Vol 6 • Number 2 April - Sep 2015 pp. 196-200 **Impact Factor: 2.5** **Available at** www.csjournals.com

Modified Approach of Multinomial Naïve Bayes for Text Document Classification

S.W. Mohod, Dr. C.A.Dhote, Dr. V.M. Thakare
Deptt. Computer Engineering, B.D. College of Engg. Sevagram, Wardha, India
Prof.,Ram Meghe Institute of Technology & Research, Badnera.Amravati, India
P.G.Department of Computer Science, Sant Gadge Baba Amravati University, Amravati, India

Abstract: This work proposes a text classification using modified approach of Multinomial Naïve Bayes for justifying and identifying the documents into a particular category. Due to the exploration of the textual information from the electronic digital documents as well as World Wide Web. Naïve Bayes theorem is effective for classification of text documents into the predefined categories by means of the probabilistic values. However, its performance is repetitively inadequate by inappropriate feature selection. The aim of this paper is to propose a method that will improve the classification accuracy decision. In addition a new feature selection method for text document classification in machine learning is also proposed. In machine learning the training set is generated for testing the documents. Scoring method is used to enhance the efficiency of both classifications with a relevance to accuracy and performance.
Keyword: Text classification, Naïve Bayes, Feature selection.

Introduction

With the increasing availability of electronic documents and rapid growth of the World Wide Web and data in digital format, the task of automatic document classification is important for organization. Proper classification of electronic documents, online news, blogs, e-mails and digital libraries requires Text Mining, Machine learning and natural language processing techniques to extract required knowledge information. Text mining makes an attempt to discover interesting information and knowledge from unstructured documents. The important task is to develop the automatic classifier to maximize the accuracy and efficiency to classify the existing and incoming documents.

Voluminous information of an organization is stored in an unstructured form of reports messages, news and email [1]. However, data mining deals with structured data, whereas text presents special characteristics and is unstructured. The important task is how these documented data can be properly retrieved, presented and classified it is difficult to machine, so there has been a growing interest in this area of research [2]. Extraction, integration and classification of text documents from different sources and knowledge information discovery which finds features from available documents are important.

In data mining, Machine learning is often used for Prediction or Classification. Classification involves finding rule that partitions the data into disjoint groups. The input for the classification is the training data set, whose class labels are already known. Classifications analyze the training data set and construct a model based on the class label. The goal of classification is to build a set of models that can correctly predict the class of the different objects. Machine learning is an area of artificial intelligence concerned with the development of techniques which allow computers to "learn". More specifically, machine learning is a method for creating computer programs by the analysis of data sets since machine learning study the analysis of data. The challenging task is of text classification performance, because many problems are due to high dimensionality of feature space and unordered collection of words in text documents. Various machine learning algorithms available and utilize in document classification. Naïve Bayes has been one of the popular machine learning algorithm because of its simplicity. [2][3][4] easy to implement and draws better accuracy in large datasets[5]. Naïve Bayes classifier performing well in classification task where the probability is calculated by the Naïve Bayes independent assumption [6] [7]. The paper mainly focuses on reducing the number of features class dependent by using the text document feature selection with new feature scoring method and using proposed feature selection method implement Modified approach of Multinomial Naïve Bayes classification model to classify testing text documents.

Thousands of term word occurs in the text document, so it is important to reduce the dimensionality of feature using feature selection process [8], to resolve this problem many feature evaluation metrics have been explored such as X2 Statistics (CHI), Information Gain (IG), mutual information, term strength, document frequency, Term Frequency Inverse Document Frequency. With the help of these approaches it is possible to reduce the high dimensionality of features. Proposed feature scoring metrics to select the feature is the most effective method to reduce the dimensionality of feature and improve the efficiency and accuracy of classifier. In this approach document preprocessing is also important to reduce the complexity and high dimensionality of features occurs in the text document.

Feature selection Approaches

Feature selection helps in the problem of text classification to improve efficiency and accuracy. In our approach we are examining different feature selection methods and then will find wheather our proposed method is effective to other studied method.

A. TF (Term Frequency)

Term frequency in the given document is simply the number of times a given term appears in that document. TF used to measure the importance of item in a document, the number of occurrences of each term in the document. Every document is described as a vector consisting of words.
Importance of the term 't' within the particular document with 'ni' being the number of occurrences of the considered term and the denominator is the number of occurrences of all terms.

$$TF = \frac{ni}{\sum_K n_k} \qquad (1)$$

B. AC(T) (Average of Class Term)

Average of class term calculated using not only the term appear in the given document number of times divided by number of document. It is calculated by how many times the term appear in the corpus documents divided by no of classes in that corpus.

C. DF (Document Frequency)

One way of calculating the document frequency (DF) is to determine how many documents contain the term't' divide by the total number of documents in the collection. |D| is the total no of documents in the document set D, and | {di tj $\in$ di $\in$ D}| is the number of documents containing term tj.

D. IDF (Inverse Document Frequency)

The inverse document frequency is a measure of the general importance of the term in the corpus. It assigns smaller value to the words occurring in the most of the documents and higher values to those occurring in fewer documents. It is the logarithm of the number of all documents divided by the number of documents containing the term.

$$IDF = \log \frac{|D|}{|(\text{di} \supset \text{ti})|} \qquad (2)$$

Where |D| is total no of documents in the corpus & $|(di \supset ti)|$ is number of documents where the term 'ti' appears.

E. TFIDF (Term Frequency Inverse Document Frequency)

TFIDF is commonly used to represent term weight numerically using multiplication of term frequency and inverse document frequency [9].

It is possible to do better term weighing by multiplying tf values with IDF values, by considering local and global information. This is commonly referred to as, TFIDF score (weighting).

TFIDF(t) =TF(d,t) X IDF(t) (3)

F. Our Approach : ACTIDFCP (Average of Class Term multiply by Inverse Document Frequency plus one minus Conditional Probability divide by Average of Class Term)

$$\text{ACTIDFCP} = \frac{(AC(T) * IDF) + (1 - \text{P}(\text{t} \mid c))}{AC(T)} \qquad (4)$$

$AC(T)$ -is the average of class term among all the training classes documents.

IDF - Inverse document Frequency

$\text{P}(\text{t} \mid c)$ -Estimate the conditional probability as the relative frequency of term 't' in test documents belonging to class c.

TFIDF is commonly used to represent term weight numerically using multiplication of term frequency and inverse document frequency [10]. Proposed ACTIDFCP is commonly used to Average of class term multiply by Inverse Document Frequency plus one minus Conditional Probability Divide by average of class term. Here we can also get the numerical value and assign to the related term. Using these we can select the relevant (important) term from the total number of features in the corpus. The proposed method of feature selection to train the classifier is described is as follows.

Step1: Collect the standard Data sets for text classification.
Step2: Text tokenization:
1. Its conversion is very important for separating the sentences into words.

2. Breaking a text into tokens. A token is a non-empty sequence of characters, excluding spaces and punctuation.
3. Lowercase conversion converting all the character in a document into the same case
4. Special character removal (+, -, !, ?, }, …., etc.) and digits then convert into word string.

Step3: Filtration:
Remove all stop words using already well defined Blockade list.
Step4: Stemming:
In this process system removes the word's prefixes and suffixes. Applying stemming algorithm.
Step5: We get the highly relevant words from the document.
Step6: All the terms and frequencies are collected from each document. Evaluate and retain values of TF.
Step7: Repeat Step2 to Step6 for all the documents of corpus.
Step8: Evaluate and retain all values of DF, IDF, and ACTIDFCP.
Step9: Obtain the word feature set of corpus.
Step10: According the ACTIDFCP perform sort operation in ascending order on ACTIDFCP score with conditional probability.
Step 11: Retain Database for testing

Figure1 Proposed Feature selection method used for training.

Classification Task

In our system we used the number of top most features from the training subset for the classification purpose. Using the number of features from training set (Modified) Hybrid Multinomial Naïve Bayes (HMNB) classifier would be decided the category of that document.

A. Multinomial Naïve Bayes Classifier

Multinomial NB model is the supervised learning method, a probabilistic learning method. The probability of a document d being in class c is computed as

$$P(c \mid d) \propto P(c) \prod_{1 \le i \le nd} P(t_i \mid c) \quad (5)$$

Where $P(t_i \mid c)$ is the conditional probability of term t_i occur in a document of class c. $P(c)$ is the prior probability of a document occurring in class c.

In the multinomial model a document di is an ordered sequence of term events, drawn from the term space T. The naive Bayes assumption is that the probability of each term event is independent of term's context, position in the document, and length of the document. So, each document di is drawn from a multinomial distribution of terms with number of independent trials equal to the length of di. The probability of a document di given its category cj can be approximated as:

$$P(d_i \mid c_j) \approx \prod_{i=1}^{|di|} P(t_i \mid c_j) \quad (6)$$

where |di| is the number of terms in document di; and ti is the i^{th} term occurring in document di. Thus the estimation of P(di|cj) is reduced to estimating each P(ti|cj) independently. The following Bayesian estimate is used for P(ti|cj) its also called as a conditional probability:

$$P(t_i \mid c_j) = \frac{1 + TF(ti, Cj)}{|T| + \sum_{tk \in T} TF(tk, Cj)} \quad (7)$$

Here, TF(ti,cj) is the total number of times term ti occurs in the training set documents belonging to category cj. The summation term in the denominator stands for the total number of term occurrences in the training set documents belonging to category cj. This estimator is called Laplace estimator and assumes that the observation of each word is a priori likely [11].

In a text classification our goal is to find the best class for the document. We do not know the true values of the parameters $P(c)$ and $P(t_i \mid c)$ but estimate them from the training set. In equation (5) many conditional probabilities are multiplied one for each position $1 \le i \le nd$.

This can result in a floating point underflow so we propose the Modified Multinomial Naïve Bayes to over come this problem.

B. Our Approach

To remove floating point underflow issue, Multinomial Naïve Bayes for text document classification is presented. We have proposed a modified version of Multinomial Naïve Bayes, This new version; we refer to it as Modified Multinomial Naïve Bayes (MMNB).

In equation (5) many conditional probabilities are multiplied one for each position $1 \le i \le nd$ therefore it is better to perform the addition of many conditional probability as per equation(5).

$$P(c \mid d) \propto P(c) \sum_{1 \le i \le nd} P(t_i \mid c) \quad (8)$$

The proposed MMNB method used in proposed feature selection to test the classifier is described is as follows.

Step1: Collect the test Data sets for testing.

Step2: Extract features from test document using Step2 to Step 4 as per Figure1.

Step3: Using Modified MNB Classifier assigns the test document class

Figure 2 Proposed Modified MNB classification model used for to classify testing.

Experimental Results

To evaluate the performance of our proposed method we have performed experiments on data set from R10 of Reuters 21578. The experiments were carried out on Pentium® Dual-Core CPU, 3GB RAM, Windows Vista 32-bit Operating System, and MATLABR2008a.

Table1 shows the detailed information of the data set. Performance of our proposed method using above mention dataset is shown in Fig. 3.

Using traditional document frequency and inverse document frequency algorithm calculates the numeric values for the term feature of the corpus. Also calculate the average term frequency within the class. With the help of these values proposed ACTIDFCP method is used to generate new numerical values for the corresponding term. These term values are ordered by their ACTIDFCP values in descending order. For creating the training set for testing the corpus documents, conduct feature selection by picking up top few terms. It has been observed that using top most minimum terms related to corpus generated using proposed ACTIDFCP method are relevant with the class of the data set.

TABLE1. DATA DESCRIPTION

DataSet R10 of Reuters 21578 (all-terms, user select 10 classes)										
Class	alum	bop	cocoa	cotton	gas	gold	jobs	livestock	oranje	rubber
Train Docs	31	22	46	15	10	70	37	13	13	31

	A	B	C	D	E	F	G	H	I	J
1	alum	bop	cocoa	cotton	gas	gold	jobs	livestock	oranje	rubber
2	aluminium	billion	cocoa	cotton	gasolin	gold	unemploy	cattl	orang	rubber
3	tonn	deficit	icco	agricultur	compani	ounc	januari	depart	juic	price
4	compani	current	buffer	crop	unlead	mine	februari	agricultur	frozen	produc
5	metal	surplu	stock	plant	expect	compani	adjust	dairi	depart	natur
6	aluminum	mln	intern	state	produc	ton	number	head	fcoj	intern
7	smelter	dlr	deleg	increas	dlr	grade	fell	announc	florida	pact
8	plant	januari	rule	mln	upgrad	product	season	program	brazil	consum
9	product	trade	organ	depart	octan	expect	total	export	citru	inra
10	price	rose	price	bale	convert	price	jobless	dlr	concentr	agreement
11	primari	balanc	produc	weather	ga	estim	rate	ccc	usda	confer
12	capac	offici	council	usda	refineri	mln	statist	bonu	final	malaysian
13	spokesma	figur	week	march	mid	or	fall	paid	estim	octob

Figure 3 Generated Feature Set Using ACTIDFCP

We want to determine which term in a given set of training feature vector is most important for discriminating between the classes to be learned. ACTIDFCP tells us how important a given term of the feature vectors.

We compare classification accuracy with two different algorithm mention in the paper to evaluate the effect of the improve MMNB algorithm shown in the Fig. 4.

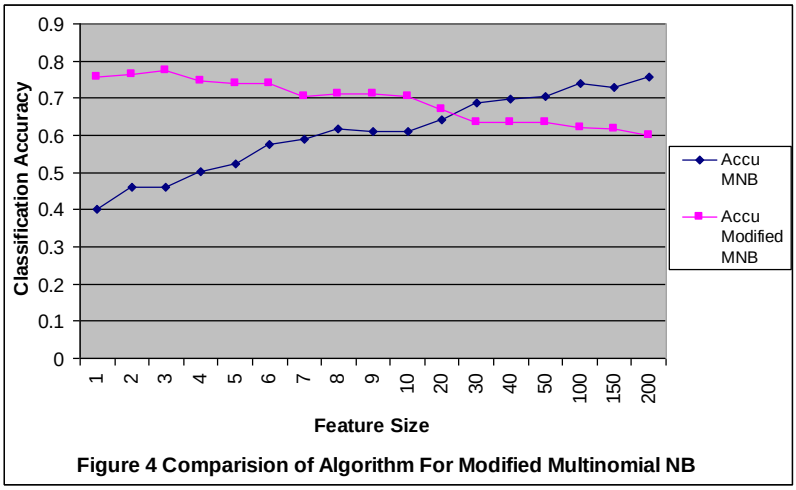

Figure 4 Comparision of Algorithm For Modified Multinomial NB

Conclusion

The proposed method is another approach for feature selection in text classification. R10 of Reuters 21578 data set s are used for experimentation. The proposed method performs well for feature selection. Hence the accuracy and performance in feature selection is enhanced by adopting the proposed method. The experimental results improve the accuracy with minimum number of features in training the class. Moreover, the minimum number of features achieves highest accuracy over the previous method which requires large number of features to achieve that accuracy.

References

[1] Raghavan P., S. Amer-Yahia and L. Gravano eds.., "Structure in Text: Extraction and Exploitation." In proceeding of the 7th Internatinal Workshop on the Web and Databases(WebDB):ACM SIGMOD/PODS 2004, ACM Press, 2004, Vol 67.

[2] Sang-Bum Kim, Kyong-soo Han, Hae-Chang rim, Sung Hyon Myaeng "Some Effective Techniques for Naïve Bayes Text Classification" IEEE Transactions on Knowledge and Data Engineering-2006

[3] D. Koller, and M. Sahami (1997), Hierarchically classifying documents using very few words. In Proceedings of ICML-97, 14th International Conference on Machine Learning (Nashville, TN, 1997), pp.170–178.

[4] D. D. Lewis, AND W. A. Gale (1994), A sequential algorithm for training text classifiers. In Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval (Dublin, Ireland, 1994), pp.3–12.

[5] Vidhya.K.A, G.Aghila "A Survey of Naïve Bayes Machine Learning approach in Text Document Classification" (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, No. 2, 2010.

[6] Tom M.Mitchell, "Machine Learning", Carnegie Mellon University, mcGraw-Hill Book Co,1997.

[7] Yi-Hsing Chang (2007), Automatically constructing a domain ontology for document classification, International Conference on Machine Learning and Cybernetics (ICMLC 2007), Hong Kong, China, Aug. 19-22, 2007.

[8] F. Beil, M. Ester, X. Xu, Frequent term-based text clustering, Proc. of Int'l Conf. on knowledge Discovery and Data Mining, KDD' 02, pp. 436–442, 2002.

[9] M. Yetisgen-Yildiz and W. Pratt, "The effect of feature representation on MEDLINE document classification," AMIA Annual Symposium Proceedings, pp. 849-853,2005.

[10] Ying Liu, Han Tong Loh, Kamal Youcef-Toumi, and Shu Beng Tor, "Handling of Imbalanced Data in Text Classification: Category-Based Term Weights," in Natural language processing and text mining, pp. 172-194.

[11] Joachims, T., "A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization", ICML-97, 1997.