centroids for every cluster. This work is done by randomly separating objects into K clusters, shaping their centroids, and obeserving whether these centroids are different from each other.

3) Iterate over all objects to determine the distance of every object to the centroid of that cluster. Each object is assigned to the cluster of the nearest centroid.

4) Again calculate the centroids of new clusters.

5) Continue step 3 until centroids doesn't change any longer.

4.5 Performance Measuring

In spam filtering, so many evaluation methods or parameters have been designed for comparing performance of different filters [1]. We evolved by four evaluation criteria (i.e. spam recall, spam precision, accuracy, and Fβ measure), in our experiments to evaluate parameter values and do a assessment between the LC approach and some prevalent approaches. Among the criteria, accuracy and Fβ measure are important, for accuracy measures the total number of messages correctly classified, and Fβ is a mixture of spam recall and spam precision.

1) Spam Recall: It calculates the percentage of spam that can be classified by a model. High spam recall ensures that the classifier can protect the users from spam effectively. It is defined as follows:

Rs = (12)

Where the number of spam is correctly classified, and is the number of spam mistakenly classified as legitimate e-mail.

2) Spam Precision: It calculate how many messages, classified as spam, are truly spam along with amount of legitimate e-mail by mistake classified as spam. The higher the spam precision is, the smaller amount legitimate e-mail have been incorrectly classified. It is defined as follows:

Ps = (13)

3) Accuracy: To some level, it can reflect the all around performance of filters. It identify the percentage of correctly classified. It is defined as follows:

A = (14)

Where the number of legitimate e-mail properly classified.

4) Fβ measure: It is a combination of Rs and Ps. It is another overall performance of filter its formula calculate is defined as follows:

Fβ = (1+β2) ( ) (15)

5. EXPERIMENTAL RESULTS AND COMPARISION

We conducted experimental on four different data sets which is publically available such as PU1,PU2,PU3,PUA and we are getting results along with comparison with SVM Local Concentration based classifier [13] whereas K mean represent K mean clustering vector like classifier.

Table 1 comparison between results of SVM classifier and proposed classifier

Corpus	Approach	Precision %	Recall %	Accuracy %	F_1 %	Feature Dimension
PU1	SVM-LC-VL	96.49	93.66	96.15	95.01	7
	K mean cluster	97.59	97.44	99.07		7
PU2	SVM-LC-VL	92.31	85.71	95.78	88.88	7
	K mean cluster	95.67	85.71	97.18	92.31	7
PU3	SVM-LC-VL	95.93	91.30	96.10	93.55	7
	K mean cluster	97.16	90.96	94.67	93.36	7
PUa	SVM-LC-VL	95.49	94.09	95.02	94.20	7

	K mean cluster	98.21	96.49	97.37	97.35	7

6. CONCLUSION

From the above comparison of results for spam filtering classifier we can say that the results of classifier using k mean clustering algorithm is better than the different aspects as we have carried out through experiment using PU1,PU2 these datasets.

7. REFERNCES

[1] Thiago S. Guzella *, Walmir M. Caminhas "A review of machine learning approaches to Spam filtering", Elsevier Expert Systems with Applications 36 (2009) 10206–10222

[2] Enrique Puertas Sanz, José María Gómez Hidalgo, José Carlos Cortizo Pérez Email Spam Filtering Universidad Europea de Madrid Villaviciosa de Odón, 28670 Madrid, SPAIN.

[3] Saadat Nazirova," Survey on Spam Filtering Techniques" Communications and Network, 2011, 3, 153-160 doi:10.4236/cn.2011.33019 Published Online August 2011 Scientific Research.

[4] Meghali Das1 and Vijay Prasad, "ANALYSIS OF AN IMAGE SPAM IN EMAIL BASED ON CONTENT ANALYSIS," International Journal on Natural Language Computing (IJNLC) Vol. 3, No.3, June 2014.

[5] Alia Taha Sabri Adel Hamdan Mohammads, Bassam Al-Shargabi, Maher Abu "HamdehDeveloping New Continuous Learning Approach for spam Detection using Artificial Neural Network (CLA_ANN)," European Journal of Scientific Research ISSN 1450-216X Vol.42 No.3 (2010), pp.511-521

[6] Meharn Shami,Susan Dumais, David Hekerman, Eric Horvitz "A Bysean Appraoch to Filtering Junk e mail "Microsoft Research.

[7] Shugang Liu & Kebin Cui "Applications of Support Vector Machine Based on Boolean Kernel to Spam Filtering" Modern Applied Science ccsenet journal Volume 3, No, 10 October 2009

[8] Andrew Secker, Alex A. Freitas, Jon Timmis "AISEC: an Artificial Immune System for E-mail Classification" Evolutionary Computation, 2003. CEC '03. The 2003 Congress on (Volume:1) 2003

[9] Chris Miller, Group Product Manager Enterprise Email Security "Neural Network-based Antispam Heuristics" Symantec Enterprise Security

[10] Surendra Kumar Rakse , Sanyam Shukla "Spam Classification using new kernel function inSupport Vector Machine " (IJCSE) International Journal on Computer Science and Engineering, Vol. 02, No. 05, 2010, 1819-1823

[11] Qiang Wang Yi Guan Xiaolong Wang SVM-Based Spam Filter with Active and Online Learning

[12] Saadat Nazirova "Survey on Spam Filtering Techniques" Communications and Network, 2011, 3, 153-160 Published Online August 2011

[13] Yuanchun Zhu and Ying Tan "A Local- Concentration-Based Feature Extraction Approach for Spam Filtering" IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, VOL. 6, NO. 2, JUNE 2011

[14] Ion Androutsopoulos,, Paliouras, G., & Michelakis, E. "Learning to filter unsolicited commercial e-mail," Tech. rep. 2004/2, NCSR ''Demokritos".

[15] A. Ishiguru, Y. Watanabe, and Y. Uchikawa, "Fault Diagnosis of Plant Systems using Immune Networks," Proc. IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems, pp. 34–42, 1994.