Volume 3, Issue 3, March 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Enhancing Naïve Bayes Performance with Modified Absolute Discount Smoothing Method in Spam Classification

Astha Chharia
Department of CSE & IT
Madhav Institute of Technology and Science, India

R.K. Gupta
Department of CSE & IT
Madhav Institute of Technology and Science, India

Abstract— In recent years, Naïve Bayes classifier has gained much popularity in spam classification due to its simplicity and superior performance. We studied the performance of naïve bayes classifier and found that it largely depends on the smoothing method, which aims to adjust the probability of an unseen event from the seen event, that arises due to data sparseness. Therefore in this paper, we aim at enhancing the performance of naïve bayes classifier in classifying spam mails by proposing a modification to Absolute Discount smoothing method against the laplace method of traditional naïve bayes classifier. In addition, we have introduced a cost metric to compare our approach with the traditional scheme. Our experimental results have shown that our method not only achieves greater accuracy as compared to Laplace but also reduces false positives, which is more serious problem in spam classification.

Keywords— Naïve Bayes, Smoothing, Cost, False positive, Absolute discount.

I. INTRODUCTION

Electronic mail is one of the fastest and most common form of communication as it saves a lot of time and cost. As every emerging technology is associated with threat so is the case with e-mail and it is increasing at a rapid rate due to the increase in number of regular internet users. One of the major threat that email suffer with, is sending of unwanted emails to a large number of users and in huge amount. This problem is commonly is known as spam or unsolicited bulk email. Spam leads to filling up the mailboxes of the recipients and consuming their significant amount of time in reading and deleting. Also, they consume much of the network bandwidth and storage space and surcharge the mail servers with unwanted processing. Sometimes, they may spread malicious software or advertise unlicensed medicines and pornography. Thus spamming, if not worked upon soon, might become a potential type of "Denial of Service" attack.

Many techniques have been proposed to solve the problem of spam from blacklisting of well-known spammers or creating dictionaries with word patterns that mostly occur in spam messages to machine learning [7]. The former is based on the creation of list from FROM address or originating server's IP address. The drawback of blacklist is that list should be up-to-date and accurate [8]. The latter is considered more efficient and accurate as it requires less human interaction and classifies document automatically. One such simple and efficient machine learning technique that has gained much attention in spam classification is Naïve bayes (NB) classifier. Hence, we studied to enhance the performance of NB classifier. This can be done by using language modelling approach, which is, assigning probabilities to word sequences. Hence, this paper attempts to apply a suitable language modelling method in spam classification. Naïve Bayes classifier is a probabilistic classifier which uses Laplace smoothing method to deal with zero probability values. Thus, we thought to replace the Laplace method with a better method and to our surprise we found many smoothing methods that have been developed in language models. Many researchers have applied language models in text classification [3,6,18]. In this paper we replace laplace with a modification to Absolute Discount (AD) smoothing method and call it as NB-MAD technique (Naïve Bayes with Modified Absolute Discount) and to the best of our knowledge this is the first paper studying smoothing methods in spam classification. The motive of this paper is not only to increase the accuracy of naïve bayes but also to reduce the number of false positives because classifying nonspam mail as spam is a more severe problem than classifying spam mail as nonspam. This we have done without using the parameter λ as is done by [5, 19], in which mail is classified as spam only if its probability being in spam is λ times more than its probability being in nonspam. Although setting λ value high, causes increase in true negative (TN) value (nonspam classified as nonspam) and at the same time causes much decrease in true positive (TP) value (spam classified as spam). This can be easily seen from the results of [19] for naïve bayes. Hence, this cannot be seen as efficient spam filter as it is allowing spam mails to pass the filter. Therefore, we focused on using a proper smoothing method in which TN value is high and also there is not much reduction in TP value as in using λ. Our results showed that NB-MAD technique achieves equal or more accuracy and classifies more nonspam or legitimate mails correctly than traditional NB. Thus, resulting in reduced false positive value. Further, to compare two schemes in terms of false positive error, we introduced a cost metric.

Rest of the paper is organized as follows: Section 2 describes literature review. Section 3 describes the basic multinomial model and NB-MAD technique. Section 4 presents the datasets we have used. Section 5 introduces a cost metric to evaluate the performance of the two approaches. Section 6 presents our experimental results on different datasets and finally we conclude our work in Section 7.

II. LITERATURE SURVEY

Sahami et al. [5] proposed a Bayesian approach for spam classification in 1998. A Bayesian approach is a probabilistic method for the classifier to learn automatically. They have employed domain-specific features along with the content of the e-mail message into the bayes classifier and showed that by considering domain-specific features, more accurate classifier can be built.

McCallum and Nigam [1] has compared two variants of naïve bayes namely, multi-variate bernoulli model and multinomial model, in text classification. They found multinomial model best as it reduces error by an average of 27% over multi-variate bernoulli model. Also, there are many researchers who have used multinomial model for text classification such as Lewis and Gale [2], Kamal Nigam et al. [4].

Since NB is based on using Laplace method for dealing with zero-probability problem, many people have shown that using a better smoothing method than laplace, result in higher accuracy. Quan Yuan et al. [3] have shown the use of 4 smoothing methods, Jelinek-Mercer (JM) smoothing, Dirichlet (Dir) smoothing, Absolute discounting (AD) and Two-stage (TS) smoothing to improve the performance of naïve bayes classifier in short text classification on different sizes of training data. They have compared all the methods in terms of Macro-F1 and Micro-F1 and found that AD generally performs best for Macro-F1 scores.

David Vilar et al. [6] has also used AD smoothing method in text classification and have shown that it achieves better results by applying it on 5 different corpus.

Furthermore, F.He and X.Ding [9] have also used various smoothing methods, namely, Absolute discounting (AD), Good-Turing, Linear and Witten-Bell smoothing in text classification. They have studied the effect of smoothing methods, adopted from language model, on training set size and feature set size and have achieved better results than laplace.

III. MULTINOMIAL MODEL

Bayesian classifier is considered "naïve" in the sense because of its assumption that the presence or absence of a particular feature or word of a class is independent to the presence or absence of any other feature given the class variable. This assumption is made to simplify the computation and therefore naïve bayes is simple and easy to understand.

Consider a set of predefined categories or classes $C = \{C_1, C_2 \ldots C_m\}$ and a document d to be classified. NB classifier classifies a document d to class C_i if and only if

$$P(C_i|d) > P(C_j|d) \quad \text{for } i \neq j \text{ and } 1 \leq j \leq m \tag{1}$$

The probability of a document to belong in class C_i is calculated by Bayes theorem:

$$P(C_i|d) = \frac{P(C_i)\, P(d|C_i)}{P(d)} \tag{2}$$

Since the denominator P(d) is constant, therefore, it is ignored and the document is classified into the class C_i that maximizes $P(C_i)\, P(d|C_i)$.

There are two variants of naïve bayes: multi-variate bernoulli model and multinomial model. As multinomial performs well as compared to bernoulli model [12] so here we will use multinomial naïve bayes event model.

In multinomial naïve bayes, $P(d|C_i)$ is calculated as:

$$P(d|C_i) = (\sum_k N_{ik})! \prod_{k=1}^{|V|} \frac{P(w_k|C_i)^{N_{ik}}}{N_{ik}!} \tag{3}$$

where |V| is the vocabulary, N_{ik} is the number of times word w_k occurred in document d, $P(w_k|C_i)$ is the conditional probability of word w_k given class C_i.

The terms in (3) $(\sum_k N_{ik})!$ and $\Pi_k N_{ik}!$ can be ignored because none of them depends on class [20]. Thus the (3) becomes:

$$P(d|C_i) = \prod_{k=1}^{|V|} P(w_k|C_i)^{N_{ik}} \tag{4}$$

NB requires each conditional probability to be non-zero, hence, laplace smoothing method is usually used to avoid zero-probability values:

$$P(w_k|C_i) = \frac{1 + \text{count}(w_k, C_i)}{|V| + \sum_{w \in V} \text{count}(w, C_i)} \tag{5}$$

where $\text{count}(w_k, C_i)$ is the count of occurrence of word w_k in category C_i.

Although laplace method is simple as it assumes there is atleast one sample for each word-class pair but this addition to every word is not considered as the effective smoothing method. However, this problem has been widely studied in language models. Many smoothing methods such as Katz smoothing[14], Jelinek-Mercer[16], Absolute discount [15], Kneser-Ney [17] and more have been developed in language models.

All the smoothing methods introduced in language models differ in likelihood estimation of probability $P(w_k|C_i)$. In this paper we propose an alternative approach to absolute discount, so before proceeding further, first we will explain Absolute Discount (AD) method. In this method, a small amount 'delta' is subtracted from every positive word count for seen words in order to redistribute the discounted probability mass on unseen words.

In Absolute discount, $P(w_k|C_i)$ is calculated as:

$$P(w_k|C_i) = \frac{\max(count(w_k, C_i) - delta, 0) + delta\ (Nuc_i)\ P(w_k)}{\sum_{w \in V} count(w, C_i)} \quad (6)$$

where delta ∈ [0,1]

Nuc_i is the number of unique words in C_i

$P(w_k)$ is the probability of word w_k in collection model and is given by:

$$P(w_k) = \frac{\sum_{j=1}^{m} count(w_k, C_j)}{\sum_k \sum_j count(w_k, C_j)} \quad (7)$$

Our approach is the variation of AD method, it differs only in calculation of $P(w_k)$. We did not consider the probability of word in collection model, rather we consider it as a function of word, which is a uniform distribution probability multiplied by the occurrence of word in collection model and is given by:

$$f(w_k) = P_{unif}(w_k) \sum_{j=1}^{m} count(w_k, C_j) \quad (8)$$

where, $P_{unif}(w_k) = \frac{1}{|V|}$

Hence, $P(w_k|C_i)$ is calculated as:

$$P(w_k|C_i) = \frac{\max(count(w_k, C_i) - delta, 0) + delta\ (Nuc_i)\ f(w_k)}{\sum_{w \in V} count(w, C_i)} \quad (9)$$

where delta ∈ [0,1]

Nuc_i is the number of unique words in C_i.

We use our approach with naïve bayes and thus it is named as NB-MAD.

Here our emphasis is that by using a proper smoothing method with NB, one can achieve quite good spam classification results just by using words as features and without getting deep into various proposed complex schemes and applying them to achieve better results. We are getting quite good results just by following the basic classification steps with better smoothing method.

IV. CORPUS DESCRIPTION

We have performed experiments on 3 different corpora, SpamAssassin, PU1 and Lingspam. SpamAssassin can be found at [13]. We use its recent added mails which consist of 1400 legitimate or nonspam mails and 1397 spam mails, in total 2797 mails. To perform our experiment we use first 1200 nonspam and spam mails. Out of which we select 600 mails of each category for training and rest for testing. Each of these mails were preprocessed by removing all HTML tags, performing word stemming and removing stop words like 'a', 'an', 'the'.

PU1 corpus can be found at [10]. It consist of 4 subfolders: bare, stop, lemm, lemm_stop. We use the lemm_stop as in this word stop list and word lemmatizer has been applied. It consist of 618 nonspam and 481 spam messages, in total 1099 messages. We split this into training and testing set of 800 and 299 messages respectively.

Lingspam corpus can be found at [11]. It consist of 2893 messages; 2412 of them are nonspam and 481 of them are spam messages. In order to test our method, we randomly select 300 messages from each category for training and use the remaining spam messages plus nonspam messages in same amount for testing. Then applied word stemming and stop-word list on this corpus. In all of these corpora, body of every message was tokenized into words where words comprises of alphabets, numbers, underscore, hyphen and apostrophes. These words are known as features or attributes. As the size of this feature space is very large so we use the most popularly used feature selection technique known as information gain [12].

V. PERFORMANCE MEASURE

To evaluate the performance of classifier commonly used measures are accuracy, precision and recall. Accuracy is given by:

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \quad (10)$$

Precision and Recall is given by:

$$Precision = \frac{TP}{TP + FP} \quad (11)$$

$$Recall = \frac{TP}{TP + FN} \quad (12)$$

where
TN = true negative, nonspam predicted as nonspam
TP = true positive, spam predicted as spam
FP = false positive, nonspam predicted as spam
FN = false negative, spam predicted as nonspam

Accuracy treats the 2 type of errors equally, that is, legitimate classified as spam and spam classified as nonspam but in email classification these 2 errors are not of same importance. A false positive is more costly than false negative. Therefore, accuracy cannot be used here to measure the performance and hence we introduced a cost metric to compare our approach NB-MAD with the traditional method. Cost can be computed as:

$$\text{Cost} = \frac{FP}{TP + FN} \quad (13)$$

The value of cost indicates the percentage of misclassified legitimate mails from the total spam mails. Thus, its value should be less. In other words, lower value of cost depicts better performance.

VI. EXPERIMENTAL RESULTS

We performed experiments on train set and test set build from 3 different datasets as described in section 4. Then we compare our method with absolute discount (AD) and Laplace method in terms of accuracy, cost and spam recall. Our main emphasis is on decreasing the cost without much decrease in spam recall. The table 1shows results for spamassassin for 500 and 1000 words selected using IG on testing set size of 800 mails.

TABLE I
COMPARISON of NB PERFORMANCE by ACCURACY, SPAM RECALL and COST for SPAMASSASSIN on 500 and 1000 WORDS

Filter used	Delta	500 words			1000 words		
		Accuracy (%)	Spam Recall (%)	Cost	Accuracy (%)	Spam Recall (%)	Cost
NB-L	-	67.37	87.25	0.525	67.13	87.5	0.533
NB-AD	0.6	67.25	87.75	0.533	67.13	88	0.538
	0.1	67.75	87.75	0.523	67.75	88.5	0.53
NB-MAD	0.6	75.75	68.75	0.173	78.13	79	0.228
	0.1	69.63	86.5	0.473	68.38	87	0.503

NB-L : naïve bayes with laplace, NB-AD : naïve bayes with absolute discount, NB-MAD : naïve bayes with modified absolute discount It is clear from table 1 that our method (NB-MAD) is more efficient in predicting legitimate mails correctly by reducing cost on an average of 33% over traditionally used laplace method and AD method for delta = 0.6. Also, we can see for delta = 0.10, NB-MAD produces good results than laplace as its cost is less and in addition to this we can observe that there is not much reduction in spam recall. Further, for delta = 0.10 AD method is also better than laplace. Hence, our method achieves better performance both in terms of accuracy and cost against AD and Laplace.
The table 2 shows results for Lingspam for 100 and 200 words selected using IG.

TABLE II
COMPARISON of NB PERFORMANCE by ACCURACY, SPAM RECALL and COST for LINGSPAM on 100 and 200 WORDS

Filter used	Delta	100 words			200 words		
		Accuracy (%)	Spam Recall (%)	Cost	Accuracy (%)	Spam Recall (%)	Cost
NB-L	-	94.198	95.03	0.066	94.75	95.58	0.061
NB-AD	0.1	94.475	95.03	0.061	94.75	95.58	0.061
NB-MAD	0.1	95.3	93.37	0.027	95.86	94.48	0.027

NB-L : naïve bayes with laplace, NB-AD : naïve bayes with absolute discount, NB-MAD : naïve bayes with modified absolute discount Again it is clear from table 2, NB-MAD achieves better accuracy and reduces cost of incorrectly classifying nonspam mails. As it can be seen, cost of NB-MAD is 3.4% greater than NB-L for 200 words, so the performance of naïve bayes is enhanced by using modified AD.
The table 3 shows results for PU1 for 500 and 1000 words selected using IG.

TABLE III
COMPARISON of NB PERFORMANCE by ACCURACY, SPAM RECALL and COST FOR PU1 ON 500 and 1000 WORDS

Filter used	Delta	500 words			1000 words		
		Accuracy (%)	Spam Recall (%)	Cost	Accuracy (%)	Spam Recall (%)	Cost
NB-L	-	96.32	98.96	0.104	96.32	98.96	0.104
NB-AD	0.15	96.65	98.96	0.094	96.65	98.96	0.094
NB-MAD	0.15	97.32	96.88	0.052	97.32	97.91	0.062

NB-L : naïve bayes with laplace, NB-AD : naïve bayes with absolute discount, NB-MAD : naïve bayes with modified absolute discount

Again, NB-MAD is proven to enhance NB performance by using a better smoothing method than laplace. The cost of NB-MAD for both 500 and 1000 words is lower than NB-L.

VII. CONCLUSION

In this paper we have proposed a modification to absolute discount smoothing method in order to enhance the naïve bayes performance. Also, to compare our method we have introduced a cost metric. However, our emphasis is not only on increasing the accuracy of naïve bayes but also on decreasing the number of legitimate mails being incorrectly classified because classifying legitimate mails as spam is more severe problem than classifying spam as legitimate. Though, at the same time we need to remember, that lesser spam mail should be classified incorrectly. Classifying legitimate mails correctly alone won't serve the complete purpose of spam filtering. Hence our approach or NB-MAD technique greatly enhances NB performance and reduces cost by an average of 33% for spamassassin corpus without much reduction in spam recall than Laplace. This can be easily seen from our results presented in section 6.

REFERENCES

[1] A. McCallum, K. Nigam, "A comparison of event models for naive Bayes text classification". AAAI/ICML-98Workshop on Learning for Text Categorization, AAAI Press 41–48, 1998.
[2] D.D. Lewis, and W.A. Gale, "A sequential algorithm for training text classifiers". In SIGIR-94, 1994
[3] Q. Yuan, G. Cong, and N.M. Thalmann, " Enhancing Naive Bayes with Various Smoothing Methods for Short Text Classification", in Proc. of the 21st international conference companion on World Wide Web, WWW '12 Companion, 2012.
[4] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell, "Learning to classify text from labeled and unlabeled documents". In AAAI-98,1998.
[5] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A Bayesian Approach to Filtering Junk e-mail". AAAI Tech. Rep.WS-98-05, pp.55–62, 1998.
[6] D. Vilar, H. Ney, A. Juan, and E. Vidal, "Effect of Feature Smoothing Methods in Text Classification Tasks". In International Workshop on Pattern Recognition in Information Systems, pages 108-117. Porto, Portugal, 2004.
[7] D. Gavrilis, I.G. Tsoulos, E. Dermatas, "Neural Recognition and Genetic Features Selection for Robust Detection of E-Mail Spam". SETN 2006, pp. 498-501, 2006.
[8] K.P. Clark, "A Survey of Content-based Spam Classifiers", 2008.
[9] F. He, and X. Ding, "Improving naive Bayes text classifier using smoothing methods", in Proc. of the 29th European conference on IR research, April 02-05, Rome, Italy, 2007.
[10] PU1 corpus [Online], Available: http://www.aueb.gr/users/ion/data/pu1_encoded.tar.gz.
[11] Lingspam corpus [Online], Available: http://labs-repos.iit.demokritos.gr/skel/iconfig/downloads/lingspam_public.tar.gz.
[12] Y. Yang, and J.O. Pedersen, "A comparative study on feature selection in text categorization". In Fisher, D.H., ed.: Proceedings of ICML-97, 14th International Conference on Machine Learning, Nashville, US, MorganKaufmann Publishers, SanFrancisco, 412–420, 1997.
[13] Spam Assassin corpus [Online], Available: http://spamassassin.apache.org/publiccorpus/
[14] S.M. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer", IEEE Transactions on Acoustics, Speech, and Signal Processing, volume ASSP-35, pages 400–401, march 1987.
[15] H. Ney, U. Essen and R. Kneser, "On structuring probabilistic dependences in stochastic language modeling". Computer Speech and Language, 8, 1-38, 1994.
[16] F. Jelinek, and R.L. Mercer, "Interpolated estimation of Markov source parameters from sparse data", in Proc. Workshop on Pattern Recognition in Practice, pages 381-397, Amsterdam, 1980.
[17] R. Kneser, and H. Ney, "Improved backing-off for M-gram language modeling". In International Conference on Acoustic, Speech and Signal Processing, pages 181-184, 1995.
[18] J. Bai, and J.Y Nie, "Using Language Models for Text Classification", in Proc. of the Asia Information Retrieval Symposium (AIRS'04), October 2004, Beijing, China.

[19] I. Androutsopoulos, G. Paliouras, V. Karkaletsis, G. Sakkis, C.D. Spyropoulos, and P. Stamatopoulos, "Learning to Filter Spam E-Mail: A Comparison of a Naive Bayesian and a Memory-Based Approach". In H. Zaragoza, P. Gallinari, and M. Rajman (Eds.), Proc. of the Workshop on Machine Learning and Textual Information Access, 4th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2000), Lyon, France, pp. 1-13, 2000.

[20] A.M. Kibriya, E. Frank, B. Pfahringer, G. Holmes, "Multinomial Naive Bayes for Text Categorization Revisited", in Proc. of Australian Conference on Artificial Intelligence, 2004, pp.488-499.

