Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 10, October 2014, pg.864 – 868

REVIEW ARTICLE

A REVIEW ARTICLE ON NAIVE BAYES CLASSIFIER WITH VARIOUS SMOOTHING TECHNIQUES

Gurneet Kaur*, Er. Neelam Oberai

*Student of Masters of Technology, Computer Science, Department of Computer Science, MAHARISHI MARKANDESHWAR UNIVERSITY, SADOPUR (AMBALA)

Assistant Professor of Masters of Technology, Computer Science, Department of Computer Science, MAHARISHI MARKANDESHWAR UNIVERSITY, SADOPUR (AMBALA)

* gur_neet_kaur66@yahoo.com; Neelamoberoi1030@gmail.com

Abstract- Naive Bayes is very popular in commercial and open-source anti-spam e-mail filters. There are, however, several forms of Naive Bayes, something the anti-spam literature does not always acknowledge. We discuss five different versions of Naive Bayes, and compare them on six new, non-encoded datasets, that contain ham messages of particular Enron users and fresh spam messages. The new datasets, which we make publicly available, are more realistic than previous comparable benchmarks, because they maintain the temporal order of the messages in the two categories, and they emulate the varying proportion of spam and ham messages that users receive over time. In this paper we have discovered various aspects of Naïve Bayes Classifier and smoothing techniques for extraction of useful data along with our research criteria.

Keywords: Naïve Bayes, methods, classification, Smoothing, Two-Stage, Absolute Discounting

1. INTRODUCTION

With large increase in the amount of text documents, spam's are increasing day by day so the manual handling is not a feasible solution and it has become necessary to categorize them in different classes. There is various classification methods developed, but the choice of using these techniques mainly depend upon the type of data collections. Some Classifiers are discussed. Few methods perform well on

numerical and text data like Naive Bayes but neural networks handle both discrete and continuous data. KNN is a time consuming method and finding the optimal value is always an issue. Decision tree reduces the complexity but fails to handle continuous data. Naïve Bayes along with its simplicity is computationally cheap also. In the second section of the paper, Naïve Bayes classifier is discussed in detail. One of the major drawback of Naïve Bayes is of unseen words, which can be eliminated by applying smoothing techniques. In the IIIrd section, various smoothing methods when applied on Naïve Byes are discussed and their performances are compared.

2. NAÏVE BAYES CLASSIFIER

A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes theorem (from Bayesian statistics) with strong (naive) independence assumptions.[3] An advantage of the naive Bayes classifier is that it only requires a small amount of training data to estimate the parameters necessary for classification.

Assumption: 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.

2.1 Naive Bayesian Classification Model

The need and requirements of the admin user's of the websites to analyze the user preference become essential, due to massive internet usage. Retrieving the decisive information about the user preferences is achieved, using Naive Bayesian Classification algorithm with quicker time and lesser memory, by means of constructive naïve bayes function. The Naive Bayesian Classification technique as shown in Fig 2, is applied on the web log data to evolve the classification of user page preferences and time spent on the pages of the respective web site (URL).

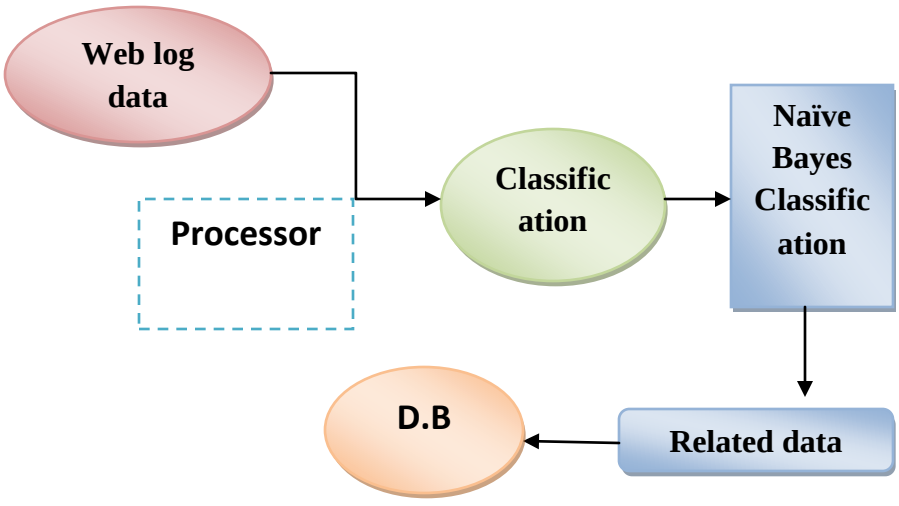

Fig 1 Framework of Naive Bayesian Classification

3. Models of Naïve Bayes Classifier: [5]

1) *Multivariate Bernoulli model:* A document is represented by a binary feature vector, whose elements (1/0) indicate presence or absence of a particular word in a given document. In this case the document is considered to be the event and the presence and absence of words are considered as attributes of the event. [4]

2) *Multinomial model:* A document is represented by an integer feature vector, whose elements indicate frequency of corresponding word in the given document. Thus the individual word occurrence is considered to be events and document is considered to be collection of word events. Multinomial model is more accurate than the multivariate Bernoulli model for many classification tasks because it considers the frequency of the words too. [4][5]

3) *Probabilistic Model:*[3]

Consider D be the set of documents and C be the set of classes. The probability of assigning a document d to a class c is given by:

$$C_{NB} = \text{argmax}_{cj \in c} P(c|d) = \text{argmax}_{cj \in c} \frac{P(c)P(d|c)}{P(d)} \quad \text{.....(4)}$$

As P(d) is independent of the class, it can be ignored.

$$C_{NB} = \text{argmax}_{cj \in c} P(c)\, P(d|c) \quad \text{……(5)}$$

According to Naïve Bayes assumption,

$$P(d|c) = P(w_1|c)\, P(w_2|c) .. P(w_d|c) = \prod_{1<k<d} P(wk|c) \quad \text{.....(6)}$$

Replacing (5) by $C_{NB} = \text{argmax}_{cj \in c} P(c) \prod_{1<k<d} P(wk|c)$ ……(7)

Where P(c) is the prior probability of the class c_j, which is calculated as $\frac{N}{n}$, where N is the total number of training documents in class c, n is the total number of training documents. P(c|d) is the posterior probability.

$P(w_k|c) = \frac{T}{\sum_{t' \in V} T'}$ ……..……(8)[5]

Where T is the number of occurrences of w in d from class c, $\sum_{t' \in V} T'$ is the total number of words in d from class c [5].

4. SMOOTHING METHODS

It refers to the adjustment of maximum likelihood estimator for the language model so that it will be more accurate. At the very first, it is not required to assign the zero value to the unseen word. It plays two important roles: 1) Improves the accuracy of the language model. 2) Accommodate the generation of common and non informative words.

General Model:

The maximum likelihood generator generally under estimate the probability of unseen words. So the main purpose of the smoothing is to provide a non-zero probability to unseen words and improve the accuracy of probability estimator. The general form of smoothed model is of the form:

$$P(w|d) = \begin{cases} P_s(w \mid d) & \text{if w is seen} \\ \alpha_d P(w|c) & \text{otherwise} \end{cases}$$

Where $P_s(w \mid d)$ is the smoothed probability word seen in the document and P(w | d) is the collection language model and α_d is the coefficient controlling the probability assigned to unseen words so that probabilities sum to one.

Generally, Smoothing methods differ in choice of P_s (w | d). A Smoothing method can be as simple as adding extra count or more complex where words of different count are treated differently.

1.) Jelinek-Mercer method: This method involves a linear interpolation of the maximum likelihood model with the collection model using a coefficient ʎ.[6] [7] [8]

$$P_\lambda(w|d) = (1-\lambda)\, P_{ml}(w|d) + \lambda\, P(w|c) \quad \ldots\ldots\ldots\ldots\ldots\ldots (9)$$

2.) Using Dirichlet Priors: A language model is a multi-nominal distribution, for which the conjugate prior for the Bayesian analysis is the Dirichlet distribution with parameters

$$(\mu p(w1|c), \mu p(w2|c), \mu p(w3|c), \ldots.., \mu p(w1|c))$$

Thus, model is given by: [6] [7] [8]

$$P_\mu(w|d) = \frac{count(w,d) + P(w|c)}{\sum_w count(w,d) + \mu} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (10)$$

3.) Absolute Discounting: It lowers the probability of seen words by subtracting a constant from their counts. It is similar to JK method but differs in that it discounts the probability by subtracting instead of multiplying.

$$P_\delta(w|d) = \frac{\max(count(w,d) - \delta, 0)}{\sum_w count(w,d)} + \sigma\, P(w|c) \quad \ldots\ldots\ldots\ldots\ldots (11)$$

Where δ is a discount constant and $\sigma = \delta\, |d|_u / |d|$, so that it equals to one. Here, $|d|_u$ is the number of unique terms in d and |d| are the total number of terms. [6] [7] [8]

4.)Two-Stage Smoothing: It combines the Dirichlet Smoothing with the Interpolation method as; [6][7][8]

$$P_{TS}(w|c_i) = (1 - \lambda)\, \frac{count(w,c) + \mu P(w|c)}{|ci| + \mu} + \lambda\, P(w|c) \quad \ldots\ldots\ldots (12)$$

Name	***Method***	***Para meter***
JM Smoothing	$P_\lambda(w|d) = (1-\lambda)P_{ml}(w|d) + \lambda\, P(w|c)$	λ
Dirichlet Smoothing	$P_\mu(w|d) = \frac{count(w,d) + P(w|c)}{\sum_w count(w,d) + \mu}$	μ
Absolute Discounting	$P_\delta(w|d) = \frac{\max(count(w,d) - \delta, 0)}{\sum_w count(w,d)} + \sigma P(w|c)$	δ
Two-Stage Smoothing	$P_{TS}(w|c_i) = (1 - \lambda)\, \frac{count(w,c) + \mu P(w|c)}{|ci| + \mu} + \lambda\, P(w|c)$	λ and μ

Fig 2. Summary of Smoothing Techniques [6]

Laplace Smoothing is replaced by various sophisticated smoothing methods like JK Smoothing, Dirichlet Smoothing, Two-Stage Smoothing, and Absolute Discounting. By applying the various Smoothing techniques, the performance of the Naïve Bayes has been increased. Dirichlet Smoothing method performed well than other methods. JM performs well mostly in case long verbose Queries instead of precise ones. Dirichlet is the most efficient type of smoothing. Absolute discounting performs well in case of short term documents. [6]

5. CONCLUSION

Text classification has become a major issue, now a days and one reason of it is the lack of single technique, which is able to produce good classification for different data sets. There are various classification methods such as Decision Trees, Neural Networks, Naïve Bayes and Centroid Based, but Naïve Bayes performs better for different data collections and is easy and computationally cheap. Along with its simplicity, Naïve Bayes also suffers from the some issues like unseen words. So, we use various smoothing techniques like JK method, Absolute Discounting method, Dirichlet Smoothing and Two-stage Smoothing to enhance the performance and accuracy of Naïve Bayes. We conclude that two-stage smoothing performs well with NB. In the next paper we will try to discover various spams and their distribution with Naïve Bayes.

REFERENCES

[1] B S Harish, D S Guru and S Manjunath," Representation and Classification of Text Documents: A Brief Review", IJCA Special Issue on Recent Trends in Image Processing and Pattern Recognition, 2010.
[2] Y. H. LI and A. K. JAIN, "Classification of Text Documents", The Computer Journal,1998.
[3] Kevin P. Murphy, " Naïve Bayes classifier", Department of Computer Science, University of British Columbia,2006.
[4] Hetal Doshi and Maruti Zalte, "Performance of Naïve Bayes Classifier-Multinomial model on different categories of documents" National Conference on Emerging Trends in Computer Science and Information Technology,IJCA,2011.
[5] Ajay S. Patil and B.V. Pawar, "Automated Classification of Naïve Bayesian Algorithm" ,Proceedings of International Multi-Conference of Engineers and Computer Scientists, March 14-16,2012.
[6] C. Zhai and J. Lafferty, "A Study of Smoothing Methods for language Models Applied to Information Retrieval" TOIS, 22:179 – 214, 2004.
[7] Jing Bai and Jian-Yun Nie. "Using Language Models for Text Classification", InAIRS, 2004.
[8] Quan Yuan , Gao Cong and Nadia M. Thalmann,"Enhancing Naïve Bayes with Various Smoothing Method for Short text Classification", Proceedings of 21st International Conference on World Wide Web, pages 645-646,2012.
[9] Colas, Fabrice, and Pavel Brazdil. "Comparison of SVM and some older classification algorithms in text classification tasks." In Artificial Intelligence in Theory and Practice, pp. 169-178. Springer US, 2006.

