A Novel Approach in Feature Selection Method for Text Document Classification

S.W. Mohod
Deptt. Computer Engineering,
B.D. College of Engg. Sevagram,
Wardha, India
sudhirwamanrao@gmail.com

Dr. C.A. Dhote
Prof., Ram Meghe Institute of Technology & Research,
Badnera.
Amravati, India
vikasdhote@rediffmail.com

Abstract—In this paper, a novel approach is proposed for extract eminence features for classifier. Instead of traditional feature selection techniques used for text document classification. We introduce a new model based on probability and over all class frequency of term. We applied this new technique to extract features from training text documents to generate training set for machine learning. Using these machine learning training set to automatic classify documents into corresponding class labels and improve the classification accuracy. The results on these proposed feature selection method illustrates that the proposed method performs much better than traditional methods.

Keywords:- *Text classification, Feature selection*

*****

I. Introduction

With the high availability of computing resources, large amount of data in digital format, the task of automatic document classification is important. Proper classification of online news, electronic documents, blogs, digital libraries and e-mails are required. Using feature selection method a number of application developed [1] with term weighing [2][3] and probabilistic model [4]. Based on the term frequency information, documents can be classified by using different classification algorithms such as naïve bayes [5][6],Decision tree[7], Support Vector Machine[8][9].

The important task is to develop a usual classifier to maximize the accuracy and efficiency to classify the existing and incoming documents.

Extraction, integration and classification of text documents from different sources and knowledge information discovery which finds pattern from available documents are important.

Various machine learning algorithms are available and work on document classification. Supervised text classification needs large quantities of labeled training data to achieve high accuracy [9]. In this paper we mainly focus 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 to implement for classifying text documents appropriately using less number of features to achieve high accuracy.

II. Background OF Document Representation

One of the pre-processing technique is feature selection is important and mostly used in text mining to reduce the number of features in the documents. Feature selection has been a field of research and in existence since 1970s in Machine learning and data mining [10][11]. Feature selection is a very important step in text classification, because inappropriate and unneeded features often degrade the performance of classification both in speed and classification accuracy. Feature reduction technique can be classified into feature extraction (FE) [12] and feature selection (FS) approaches given below.

A. *Feature Extraction*

FE is the first step of pre-processing which is used to presents the text documents into plain word format. So removing stop words and stemming words is the pre-processing tasks [13] [14]. The documents in text classification are represented by a great amount of features and most of them could be irrelevant or noisy [15]. DR is the exclusion of a large number of keywords, stand preferably on a arithmetical process, to create a low dimension vector [16]. Commonly the steps taken for the feature extractions (Fig.1) are:

Tokenization: A document is treated as a string, and then partitioned into a list of tokens.

Removing stop words: Stop words such as “the”, “is”, “and”… etc are frequently occurring, so the insignificant words need to be removed.

Stemming word: Applying the stemming algorithm that converts different word form into similar canonical form. This step is the process of different tokens to their root form, e.g. connection to connect, computing to compute etc.

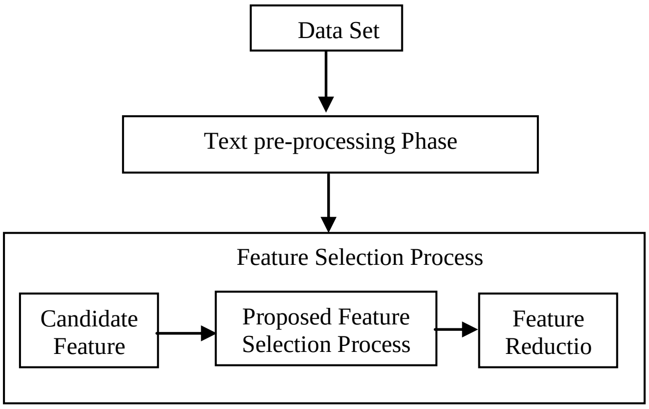
Data Set
Text pre-processing Phase
Feature Selection Process
Candidate Feature
Proposed Feature Selection Process
Feature Reductio

Figure 1. Feature selection Process

B. Feature Selection

After feature extraction the important step in preprocessing of text classification, is feature selection to create vector space, which improve the efficiency, scalability and accuracy of a text classifier. In general, a good feature selection method should believe domain and algorithm characteristics[17]. The main idea of FS is to select subset of features from the original documents. Feature selection is performed by keeping the words with highest score according to predetermined measure of the significance of the word [15]. The selected feature retains the unique substantial meaning to provide a better understanding for the data and learning process [18]. For text classification a major problem is the high dimensionality of the feature space. Nearly every text domain has a large amount number of features, most of these features are not relevant and beneficial for text classification task, and even some noise features may stridently reduce the classification accuracy [19]. Hence FS is commonly used in text classification to reduce the dimensionality of feature space and improve the efficiency and accuracy of classifiers.

In our approach TFIDF and Proposed method is compared with data set R10 of Reuters 21578 (User selected randomly ten classes from Reuters 21578).

III. Feature selection Approaches

Feature selection helps in the problem of text classification to improve efficiency and accuracy. We are examining different feature selection methods and then analyzing that the proposed method is effective in comparison with the other existing methods.

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.

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 ∈ di ∈ 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.

$$IDF = \log \frac{|D|}{|\{\mathrm{di\ tj} \in \mathrm{di} \in \mathrm{D}\}|}$$

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

E. IDFDDF (Inverse Document Frequency Divide by Document Frequency)

Using the inverse document frequency and document frequency it is the division of IDF and DF, it is denoted as IDFDDF [20].

$$\mathrm{IDFDDF} = \frac{\log \dfrac{|D|}{|\{\mathrm{di\ tj} \in \mathrm{di} \in \mathrm{D}\}|}}{|\{\mathrm{di\ tj} \in \mathrm{di} \in \mathrm{D}\}|}$$

TFIDF is commonly used to represent term weight numerically using multiplication of term frequency and inverse document frequency [18]. IDFDDF is commonly used to represent inverse document frequency divided by document frequency [20]. 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 is described is as follows.

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

$$\mathrm{ACTIDFDDFC\ P} = \frac{(AC(T) * IDFDDF) + (1 - \mathrm{P}(\mathrm{t}\,|\,c))}{AC(T)}$$

$AC(T)$ -is the average of class term among all the training classes' documents.

IDFDDF - Inverse document Frequency Divide by Document Frequency.

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

TFIDF is commonly used to represent term weight numerically using multiplication of term frequency and inverse document frequency [12]. Proposed ACTIDFDDFCP is commonly used to Average of class term multiply by Inverse document Frequency Divide by 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 important term from the total number of features in the corpus.

IV. Classification Task

Here we use the Multinomial Naïve Bayes classifier to decide the class of testing documents. We use the number of top most features from proposed feature selection method and traditional method for the classification purpose.

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

$$\mathrm{P(c\,|\,d)} \propto P(c)\prod_{1\le i\le nd} P(t_i\,|\,c) \qquad (1)$$

Where $P(t_i\,|\,c)$ 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:

$$\mathrm{P(d_i\,|\,c_j)} \approx \prod_{i=1}^{|di|} P(t_i\,|\,\mathrm{c_j}) \qquad (2)$$

where |di| is the number of terms in document di; and ti is the ith 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:

$$\mathrm{P(t_i\,|\,c_j)} = \frac{1+TF\,(ti,Cj)}{|\,T\,|+\sum_{tk\in T} TF\,(tk,Cj)} \qquad (3)$$

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 [21].

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 $P(t_i\,|\,c)$ but estimate them from the training set.

V. Experimental Results

In this paper, the document classification experiments performed to compare proposed feature selection method with traditional feature selection method. The data set used in this research is Reuter 21578 belongs to number of different topics. Here, 10 topics as alum, bop, cocoa, cotton, gas, gold, jobs, livestock, oranje and rubber are selected. Table1 shows the detailed information of the data set.

Using preprocessing step to reduce the dimensionality of original text documents the documents are converted into the vectors after which removes the non-alphanumeric characters and then insignificant words (keywords) called as filtration process hence removes the noisy elements from the term vector. After removal of noisy elements perform stemming process which is important for the text document feature selection. In this process remove the lexicons i.e. s, es, ing, ed, est etc. Stemming process generates the different word in single form in which we get the original term features of the corpus. 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 ACTIDFDDFCP method is used to generate new numerical values for the corresponding term. These term values are ordered by their ACTIDFDDFCP 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 ACTIDFDDFCP method are relevant with the class of the data set.

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. ACTIDFDDFCP tells us how important a given term of the feature vectors.

We compare classification accuracy with different feature selection methods mention in section III.

TABLE I. 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

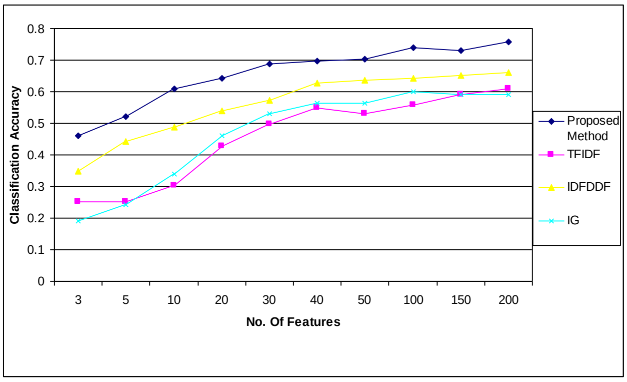

Figure 2. Comparison of Traditional Feature Selection Method With Propose method usingMNB

VI. Conclusions

The proposed method is an another approach for feature selection for text classification. Reuter 2157 data set used for experimentation. Experiment performs only 10 classes of given dada set. 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 than traditional methods such as TFIDF and IG. Proposed method improves the classification accuracy using MNB classification algorithm.

References

[1] Y. Yang and J. O. Pedersen, A comparative study on feature selection in text categorization, in Proceedings of the 14th International Conference on Machine Learning, Nashville, US, 1997.

[2] M. Lan, C. Tan, H. Low and S. Sungy, A comprehensive comparative study on term weighting schemes for text categorization with support vector machines, in

Proceedings of the 14th international conference on World Wide Web, pages 1032-1033, 2005.

[3] M.Ikonomakis, S. Kotsiantis et.al., “Text Classification Using Machine Learning Techniques”, WSEAS Transactions on Computers, Issue 8, Volume 4, August 2005, pp. 966-974.

[4] M.E. Maron and J.L. Kuhns, “On relevance, probabilistic indexing and information retrieval,” Journal of the ACM, vol. 7, pp. 216–244, 1960.

[5] Anirban Dasgupta et.al “Feature Selection Methods for Text Classification”, KDD’07, San Jose, California, USA. Copyright 2007 ACM 978-1-59593-609-7/07/0008 August 12–15, 2007

[6] Y. Yang and X. Liu, A re-examination of text categorisation methods, in Proc. 22nd Int. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR99), pages 67-73, 1999.

[7] C. Apte, F. Damerau, and S. Weiss, Text mining with decision rules and decision trees, in Workshop on Learning from text and the Web, Conference on Automated Learning and Discovery, 1998.

[8] S. Tong and D. Koller, Support Vector Machine Active Learning with Applications to Text Classification, Journal of Machine Learning Research, Volume 2, pages 45-66, 2001.

[9] M.M. Larry and Y. Malik, 2001, “One-class SVMs for document classification,” Journal of Machine Learning Research, 2:139-154.

[10] HAN Hong-qi,ZHU Dong-hua, WANG Xue-feng,”Semi-supervised Text Classification from Unlabeled Documents Using Class Associated Words”, IEEE, 2009.

[11] Tseng V. S.,Ja-Hwung Su, Hao-Hua Ku, Bo-Wen Wang, “ Intelligent Concept-Oriented and Content base Image Retrival by using data mining query decomposition techniques” IEEE International Conference on Multimedia and Expo. June23 2008-April 26 2008 Page(s):1273-1276.

[12] Tao Jiang, Ah-hwee tan, Senior Member, IEEE, and Ke Wang “Mining Generalized Associatations of Sementic Relations from Textual Web Content” IEEE Transaction on Knowledge and Data Engineering, Vol 19, no. 2, February 2007.

[13] 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.

[14] Wang, Y., and Wang X.J., “ A New Approach to feature selection in Text Classification”, Proceedings of 4th International Conference on Machine Learning and Cybernetics, IEEE- 2005, Vol.6, pp. 3814-3819, 2005.

[15] Lee, L.W., and Chen, S.M., “New Methods for Text Categorization Based on a New Feature Selection Method a and New Similarity Measure Between Documents”, IEA/AEI,France 2006.

[16] Montanes,E., Ferandez, J., Diaz, I., Combarro, E.F and Ranilla, J., “ Measures of Rule Quality for Feature Selection in Text Categorization”, 5th international Symposium on Intelligent data analysis , Germeny-2003, Springer-Verlag 2003, Vol2810, pp.589-598, 2003.

[17] Manomaisupat, P., and Abmad k., “ Feature Selection for text Categorization Using Self Orgnizing Map”, 2nd International Conference on Neural Network and Brain, 2005,IEEE press Vol 3, pp.1875-1880, 2005.

[18] Zi-Qiang Wang, Xia Sun, De-Xian Zhang, Xin Li “An Optimal Svm-Based Text Classification Algorithm” Fifth International Conference on Machine Learning and Cybernetics, Dalian,pp. 13-16 , 2006.

[19] Liu, H. and Motoda., “Feature Extraction, constraction and selection: A Data Mining Perpective.”, Boston, Massachusetts(MA): Kluwer Academic Publishers.

[20] Jingnian Chen a,b,, Houkuan Huang a, Shengfeng Tian a, Youli Qua Feature selection for text classification with Naïve Bayes” Expert Systems with Applications 36, pp. 5432–5435, 2009.

[21] S.W. Mohod, Dr.C.A. Dhote “Feature Selection Technique for Text Document Classification: An Alternative Approach”, International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169-2914 – 2917 Volume: 2 Issue: 9, September 2014,

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