TECHNICAL SEMINAR REPORT

“NAIVE BAYES ALGORITHM ON SMALL SAMPLE SET”

PAVAN KUMAR D
1PE14EC094

VISVESVARAYA TECHNOLOGICAL UNIVERSITY
Belgavi-590014

SEMINAR REPORT
ON
“NAIVE BAYES ALGORITHM ON SMALL SAMPLE SET ”

Submitted in Partial Fulfilment of the Requirement for VIII Semester

BACHELOR’S DEGREE
IN
ELECTRONICS AND COMMUNICATION ENGINEERING

For the Academic year
2017-2018

BY
PAVAN KUMAR D
(1PE14EC094)
UNDER THE GUIDANCE OF

PROF. SHWETHA S BHAT
Dept.of ECE,PESIT,BSC.

Department of Electronics and Communication Engineering
PESIT Bangalore South Campus
HOSUR ROAD, BANGALORE-560100

PESIT BANGALORE SOUTH CAMPUS

HOSUR ROAD

BANGALORE-560100

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

CERTIFICATE

*This is to certify that the seminar entitled **"NAIVE BAYES ALGORITHM ON SMALL SAMPLE SET"** is a bonafide work carried out by **PAVAN KUMAR D** bearing the register number **1PE14EC094** in partial fulfillment of the requirement for the award of Degree of Bachelors of Engineering in Electronics and Communication Engineering under Visvesvaraya Technological University,Belagavi during the year 2017-2018.*

Signatures:

Seminar Guide	*Head of the Department*
Shwetha S Bhat	***Dr.Subhash Kulkarni***
Assistant Professor,ECE	***HOD,ECE***
	PESIT- BSC

Examiners:

1.

2.

ACKNOWLEDGEMENTS

I am grateful to my college **PES Institute of Technology - BSC** for providing me the opportunity and would like to express a sense of gratitude to my Principal/Director **Dr. J Surya Prasad** for the continued effort in creating a competitive environment in my college.

I would also like to convey my heartfelt thanks to my H.O.D **Dr. Subhash S Kulkarni**, for his encouragement. I would also like to extend my sincere and heartfelt thanks to my guide **Asst. Prof. Shwetha S Bhat** and I am ineffably indebted to him for his conscientious guidance and encouragement.

I also wish to thank all the staff members of the department of Electronics & Communications for helping in completing this work successfully.Any omission in this brief acknowledgement does not mean lack of gratitude.

PAVAN KUMAR D

1PE14EC094

ABSTRACT

Naive Bayes algorithm is one of the most effective methods in the field of text classification,but only in the large training sample set can it get a more accurate result. The requirement of large number of samples not only brings heavy work for previous manual classification, but also puts forward a higher request for storage and computing resources during the computer post-processing.

This paper mainly studies Nave Bayes classification algorithm based on Poisson distribution model, and the experimental results show that this method keeps high classification accuracy even in small sample set.

Nave Bayes: Text classification, Poisson distribution, Classification accuracy, small sample set.

Contents

List of Figures

Chapter 1

Introduction

1.1 BACKGROUND

With the rapid development of Internet applications, e-commerce and network communication, there is a geometric multiples growth for information, which has brought our lives more and more important influence. Almost all information we want can be found in the network. But what people care about most is how to dig out the most valuable information from this large quantity of information.

The technology of automatic text classification is one of the basic ways to solve these problems, and it is an important research subject in information storage and retrieval. Automatic text classification has many advantages, such as needing no human intervention, saving a lot of manpower and updating quickly.

1.2 MAJOR CONCEPTS FOR IMPLEMENTATION

Traditional Naive Bayes algorithm, Poisson distribution for text classification, Parameter estimation and weight-enhancing, Comparision with Evaluation standards.

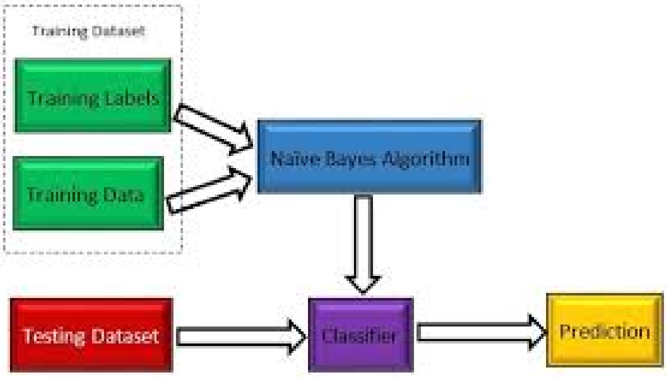

Figure 1.1: Basic Block Diagram

Chapter 2

IMPROVEMENT OF CLASSIFICATION ALGORITHM FOR SMALL SAMPLE SET

2.1 DEDUCTIVE REASONING OF NAIVE BAYES

From the above basic principles of Naive Bayesian classifier, the probability of document dj belonging to the C class is calculated as:

$$p(c|d_j) = \frac{p(d_j|c)p(c)}{p(d_j)}$$
$$= \frac{p(d_j|c)p(c)}{p(d_j|c)p(c) + p(d_j|\bar{c})p(\bar{c})}$$

Where P(dj) denotes probabality of document dj.P(cbar) denotes probabality of non occuring classes.

If we define Pjc,

$$P_{jc} = \log \frac{p(d_j|c)}{p(d_j|\bar{c})}$$

Then above expression can be modified as,

$$p(c|d_j) = \frac{e^{P_{jc}} \cdot p(c)}{e^{P_{jc}} \cdot p(c) + p(\bar{c})}$$

Therefore, we can get the posterior probability P(c/dj) by calculating Pjc.

2.2 POISSON DISTRIBUTION FOR TEXT CLASSIFICATION

Poisson distribution is suitable to describe the times of random events happening at the unit time (or space).For example, the number of telephone exchange receiving the calls, the number of guests waiting for the train on the platform, the number of faulty machines, the frequency of natural disasters, the number of defects on a product etc.

We assume that a document is generated by a multivariate Poisson model. A document dj represented as a random vector which consists of Poisson random variables Xij, where Xij has th value of within-document-frequency fij for ith term. Therefore, P(dj) can be expressed as:

$$p(d_j) = p(X_{1j} = f_{1j}, X_{2j} = f_{2j}, \cdots, X_{|V|j} = f_{|V|j})$$

Assuming each of the variables Xij is independent of each other, the probablity of dj calculated as,

$$p(d_j) = \prod_{i=1}^{|V|} p(X_{ij} = f_{ij})$$

Where,

/V/ denotes vocabulary size.

Using Poisson model P(Xij = fij) can be calculated as:

$$p(X_{ij} = f_{ij}) = \frac{e^{-\lambda_{ic}} \lambda_{ic}^{\ f_{ij}}}{f_{ij}!}$$

λ is the Poisson mean.

Therefore, Pjc can be estimated from Poisson model as:

$$P_{jc} = \sum_{i=1}^{|V|} \log \frac{p(X_i = f_{ij} | c)}{p(X_i = f_{ij} | \bar{c})}$$

$$= \sum_{i=1}^{|V|} \log \frac{e^{-\lambda_{ic}} \lambda_{ic}^{f_{ij}}}{e^{-\mu_{ic}} \mu_{ic}^{f_{ij}}}$$

$$= \sum_{i=1}^{|V|} (\mu_{ic} - \lambda_{ic}) + \sum_{i=1}^{|V|} f_{ij} \cdot \log \frac{\lambda_{ic}}{\mu_{ic}}$$

Where λ_{ic} and μ_{ic} are the Poisson means in positive class and negative class, respectively.

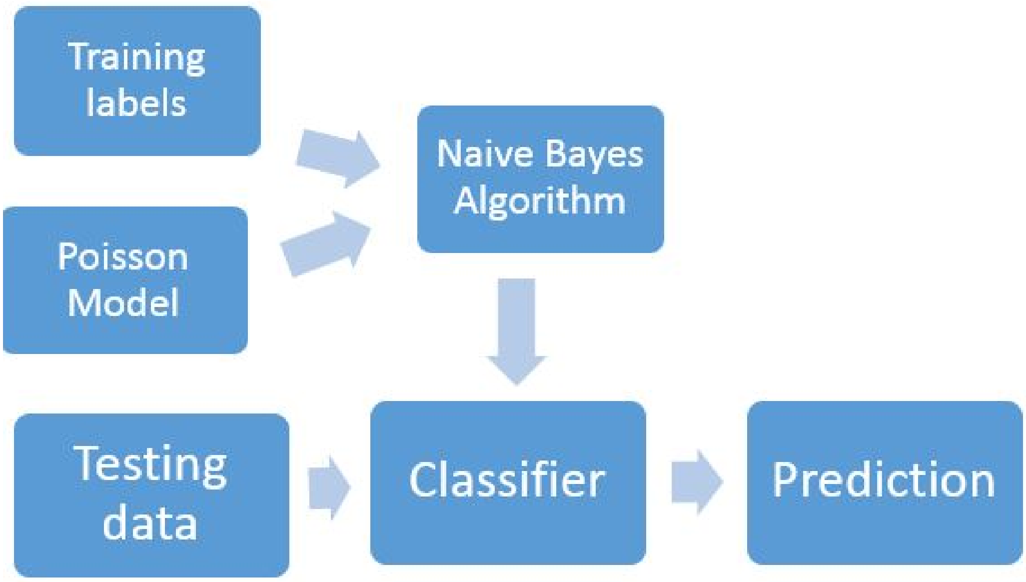

Figure 2.1: Implementation of Naives Bayes for Small Sample Set

2.3 PARAMETER ESTIMATION AND WEIGHT ENHANCING

From the definition of the Poisson distribution, that Poisson parameter is the average rate of random events in unit time or unit area. Thus, we define the average number of occurrences of the positive documents, and the negative documents as:

$$\lambda_{ic} = \frac{1}{|D_c|} \cdot \sum_{j=1}^{|D_c|} f_{ij}$$

$$\mu_{ic} = \frac{1}{|D_{\bar{c}}|} \cdot \sum_{j=1}^{|D_{\bar{c}}|} f_{ij}$$

Where,

/Dc/ and /Dc'/ are positive and negative documents.

fij is the actual frequency.

Thus, we define Pjc as:

$$P_{jc} = \sum_{i=1}^{|V|} f_{ij} \cdot \left(\frac{\lambda_{ic}}{\mu_{ic}} + \frac{\mu_{ic}}{\lambda_{ic}} \right) \cdot \log \frac{\lambda_{ic}}{\mu_{ic}}$$

Thus, for words which have obvious category feature, it further strengthened its classification weight.

Chapter 3

EVALUATION STANDARD

Here, we use Internationally accepted classification evaluation system to assess the performance, including the recall rate R, the precision rate P, F1 assessed value, the macro average accuracy MacroP, the macro average recall rate MacroR and the macro average F1 value MacroF1 and the corresponding formulas are as follows:

3.1 PRECISION AND RECALL RATES

$$P_j = \frac{l_j}{m_j} \times 100\%$$

Where,

lj is the correct number of text classification in category j.

mj is the actual number of text classified by Classification system.

Similarly, we have

$$R_j = \frac{l_j}{n_j} \times 100\%$$

Where,

Where, nj is the number of category j by an expert.

3.2 MACRO F1 VALUE

Finally, MacroF1 value computed as:

$$MacroF1 = \frac{MacroP \times MacroR \times 2}{MacroP + MacroR}$$

Where,

MacroP and MacroR are given by,

$$MacroP = \frac{1}{n} \sum_{j=1}^{n} P_j$$

P_j is the accuracy of category j.

$$MacroR = \frac{1}{n} \sum_{j=1}^{n} R_j$$

R_j is the recall rate of category j.

Chapter 4

EXPERIMENTAL RESULTS

The data used in the experiment comes from text classification corpus provided by the laboratory of Sogou. The corpus comes from the Sohu news site edited and saved with a lot of manual sorting and classification.

Here, lot of preprocessing work in the original corpus, including word segmentation, removing stop words and single words, computing word frequency in a document and computing word frequency in a category. We select eight categories to do the test, including finance, health, sports, tourism, education, recruitment, culture and military.

Category	TFNB			PDNB		
	P	R	F1	P	R	F1
Finance	0.974468	0.773649	0.862524	0.95053	0.908784	0.929188
Health	0.953782	0.761745	0.847015	0.919872	0.963087	0.940984
Education	0.893238	0.836667	0.864028	0.898734	0.946667	0.922078
Tourism	0.902344	0.791096	0.843066	0.847826	0.934932	0.889251
Sports	0.510345	0.996633	0.675029	0.989831	0.983165	0.986486
Culture	0.933333	0.143836	0.249258	0.876448	0.777397	0.823956
Recruitment	0.863492	0.918919	0.890344	0.977612	0.885135	0.929078
Military	0.991935	0.82	0.89781	0.957792	0.983333	0.970395

Figure 4.1: Results on Small Sample Set

From above table, in terms of precision and recall rates, TFNB may be better than PDNB in some categories, while in some other categories PDNB method works much better.

Chapter 5

ANALYSIS

Here, we consider Bayesian classifier based on word frequency (TFNB) and Poisson distribution-based Bayesian classifier (PDNB). And several sets of comparative experiments in the large-scale data set (1200 training documents for each category) and small-sample data set (100 documents in each (category) obtained by the above-mentioned clustering.

Classification method	TFNB			PDNB		
Macro average	MacroP	MacroR	MacroF1	MacroP	MacroR	MacroF1
Small data set	0.877867	0.755318	0.811995	0.927331	0.922812	0.925066
Large data set	0.919698	0.883435	0.901202	0.942408	0.939173	0.940788

Figure 5.1: Macro average

From the above macro average of Table , it can again be seen that PDNB significantly better than TFNB method, especially in the small data set the MacroF1 value of PDNB is even 10 percentage points higher than TFNB method. The most important thing is that, in TFNB classification algorithm, the MacroF1 on small sample data set is lower than the large data sets nearly 9 percentage points. However, in PDNB classification algorithm, although the small sample data sets is 1/12 of the large-scale data set, its MacroF1 value only 1.5 percentage points lower than the large data set. Meanwhile, they almost have the same MacroR and MacroP value.

Here, for this analysis we select eight categories to do the test, including finance, health, sports, tourism, education, recruitment, culture and military etc as an example which has 1990 documents in each category. We select 1200 documents as the large-scale training data set for each category, and then select approximately 300 documents as the testing corpus form the remaining 790 documents of each category. In addition, we extract 100 represent.ative documents in each category as a small-scale training data set

Chapter 6

CONCLUSIONS

By introducing Poisson probability model for Naive Bayes Algorithm, each document is regarded as Poisson random variable generated by the multivariate Poisson model, and we did a series of comparative experiments in certain data set using the combination method of Poisson distribution model and Naive Bayes. The experimental results show that it has good classification performance in the small sample data set. However, the traditional Bayesian classification algorithm based on word frequency in the two different data sets has a very different classification effect.

The obvious advantages of Bayesian method are efficient, fast, and SVM is regarded as a good classification method, but consuming too much time and space. Therefore, the future work will focus on the comparative analysis of the differences between them in effectiveness and efficiency.

References

[1] **www.ieeexplore.ieee.org**

[2] **https://youtu.be/sjUDlJfdnKM-hackerearth.**

[3] .wiki.orgwww.wiki.org

[4] **Introduction to Machine Learning-Alex Simola and S.V.N. Viswanatha**

[5] **Yuguang Huang, Lei Li ,Beijing University of Posts and Telecommunications, Beijing, China,NAIVE BAYES CLASSIFICATION ALGORITHM BASED ON SMALL SAMPLE SET**