Bayesian Statistical Analysis for Spams

Youcef Begriche
Institut Telecom, Telecom ParisTech
Youcef.Begriche@telecom-paristech.fr

Ahmed Serhrouchni
Institut Telecom, Telecom ParisTech
Ahmed.Serhrouchni@enst.fr

***Abstract*—This paper presents a Bayesian statistical analysis applied to the spam problem. In most anti-spam related research, generally it is assumed that the probability of a spam occurrence is equal to 0.5, which is in our opinion unrealistic. It is also assumed that in the spam message, words are considered as an independent family of words. This makes us look at how the posterior probability behaves when the a priori probability is different from 0.5 and derive the consequences of the assumption of independent words on the posterior probability. The first assumption pushes us to define a prior and find a posterior probability laws to enhance the spam detection and increase the reliability decision. This analysis differs from previous results, that used the Bayesian approach to the anti-spam issue, especially through refinement and enhancement of various probability laws.**

***Keywords* – Conditional density, Distribution attachment, Binomial law, Spam($\mathcal{S}$), Ham($\mathcal{H}$), Bayesian statistical model, Classification.**

I. Introduction

This article aims to contribute to the study of anti-spam methods by improving the application of Bayes theorem. The negative effects of spam on information systems have been widely described in [1]. Several research works and efforts are being made to reduce its scale and impact. Solutions have been developed but are still not sufficient for a total eradication. These solutions are based on several methods of treatment, and can be found in [2]. Many of these methods require a specific network infrastructure or cooperation between network elements (e.g. mail server directories) based on particular organizations that require specific costs. Anti-spam Bayesian methods operate as "standalone" and are based on statistical and probabilistic mathematical models. These methods benefit from an important research in the field of processing language. Indeed, fundamental results in this field have been obtained [3] and make it possible to apply it to anti-spam field. Paul Graham is probably the pioneer who suggested a complete solution to filter messages based on Bayesian mathematical method [6]. Several classification-based solutions also exist for which a summary can be found in [2]. The author proposes three main classes :

- Filtering the content amounts to distinguish the spam (noted $\mathcal{S}$) from the non spam (called ham and noted $\mathcal{H}$). As mentioned in [2], this type of process is prone to errors, and subject to false negatives [1] and false positives [2]. The goal is to reduce the number of false positives. For this purpose there are several filtering methods. Heuristic, adaptive or Bayesian filtering, collaborative and honey pot.
- The identification of one or more fields of the message. Clasical cryptographic functions may be used (digital signature and authentication). This method is effective against phishing [3] or any approach of spoofing. Several approaches related to this category are deployed
- Establishing a cost for message transmission. This method reduces a significant amount of resources at spammer side. Several approaches based on costs exist.

Because of their limitations when they are used separately, administrators and users sometimes combine several of these methods. Paul Graham [4] and Gary Robinson [5] worked on messages content by assuming the probability of a message being a spam equal to that of a ham (Pr (spam) = Pr (ham) = 0.5) and all classifiers have used this formula worked under the same assumption.

In this paper, we focus on the generalization of the Bayes' theorem for the application of anti spam taking into account that the probabilities mentioned above are different.

In [6], some naive Bayes classifiers are proposed : multi-variate Bernoulli NB [1], multinomial NB, Boolean attributes, multi-variate Gauss NB.

The classifiers are built on Bayes theorem : if a message is represented by a vector $\overrightarrow{x} = (x_1, \ldots, x_n)$, then the probability that the message belongs to the class c (spam($\mathcal{S}$) or ham($\mathcal{H}$) is given by the formula:

$$P(c/\overrightarrow{x}) = \frac{P(c).P(\overrightarrow{x}/c)}{P(\overrightarrow{x})} \tag{1}$$

As there are two classes (spam and ham) the denominator is written :

$$P(\overrightarrow{x}) = P(\overrightarrow{x}/\mathcal{S}).P(\mathcal{S}) + P(\overrightarrow{x}/\mathcal{H}).P(\mathcal{H}) \tag{2}$$

The criterion for classifying a message as spam is given by :

$$\frac{P(\mathcal{S}).P(\overrightarrow{x}/\mathcal{S})}{P(\overrightarrow{x}/\mathcal{S}).P(\mathcal{S}) + P(\overrightarrow{x}/\mathcal{H}).P(\mathcal{H})} > T \tag{3}$$

Where T is a fixed threshold. The criterion for classifying is adapted for each classifier :

1) For the multi-variate Bernoulli NB, the criterion becomes :

$$\frac{P(\mathcal{S}).\prod_{i=1}^{m} P(t_i/\mathcal{S})^{x_i}.(1–P(t_i/\mathcal{S}))^{(1-x_i)}}{\sum_{c\in\{\mathcal{S},\mathcal{H}\}} P(c).\prod_{i=1}^{m} P(t_i/c)^{x_i}.(1–P(t_i/c))^{(1-x_i)}} > T \tag{4}$$

[1] Naive Bayes

where :

- $x_i = 1$ or 0 if the token t_i occurs in the message.
- $P(t/c) = \frac{1+M_{t,c}}{2+M_c}$ where $M_{t,c}$ is the number of training messages of category c that contain token t, while M_c is the total number of training messages of category c.

2) For the multinomial NB, the criterion of classifying a message as spam is :

$$\frac{P(\mathcal{S}).\prod_{i=1}^{m} P(t_i/s)^{x_i}}{\sum_{c\in\{\mathcal{S},\mathcal{H}\}} P(c).\prod_{i=1}^{m} P(t_i/c)^{x_i}} > T \tag{5}$$

3) For the multi-variate Gauss NB, the criterion is :

$$\frac{P(\mathcal{S}).\prod_{i=1}^{m} g(x_i;\mu_{i,s},\sigma_{i,s})}{\sum_{c\in\{\mathcal{S},\mathcal{H}\}} p(c).\prod_{i=1}^{m} g(x_i;\mu_{i,s},\sigma_{i,s})} > T \tag{6}$$

Where :

$$g(x_i;\mu_{i,s},\sigma_{i,s}) = \frac{1}{\sigma_{i,s}\sqrt{2\Pi}} e^{-\frac{(x_i-\mu_{i,s})^2}{2\sigma_{i,s}^2}} \tag{7}$$

and the mean $(\mu_{i,s})$ and typical deviation $(\sigma_{i,s})$ of each distribution are estimated from the training data.

In those classifiers, $p(\mathcal{S})$ and $p(\mathcal{H})$ are typically estimated by dividing the number of training messages of category (s or h) by the total number of training messages.

In [7], the criterion for classifying a message as spam is :

$$\frac{P(C=\mathcal{S}/\overrightarrow{X}=\overrightarrow{x})}{P(C=\mathcal{H}/\overrightarrow{X}=\overrightarrow{x})} > \lambda \tag{8}$$

Where :

$$P(C{=}c/\overrightarrow{X}{=}\overrightarrow{x}){=}\frac{P(C{=}c){\cdot}\prod_{i=1}^{n} P(X_i{=}x_i/C{=}c)}{\sum_{k\in\{\mathcal{S},\mathcal{H}\}} P(C{=}k){\cdot}\prod_{i=1}^{n} P(X_i{=}x_i/C{=}k)} \tag{9}$$

For this classifier, $P(C = \mathcal{S})$ and $P(C = \mathcal{H})$ are also estimated by by dividing the number of training messages of category (spam or ham) by the total number of training messages.

In [8], authors work on the same formula (1)(Bayes theorem) as follows :

$$\begin{aligned} P(\mathcal{S}/\overrightarrow{x}) &= \frac{P(\mathcal{S}).P(\overrightarrow{x}/\mathcal{S})}{P(\overrightarrow{x})} {=} \frac{P(\mathcal{S}).P(\overrightarrow{x}/\mathcal{S})}{P(\overrightarrow{x}/\mathcal{S}).P(\mathcal{S}){+}P(\overrightarrow{x}/\mathcal{H}.P(\mathcal{H})} \\ &= \frac{\frac{P(\overrightarrow{x}/\mathcal{S})}{P(\overrightarrow{x}/\mathcal{H})}P(\mathcal{S})}{\frac{P(\overrightarrow{x}/\mathcal{S})}{P(\overrightarrow{x}/\mathcal{H})}P(\mathcal{H}) + P(\mathcal{H})} \end{aligned} \tag{10}$$

In this paper, they worked only on the ratio $\frac{P(\overrightarrow{x}/\mathcal{S})}{P(\overrightarrow{x}/\mathcal{H})}$ and $P(C = \mathcal{S})$ and $P(C = \mathcal{H})$ are also estimated by dividing the number of training messages of category (spam or legitime) by the total number of training messages.

In [9], authors use the following criterion :

Given an input document d, its target class (spam, ham) can be found by choosing the one with the hig posterior probability.

$$H(d){=}\arg\max_{c_j\in C} \sum_{w\in F} \frac{P(w/c_j)P(c_j)}{\sum_{c'\in C} P(w/c')P(c')} P(w/d')CHI(w,c_j) \tag{11}$$

where F is text feature vector; $P(w/c_j)$ is a ratio of frequency of w in d of the total number of words, $CHI(w, c_j)$ is χ^2 statistic of word w and c_j which measures the lack of independence between word w and c_j.$P(c_j)$ is the number of documents with class label c_j divided by the total number of documents.

In [10], authors determine the maximum a posterior (MAP) class by calculating :

$$e_{MAP} = \arg\max_{e_i\in E} P(e_i) \prod P(k_j/e_i)$$

where k_j is the word found in the j^{th} position in the unseen query. e_i is the tutorial example/page.

$$P(e_i) = \frac{Number\ of\ queries\ in\ Q\ with\ e_i}{Number\ of\ queries\ in\ Q} \tag{12}$$

Where Q is the queries set.

In [11,12,13], the degree of confidence is introduced $W_S^{NB}(\overrightarrow{x})$ that $\overrightarrow{x}$ is spam by :

$$W_S^{NB}(\overrightarrow{x}){=}P(\mathcal{S}/\overrightarrow{x}){=}\frac{P(\mathcal{S}){\cdot}\prod_{i=1}^{m} P(x_i/\mathcal{S})}{\sum_{k\in\{\mathcal{S},\mathcal{H}\}} P(k){\cdot}\prod_{i=1}^{m} P(x_i/k)} \tag{13}$$

and $P(C = \mathcal{S})$ and $P(C = \mathcal{H})$ are also estimated by dividing the number of training messages of category (spam or legitime) by the total number of training messages.

In [14], authors deliver a discussion about the implementation of Binomial Distribution and Poisson Distribution in Bayesian spam filter, to calculate the probability of a mail being spam, containing words that are not already stored in a database (i.e., encountered by the filter for the first time). They use :

- Binomial random variable with parameters n and p, then the probability function of R is given as follows :

$$f(r){=}P(R){=}\mathrm{C}_n^k \cdot p^r q^{n-r}, r{=}0,1,2,3,\ldots,n \tag{14}$$

where $p + q = 1$.

- Poisson random variable R, the probability function of R is then :

$$f(r){=}P(R){=}e^{-m}m^r/r!, r{=}0,1,2,3,\ldots,n, \tag{15}$$

where $m = np$

p is the probability to have a spam word, which is assumed as a constant constant from trial to trial (0.4 for the Binomial distribution and 0.04 for the Poisson distribution). // [15] defines a local probability. For example :

$$P_{local-\mathcal{S}}=0.5+\frac{(N_{\mathcal{S}}-N_{\mathcal{H}})}{C_1\cdot(N_{\mathcal{S}}+N_{\mathcal{H}}+C_2)} \quad (16)$$

Each word feature generates one such local probability. These local probabilities are then used in a Bayesian chain rule to calculate an overall probability that an incoming text is spam. The Bayesian chain rule is :

$$P(in\ class/feature)=\frac{P(feat/inclass)\times P(in\ class)}{P(feat/in\ class)\times P(in\ class)+P(feat/not\ in\ class)\times P(not\ in\ class)}$$

which is applied iteratively to chain each local probability feature into an overall probability for the text. Here $P(in\ class)$ is also estimated by dividing the number of training messages of class (spam or legitimate) by the total number of training messages.

In [16], they introduce a score notion as follows. If a mail is represented by $\overrightarrow{x}=(x_1,\ldots,x_n)$ then :

$$score(\overrightarrow{x})=\log P(\mathcal{S})+\sum_k \log P(x_k/\mathcal{S}) - (\log P(leg)+\sum_k \log P(x_k/leg)) \quad (17)$$

Therefore, if $score(\overrightarrow{x})>0$, the e-mail will be a spam, and legitime otherwise.
$P(C=\mathcal{S})$ and $P(C=\mathcal{H})$ are also estimated by dividing the number of training messages of category (spam or legitime) by the total number of training messages.
This work differs from previous works listed above; in this work, we give the generalization of the Bayes' theorem applied to spam and its proof. This generalization gives a relationship between the posterior probability, the prior probability and the size n of the sample that characterized the message. This relationship allows to find the value of the posterior probability for any value of the prior probability and all values of n.
We also explain and compare some results found previously, ie when the prior probability is equal to 0.5.

II. Explication, formula, proof

Afterwards, spam and ham will be denoted respectively $\mathcal{S}$ and $\mathcal{H}$

A. Explication

If A and B are two events, the Bayes'formula :

$$P(A/B)=\frac{P(B/A).P(A)}{P(B)} \quad (18)$$

gives the conditional probability of A knowing B
This formula is applying to anti-spam as follows:
If a certain message is represented by the vector $\overrightarrow{x}=(x_1,\ldots,x_n)$, the probability that this message is a spam is calculated as follows :

$$P(\mathcal{S}/x_1,\ldots,x_n)=\frac{P(x_1,\ldots,x_n/\mathcal{S})\cdot P(\mathcal{S})}{P(x_1,\ldots,x_n/\mathcal{S})\cdot P(\mathcal{S})+P(x_1,\ldots,x_n/\mathcal{H})\cdot P(\mathcal{H})} \quad (19)$$

Here, in previous works authors assume that $P(\mathcal{S})=P(\mathcal{H})=0.5$.

B. Formula

Taking into account that these probabilities are different, we propose the following formula and its proof is given

$$P(\mathcal{S}/x_1,\ldots,x_n)=\frac{P(\mathcal{S}/x_1)\cdot\ldots\cdot P(\mathcal{S}/x_n)}{P(\mathcal{S}/x_1)\cdots P(\mathcal{S}/x_n)+[\frac{P(\mathcal{S})}{(1-P(\mathcal{S}))}]^{(n-1)}\cdot P(\mathcal{H}/x_1)\cdot\ldots\cdot P(\mathcal{H}/x_n)} \quad (20)$$

III. Method of Calculation

A. programmatically

Using the MATLAB software, we wrote a program that allows us, using the formula to calculate whether a given message is spam. Moreover, the probabilities $P(\mathcal{S}/x_i)$ involved in the formula are numbers randomly drawn by the MATLAB software.

B. Classification criteria:

When a threshold α is fixed, a messge is classified as spam when:

$$P(\mathcal{S}/x_1,\ldots,x_n)\geq\alpha$$

IV. Results and Interpretation

The formula (23) binds the posterior probability $P(\mathcal{S}/x_1,\ldots,x_n)$, the prior probability ps and the size of the sample n.
By fixing n and considering in this formula the posterior probability $P(\mathcal{S}/x_1,\ldots,x_n)$ as a function, denoted f, the prior probability ps, so the derivative of this function is equal to:

$$\frac{-(n-1)\cdot\mathcal{E}\cdot(P(\mathcal{S}))^{(n-2)}}{(1-P(\mathcal{S}))^n\cdot[\mathcal{E}+(\frac{P(\mathcal{S})}{(1-P(\mathcal{S}))})^{(n-1)}\cdot\mathcal{F}]^2} \quad (21)$$

This derivative is negative then the function cited above is necessarily decreasing, therefore the two probabilities prior and posterior vary inversely, when the prior probability ps increases from 0 to 1 posterior probability $P(\mathcal{S}/x_1,\ldots,x_n)$ decreases from 1 to 0 . This shows in figures 1 and 2

Figure 1 contains a family of histograms. For each value of the prior probability ps, we have a histogram which gives the value of the probability that the message represented by $(x_1,\ldots,x_n)$ is a spam, noted $P(\mathcal{S}/x_1,\ldots,x_n)$, by varying the size of the sample n. It may be noted that this was only observed from 0.5 sticks below 0.1, that is to say that the posterior probability is less than or equal to 0.1.
Figure 2 contains a family of curves. For each value of the sample size n, there is a curve showing variations of

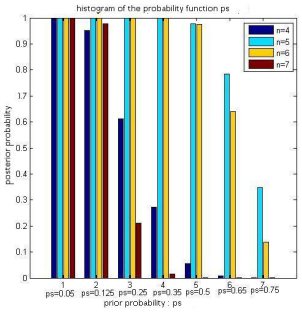

Figure 1: Histogram of posterior probability according to the prior probability

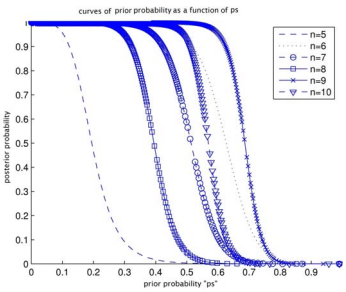

Figure 2: curves of posterior probability according to the prior probability

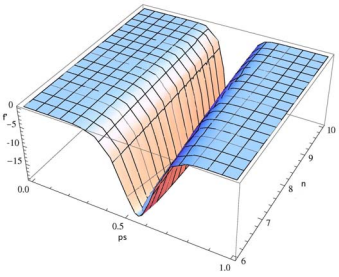

Figure 3: curve of the derivative function of f given by the formula 24

the posterior probability $P(\mathcal{S}/x_1, \ldots, x_n)$ based on the prior probability ps. We see that these curves are sharply decreasing in a neighborhood of 0.5.

Figure 3 contains the graph of the derivative (24) according to ps and n. We note that this derivative is zero outside a certain neighborhood of 0.5, which explains the variations of the functions in Figure 2.

V. CONCLUSION AND FUTURE WORK

According to remarks made in the previous section, we can say we can have interesting results in terms of spam filtering that when it is assumed that the a prior probability ps takes values that are in a neighborhood of 0.5 .

Obviously taking ps = 0.5 is not very realistic, because we all know that we do not receive one spam on two messages. The histograms in figure 1, the curves in figure 2 and the comments made on them allow us to say that if we want to be realistic by taking values of the prior probability ps far from 0.5, we would have results very unreliable.

To overcome these difficulties, it is useful to make a more specific study by proposing laws for prior probabilities , which summarize the best behavior of ps and must respect some theoretical criteria. Once these laws are established, it will lead to laws of posterior probability that will identify the best developments of $P(\mathcal{S}/x_1, \ldots, x_n)$.

These proposals fall within the scope of future work.

REFERENCES

[1] D. W. K. Khong. *An Economic Analysis of Spam Law.* Erasmus Law & Economics Review, Vol. 1, pp. 23-45, February 2004.

[2] A. Herzberg. *Controlling Spam by Secure Internet Content Selection.* Secure Communication Networks (SCN) 2004, LNCS vol. 3352, Ed. Springer-Verlag.

[3] M. Sahami, M. Dumais, S. Heckerman and E. Horvitz. *A Bayesian Approch to Filtering Junk E-mail.* AAAI Workshop, 1998, Madison Wisconsin.

[4] P. Graham. *A Plan for Spam.* http:// paulgraham.com/.

[5] G. Robinson. *A Statistical Approch to the Spam Problem.* Linux journal, 01-03-2003, Issue 107,2003.

[6] V. Metsis, I. Androutsopoulos and G. Paliouras. *Spam Filtering with Naive Bayes – Which Naive Bayes?* 3rd Conference on Email and Anti-Spam (CEAS 2006), Mountain View, CA, USA, 2006.

[7] Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim, Sung Hyon Myaeng. *Some Effective Techniques for Naive Bayes Text Classification.*IEEE Transactions on Knowledge and Data Engineering, Volume 18 Issue 11, November 2006.

[8] Yanhui Guo; Yaolong Zhang; Jianyi Liu; Cong Wang. *Research on the Comprehensive Anti-Spam Filter,* Industrial Informatics, 2006 IEEE International Conference on 16-18 Aug. 2006 Page(s):1069 - 1074

[9] Yan Zhou, Madhuri S. Mulekar, Praveen Nerellapalli. *Adaptive Spam Filtering Using Dynamic Feature Spaces.*International Journal on Artificial Intelligence Tools 16(4): 627-646 (2007)

[10] Hernes, O.; Jianna Zhang. *A tutorial search engine based on Bayesian learning.*Machine Learning and Applications, 2004. Proceedings. 2004 International Conference on 16-18 December, 2004 Page(s):418 - 422

[11] G. Sakkis, I. Androutsopoulos, G. Paliouras, V. Karkaletsis, C. D. Spyropoulos, P. Stamatopoulos. *Stacking classifiers for anti-spam filtering of e-mail.*"Empirical Methods in Natural Language Processing" (EMNLP 2001), L. Lee and D. Harman (Eds.), pp. 44-50, Carnegie Mellon University, Pittsburgh, PA, 2001

[12] Georgios Sakkis , Ion Androutsopoulos , Georgios Paliouras , Vangelis Karkaletsis , Constantine D. Spyropoulos and Panagiotis Stamatopoulos. *A Memory-Based Approach to Anti-Spam Filtering for Mailing Lists.*Information Retrieval, Volume 6, Number 1, Pages 49-73, / janvier 2003, Kluwer Academic Publishers Hingham, MA,USA,diteur Springer Netherlands

[13] I. Androutsopoulos, G. Paliouras and E. Michelakis. *Learning to Filter Unsolicited Commercial E-Mail.*Technical report 2004/2, NCSR "Demokritos", revised version (October 2004), with additional minor corrections (October 2006).

[14] Redwan Zakariah, Samina Ehsan.*Detecting junk Mails by Implementing Statistical Theory.* 20th International Conference on Advanced Information Networking and Applications (AINA 2006), 18-20 April 2006, Vienna, Austria. IEEE Computer Society 2006.

[15] Yerazunis, W.S.*"The Spam-Filtering Accuracy Plateau at 99.9% Accuracy and How to Get Past It".*MIT Spam Conference, January 2004

[16] Zhen Yang, Xiangfei Nie, Weiran Xu, Jun Guo.*An Approach to Spam Detection by Naive Bayes Ensemble Based on Decision Induction.* Intelligent Systems Design and Applications, 2006. ISDA '06. Sixth International Conference on, 16-18 Oct. 2006, Volume: 2, On page(s): 861-866.