RESEARCH ARTICLE OPEN ACCESS

A Comparative Study of Centroid-Based and Naïve Bayes Classifiers for Document Categorization

Rupali P. Patil*, R. P. Bhavsar** and B. V. Pawar**
**(Department of Computer Science, S. S. V. P. S's L.K. Dr. P. R. Ghogrey Science College Dhule, Maharashtra, India*
*** (School of Computer Sciences, North Maharashtra University Jalgaon, Maharashtra, India*

ABSTRACT
Assigning documents to related categories is critical task which is used for effective document retrieval. Automatic text classification is the process of assigning new text document to the predefined categories based on its content. In this paper, we implemented and performed comparison of Naïve Bayes and Centroid-based algorithms for effective document categorization of English language text. In Centroid Based algorithm, we used Arithmetical Average Centroid (AAC) and Cumuli Geometric Centroid (CGC) methods to calculate centroid of each class. Experiment is performed on R-52 dataset of Reuters-21578 corpus. Micro Average F1 measure is used to evaluate the performance of classifiers. Experimental results show that Micro Average F1 value for NB is greatest among all followed by Micro Average F1 value of CGC which is greater than Micro Average F1 of AAC. All these results are valuable for future research.

Keywords: A Naïve Bayes (NB), Centroid Based (CB), Arithmetical Average Centroid (AAC), Cumuli Geometric Centroid (CGC)

I. INTRODUCTION

Information stored on the web is in digital form. This digital information stored on the web is growing rapidly. Every day more and more information are available. Documents on the web consist of huge amount of information which can be easily accessible. Today web is the main source of information. Generally textual information is available on web. Day by day these textual data are increasing continuously. Searching information in this huge collection is very difficult and there is a need to proper organization of this information. It can be handled by automatic text categorization. Automatic text classification is the process of assigning new text document to the predefined categories based on its content. Automatic classification is the primary requirement of text retrieval system [1]. Text classification has number of applications such as email filtering, topic spotting for news wire stories, language identification, question answering, business document classification, web page classification, document management and so on. In the early days each incoming document analyzed and categorized manually by domain expert. In order to carry this work large amount of human resources have been spent. It is very expensive to assist the process of classification. Automatic classification schemes are required. The goal of classification is to learn such classification algorithms that can be used to classify text document automatically.

There are number of text classification algorithms are available. This includes K-Nearest-Neighbor, Decision tree, Rocchio algorithm, Neural Network, Support Vector Machines and Naïve Bays etc. This paper describes Naïve Bayesian (NB) approach and Centroid Based approach for automatic classification of Reuters-21578. The Naïve Bayes is one of the simple, efficient and still effective algorithms for text document classification and has produced good results in previous studies. The Centroid Based approach required less computational time therefore it is widely used in different web application like language identification, Pattern recognition etc. The rest of the paper organized as follows. Section II reviews previous works. Related terms are described in section III. Section IV gives the concept of Centroid-based and Naïve Bayes classifiers. Experimental setup is described in section V. In section VI our experimental results are discussed. The last section concludes the paper.

II. REVIEW OF PREVIOUS WORK

This section briefly reviews related work on text classification. Text classification is a Natural Language processing problem. In [2] researcher applied supervised classification using NB classifier to Telugu news articles and it is found that the performance obtained is comparable to published results for other languages. Previous studies suggest that NB is simple till produces good results. In [3] authors compared their work to automatically classify Arabic documents using NB algorithm with previous study of [4]. In this work the average accuracy over all categories is 68.78% in cross validation and 62% in evaluation set experiments, it is comparable to the corresponding performance in [4] which are 75.6% and 50% respectively on the same dataset, categories & different root extraction algorithms. Web site classification using machine learning is necessary to automatically maintain directory services for the web for that author in [5] applied NB approach to classify web sites based on the contents of their home pages & yielded 89.05% accuracy. It is also observed that the classification accuracy of the classifier is proportional to number of training documents. To solve the problem that multiple occurrences of the same word in a document could reduce probability of other important features which have few occurrences, some modifications are done by researcher in [6] for that researcher adopt a new expression for words counts and thus tried to improve the performance of Naïve Bayes and the effect on categorization. Researcher adopts this algorithm in spam filter categorization the experimental result shows that the improved Naive Bayes algorithm is more effective than traditional Naïve Bayes. Thus, to improve the classification performance some researchers made modification in existing text classification techniques where as some researchers tried to combine different text classification techniques and generate a new hybrid technique. Paper in [7] presented the study of hybrid feature selection techniques and hybrid text classification techniques found in literature. In [8] researcher design Class-Feature-Centroid (CFC) classifier for multiclass, single-label text categorization and compared its performance with SVM and centroid based approaches. Their experiment on Reuters-21578 corpus and 20-newsgroups email collection shows that CFC outperforms SVM and centroid based approaches with both micro-F1 and macro-F1 scores. Additionally researchers show that when data is sparse, CFC has much better performance and is more robust than SVM. A simple linear-time centroid –based document classification algorithm is focused in [9]. Their experiments show that centroid-based classifier consistently and substantially outperforms other algorithms such as NB, KNN and C4.5 on a wide range of datasets. Their analysis shows that the similarity measure of the centroid-based scheme accounts for dependencies between the terms in the different classes and this is the reason why it consistently outperforms other classifiers. In [10] author describes three term distribution (among classes, within classes and in the whole collection) and investigates how this term distribution contributes to weight each term in documents. Several centroid-based classifiers are constructed with different term weighting using various datasets; their performances are investigated compared to a standard centroid-based classifier (TFIDF) and centroid-based classifier modified with information gain, NB and KNN. They showed that the effectiveness of term distribution to improve classification accuracy is explored with regard to the training set size and the number of classes. English is dominant language on web; most of the works in the area of text classification are done for English language text. Now a day, because of the rapid growth in use of Internet in India, work on Indian languages text classification is also started. Discussion on Indian languages text classification has been presented in [11]. From literature it is found that text classification is an important research area.

III. RELATED TERMS

Vector Space Model: vector space model is commonly used in information retrieval and specifically for document categorization. In this model a set of m unique terms are represented by m dimensional vectors $\vec{w}$ = (w_1, w_2, w_3, ..,w_m). To calculate weight w_i of term t_i is generally calculated by using $tf_i idf_i$ weighting scheme where tf_i is term frequency of term t_i i.e. the number of occurrence of term t_i in any document x and inverse document frequency idf_i which is calculated as log (N/D) where N is the total number of documents in the collection and D is the number of documents containing the term t_i. Therefore weight w_i of term t_i belonging to the document x can be calculated as

$$w_i = tf_i * idf_i$$

Similarity Measure:Cosine similarity is one of the similarity measures that can be used to measure similarity between two documents d_1 and d_2. The function is written as

$$\cos(d_1, d_2) = \frac{d_1 . d_2}{|d_1| * |d_2|}$$

Where $d_1 . d_2$ is the dot product of the two document vectors d_1 and d_2. $|d_1|$ and $|d_2|$ are the length of vector d_1 and vector d_2 respectively. In our paper we have used cosine similarity as the similarity measure.

IV. CLASSIFIERS

Centroid Based Classifier: CB classifier is one of the commonly used simple but effective document classification algorithm. In CB Classifier, centroid vector for each category in training phase is calculated from set of documents of the same class. Suppose there are m predefined classes in training data set then there are total m centroid vectors {$\vec{c_1}, \vec{c_2}$, ,..,$\vec{c_m}$} and each $\vec{c_i}$ is the centroid of class i.

For calculation of centroid vector two commonly used methods are

a. Arithmetical Average Centroid (AAC): Most commonly used initialization method for centroid based classifier

$$\vec{c_i} = \frac{1}{|c_i|}\sum_{d\in c_i}\vec{d}$$

where centroid is the arithmetical average of all document vectors of class c_i

b. Cumuli Geometric Centroid (CGC):

$$\vec{c_i} = \sum_{d\in c_i}\vec{d}$$

where each term will be given a summation weight.

In testing phase, test document d_t is classified by finding similarity of testing document vector $\vec{d_t}$with centroid vector $\vec{c_i}$ of each category $\{\vec{c_1},\vec{c_2}, .. ,\vec{c_m}\}$ using cosine similarity and assign the test document d_t to the category having maximum similarity. That is d_t is assign to the class by using

$$\arg\max_{i=1..m}\left(cos\left(\vec{d_t},\vec{c_i}\right)\right)$$

NAÏVE BAYES CLASSIFIERS (NB)

Bayesian is based on Bayes theorem. It is a supervised learning method as well as statistical method. It is used as a probabilistic classifier. It is one of the most successful classifier applied to text document classification. It is very simpler classifier. Previous studies suggest that its performance is comparable with decision tree classifiers and Neural Network classifiers. The basic approach of NB is to use the joint probabilities of words and categories to estimates the probabilities of categories given a document. The conditional probability of a word given a category is assumed to be independent from the conditional probabilities of other words given that category. NB algorithm computes the posterior probabilities that document belong of different classes and assign it to the class with the highest posterior probability.

Let D= $\{d_1,d_2,\ldots,d_p\}$be set of documents. C= $\{c_1, c_2,\ldots,c_q\}$ be set of classes. The probability of a document d being in class c using Bayes theorem is

$$C = arg\max_{c\in C}\frac{P(c)*P(d|c)}{P(d)}$$

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

$$=\arg\max_{c\in C} P(c) * P(d|c)$$

$$=\arg\max_{c\in C} P(c)\prod_{1\le k\le n_d} P(t_k|c)$$

Where $P(t_k|c)$ is the conditional probability of term t_k occurring in a document of class c. $<t_1, t_2, \ldots, t_{n_d}>$ are n_d number of tokens in d which are included in vocabulary used for classification.

$P(c)$, the prior probability of c

$$P(c) = \frac{N_c}{N}$$

where N_c the number of training documents in class c, N is total number of training documents.

Conditional probability $P(t|c)$ as the relative frequency of term t in a document d belonging to class c:

$$P(t|c) = \frac{T_{ct}}{\sum_{t'\in V} T_{ct'}}$$

Here T_{ct} is the number of occurrences of t in training documents from class c. And $\sum_{t'\in V} T_{ct'}$ is the total number of term in all positions k in the documents in training set.

Laplace Smoothing:

A term class combination that does not occur in the training data makes the entire result zero. To solve this problem we use add-one smoothing or Laplace Smoothing.

$$P(t|c) = \frac{T_{ct}+1}{\sum_{t'\in V} T_{ct'} + |V|}$$

$|V|$is the no. of terms in the vocabulary.

Underflow Condition:

Many conditional probabilities between 0 & 1 are multiplied this can result in a floating point underflow. Since log(xy) =log(x) +log(y) so rather than multiplying we add logs of probabilities. Class with highest probability score is most suitable.

The Performance Measure:

In our experiment, for evaluation of classifiers commonly used performance measure Micro Averaged F1 has been used. The detail explanation of the performance measure has been presented in [12].

V. EXPERIMENTAL SETUP

In our experiment we used standard R-52 dataset of Reuters-21578 corpus. It is available on web link in [13]. We applied stop words removal. There are total 52 categories available for training and testing purpose. There are total 6532 documents in training folder and total 2568 documents in testing folders of 52 categories. The Centroid-based and Naïve Bayes classifier software has been developed by us using Java Programming. Distribution of documents across the categories is as shown in Figure1.

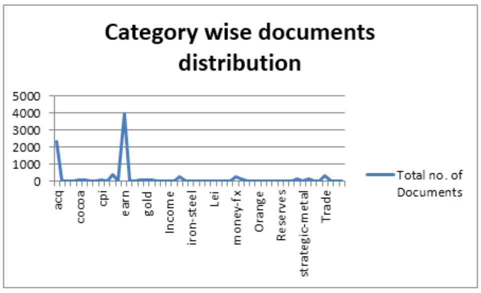

Figure 1. Category wise distribution of documents for Reuters-21578 corpus

Comparison of classifier:

Table 1: summary of performance of NB, AAC and CGC classifier

Classifier	Micro Avg. Precision (%)	Micro Avg. Recall (%)	Micro Avg. F1 (%)
NB	87.69	91.21	89.42
AAC	85.51	85.51	85.51
CGC	85.98	85.71	85.85

Table 1 summarizes the global performance scores. In our experimental trials we have obtained Micro Average Precision of 87.69%, 85.51% and 85.98% for NB, AAC and CGC respectively. The Micro Average Precision for NB is higher than other two methods of Centroid Based classifiers. This indicates that the NB method perform high precision. This is shown graphically in Figure 2.

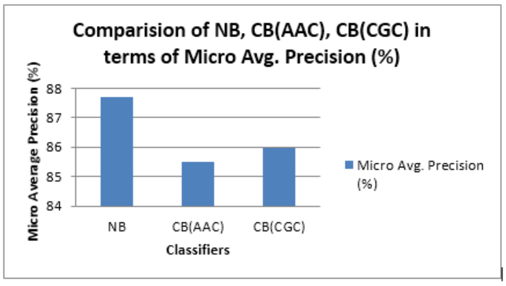

Figure 2. Comparison of NB, CB (AAC) and CB (CGC) in terms of Micro Average Precision

The Micro Average Recall of NB, AAC and CGC are 91.21%, 85.51% and 85.71% respectively. Micro Average Recall of NB is higher than Micro Average Recall of other two methods of Centroid Based classifier. This indicates that NB outperforms AAC and CGC classifiers in terms of Recall which is shown in Figure 3.

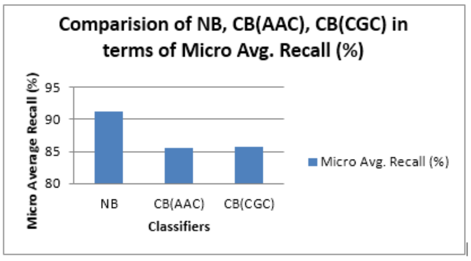

Figure 3 Comparison of NB, CB (AAC) and CB (CGC) in terms of Micro Average Recall

Micro Average F1 of NB is 89.42% which is greater than Micro Average F1 of AAC and Micro Average F1 of CGC which is 85.51% and 85.85% respectively. It indicates that NB outperforms AAC and CGC in terms of Micro Average F1. This is shown graphically in Figure 4.

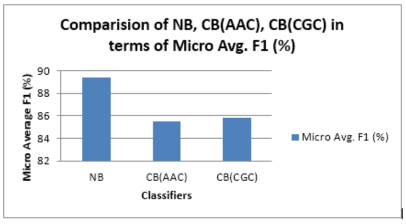

Figure 4 Comparison of NB, CB (AAC) and CB (CGC) in terms of Micro Average F1

VI. CONCLUSION

In this paper, we have described our experimental study of two well-known classifiers NB and CB (In CB for centroid calculation AAC and CGC methods are used.) for English language text categorization on R-52 of Reuters21578 corpus. We have compared the performance of NB and CB classifiers. No feature selection was applied. The experimental results showed that the performance of the Naïve Bayes classifier is best among all classifiers with 89.42% Micro Average F1 measure. Out of AAC and CGC centroid calculation methods, CGC has obtained better performance with 85.85% Micro Average F1 measure than AAC (85.51%). We have observed that the classification speed of NB is very fast among all. In future we will test the effect of different weighting scheme on the performance of CB (AAC and CGC). Also we will try to adapt some different document classification algorithm to solve the document categorization problem more efficiently.

REFERENCES

[1] *S. M.Kamruzzaman, "Text Classification using Data Mining", Proc. International Conference on Information and Communication Technology in Management (ICTM-2005), Multimedia University, Malaysia, May 2005*

[2] *Kavi Narayana Murthy, "Automatic Categorization of Telugu News Articles", doi: 202.41.85.68.*

[3] *Kourdi Mohmed EL, Benasid Amine, Rachidi Tajje-eddine, "Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm", Semitic '04 Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages Pages 51-58, Association for Computational Linguistics Stroudsburg, PA, USA 2004.*

[4] *M. Yahyaoui, "Toward an Arabic web page classifier," Master project. AUI. 2001.*

[5] *Ajay S. Patil, B. V. Pawar, "Automated Classification of Web Sites using Naïve Bayesian Algorithm" , Proceedings of the International MultiConference of Engineers*

and Computer Scientists 2012 Vol I, IMECS 2012, March 14-16, 2012, Hong Kong.

[6] *Guoqiang, "An Effective Algorithm for Improving the Performance of Naïve Bayes for Text Classification", Second International conference on Computer Research and Development DOI 10.1109/ICCRD.2010.160, 2010 IEEE.*

[7] *Rupali Patil, R. P. Bhavsar and B. V. Pawar, "Holy Grail of Hybrid Text Classification", IJCSI International Journal of Computer Science Issues, Volume 13, Issue 3, May 2016.*

[8] *Hu Guan, Jingyu Zhou, Minyi Guo, "A Class-Feature-Centroid classifier for Text Categorization" WWW 2009, April20-24, 2009, Madrid, Spain. ACM 978-1-60558-487-4/09/04*

[9] *Eui-Hong(Sam) Han and George Karypis, "Centroid-Based Document Classification: Analysis & Experimental Results", Principles of Data Mining and Knowledge Discovery, p 424-431, 2000.*

[10] *Veryuth Lertnattee, Tjanaruk Theeramunkong, "Effect of term distributions on centroild-based text categorization", Information Sciences 158 (2004) 89-115. Doi:10.1016/j.ins.2003.07.007.*

[11] *Rupali P. Patil, R. P. Bhavsar, B. V. Pawar, "A NOTE ON INDIAN LANGUAGES TEXT CLASSIFICATION SYSTEMS", Asian Journal of Mathematics and Computer Research, 15(1): 41-55, 2017,* ***ISSN No. : 2395-4205 (Print), 2395-4213 (Online)****.*

[12] *Rupali P. Patil, R. P. Bhavsar, B. V. Pawar, "A Comparative Study of Text Classification Methods: An Experimental Approach", International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC), March 16 Volume 4 Issue 3, ISSN: 2321-8169, PP: 517 – 523.*

[13] *www.cs.umb.edu/`smimarog/textmining/datasets/SomeTextDatasets.html.*