# Comparative Study and Analysis of Supervised and Unsupervised Term Weighting Methods on Text Classification

Mahak Motwani
Asst. Prof., TIEIT
Bhopal

Aruna Tiwari
Asst.Prof., IIT
Indore

## ABSTRACT

Text Classification is one of the booming area in research with the availability of huge amount of electronic data in the form of news article, research articles, email message, blog, web pages etc. Text Representation is a vital step for text classification. In text representation, term weighting method assigns appropriate weights to the term to get better performance; the term weighting method which uses known information on membership of training document is supervised Term weighting method. Unsupervised term weighting method tf is compared with supervised Term weighting method tf.rf with Back Propagation Neural Network, results of experiment demonstrates that term weighing method (tf.rf) performs better than (tf) term frequency.

## General Terms

Text Classification, Text Preprocessing, BPNN,

## Keywords

Term Weighting Method, Relevance Factor, Term Frequency.

## 1. Introduction

With the immense growth of digital data searching data and deriving useful knowledge has become burdensome, Text Classification is one of the most essential requirements for effective navigating, summarizing and organizing data, Text classification is the task of assigning Predefined categories to free text documents.

Text classification method includes decision tree, naïve Bayesian[1],support vector machine[2],genetic programming [3], fuzzy k means[4],neural network[5],KNN[6]rule based classification etc. Neural network have the ability to recognize and response to pattern of information in the environment, they can be made to mimic the functioning of certain aspects Text document consist of sentences, special character, numbers, graph and images etc, in case of web data it has number of tags, words are extracted from these document which need to be tokenized and separated by space. Word extraction involves removal of special characters, numbers, graphs and images, in other word proper words are extracted from document.

## 2.2Removal of stop words

A set of functional words like is, a , an, the, of, for, at etc occur in almost all document and have low discriminating and representation value, these words are considered useless for classification. Removal of these stop words [13][14]15] results in proper set of words for further analysis

of human mind[7],various neural network approaches have been proposed and proved efficient for the task of text classification[8][9], Back Propagation Neural Network (BPNN) has been proved as one of the most successful algorithm in the field of neural network for text classification, BPNN with different approaches are used for text classification[10][11].

The text which basically is in unstructured form need a lot of preprocessing to be represented in the compact form of useful terms, Text representation is one of the most important step in text classification which represent the text document d= $(w_1,w_2,w_3,w_4,w_5….w_n)$here $w_1,w_2,w_3,w_4…$ represent the numeric value associated with k number of features. The process of allocating this value to words is called term weighting method. Term weighting methods assign appropriate weight to the terms, to improve the performance of text classification. Term weighting method for which information related to the category of document is available is supervised term weighting Method where as the traditional term weighting Binary, tf, tf.idf and its variant belongs to unsupervised term weighting method.

This paper Compares the result on BPNN for Unsupervised Term Weighting Method tf, with new supervised Term weighting method tf.rf proposed by Man Lan,Chew-Lim Tan And Hwee-Boon Low[12] ,Section 2 discusses various steps of preprocessing, section 3 discusses the approaches of term weighting method, section 4 describes the BPNN algorithm, section 5 presents experiments performed and reports result.

## 2. Text Preprocessing

Text document are unstructured and have a very huge size since each word is considered a dimension for text data, to get accurate, complete and consistent result systematic preprocessing need to be applied

## 2.1Word Extraction

## 2.3Stemming

Stemming is the action of reducing words to their root or base form. For English language, the Porters stemmer is a popular algorithm [15] [16], which is a suffix stripping sequence of systematic steps for stemming an English word, reducing the vocabulary of the training text by approximately one-third of its original size [16].For example, using the Porter’s stemmer, the English word “globalizations” would subsequently be stemmed as “globalizations → globalization → globalize → global”. In this paper porter’s stemming algorithm is used for suffix stripping

## 2.4 Dimension reduction

After eliminating stopwords and suffix stripping ,still text have huge number of words, the words which occur very less number of times and those which occur very frequently in document do not participate much in representing the document, such words are filtered which occur less than 2 times and which occur more than 35 time in our text document. Such words are filtered

# 3. Term Weighting Method

## 3.1Traditional Term Weighting Method

The traditional term weighting methods for text categorization are usually borrowed from information retrieval field and belong to the unsupervised term weighting methods, the simplest binary representation is used to represent 1 as presence & 0 as absence of the term in document, the term frequency is one of the most used method where each word is associated with the frequency of occurrence of that term in the document

The most popular term weighting approach is tf.idf[17] , which has been widely used in information retrieval and has consequently been adopted by researchers in text categorization. There are several variants of tf.idf, such as log(tf ).idf, tf.idf-prob , term relevance[18].Studies also showed there is no significant difference among them. Consequently, we adopt the most popular tf for our analysis

## 3.2 Supervised term weighting Method

In text classification of multiple classes, a term may have high term frequency (tf) and may belong to almost all the classes in this case the term actually do not possess a high discriminating power and so the inverse term document frequency factor and its variant has been used, although the tf.idf is not completely able to specify the discriminating power of a term since a term is assigned as positive category if it belongs to the document that belongs to the category and all other categories combined together as negative category it considers the occurrence of term in all the document positive as well as negative category. A supervised term weighting method where positive category is given more weight than negative category since the value of term in negative category is quite high and does not accurately represent the discriminating power of the term.

Supervised term weighting method used in this paper is a multiplication of term frequency (tf) and relevance factor (rf) where relevance factor is defined as

rf= log (2+ (a/max (1, c)) ;

Here

a: Total Number of document in the positive category that contain this term

c: Number of document in the negative category that contain this term

# 4. Back Propagation Neural Network

Back propagation neural network (BPNN) is the most popular in all of the neural network applications. It has the advantages of yielding high classification accuracy. The training of a network by back propagation involves three stages: the feed-forward of the input training pattern, the calculation and back-propagation of the associated error, and the adjustment of the weight and the biases.

Input pattern feed-forward. Calculate the neuron’s input and output. For the neuron j, the input Ij and output Oj are

$I_j=\sum W_{ij} * O_{j;}$

$O_{j=}f(I_j+ \theta_j)$

where $W_{ij}$ is the weight of the connection from the ith neuron in the previous layer to the neuron j, $f(I_j+ \theta_j)$is an activation function of the neurons, Oj is the output of neuron j, and $\theta_j$ is the bias input to the neuron. In this paper, a tanh(n)is used as sigmoid activation function defined with the equation:

tansig(n) = 2/(1+exp(-2*n))-1;

This function is a good trade off for neural networks. The error, E, is calculated in this paper, the mean absolute error function is used in the output layer The mean absolute error is used to evaluate the learning effects and the training will continue until the mean absolute error falls below some threshold or tolerance level.

$$\mathrm{E} = \frac{1}{2\pi}\sum_{\mathrm{n}}\sum_{\mathrm{l}}\sqrt{(\mathrm{T_{nl}} - \mathrm{Q_{nl}})^2}\,\mathrm{q}$$

Here n is the number of training patterns,l is the number of output nodes, and $O_{nl}$ and $T_{nl}$ are the output value and target value ,respectively. The mean absolute error is used to evaluate the learning effects and the training will continue until the mean absolute error falls below some threshold or tolerance level. The back propagation errors both in the output layer, $\delta_l$ and the hidden layer, $\delta_j$, are then calculated with the following formulas:

$$\delta_l = \lambda(T_l - O_l)\, f'(O_l)$$

$$\delta_j = \lambda\sum_i \delta_l W_{ij}\, f'(O_j)$$

Here $T_l$ is the desired output of the $l_{th}$ output neuron, $O_l$ is the actual output in the output layer, $O_j$ is the actual output value in the hidden layer, and k is the adjustable variable in the activation function. The back propagation error is used to update the weights and biases in both the output and hidden layers.

Weights and biases adjustment: The weights, wji, and biases, $\theta$i, are then adjusted using the following formulas:

$$W_{ji}(K+1) = W_{ji}(k) + \eta\delta_j O_i$$

$$\theta_i(k+1) = \theta_i(k) + \eta\delta_i$$

Here k is the number of the epoch and g is the learning rate.

The back propagation error is used to update the weights and biases in both the output and hidden layers [19].

# 5. Experiment Result

## 5.1Dataset

The mini newsgroup corpus is a popularly used dataset for text classification, it has a collection of 100 articles for each newsgroup. Our experiment is on three classes electronic, Politics and Space, 50 text file of each category are used, the dataset of these category is initially divided into test data and training data, 80% data as training data and 20% data as test data i.e. 40 text file from each category as training data and 10 files from each category as test data.

All text files undergo the method of pre processing i.e extract words ,removal of stop words, porter stemming algorithm is applied, and filter the data by considering words having term frequency greater than or equal to 2, and less than 35.

Total number of features extracted after preprocessing are 1993. The term weighting methods that are applied are term frequency (tf) and product of term frequency and relevance frequency (tf.rf)

The parameters used for BPNN are 1993 neurons in input layer and 20 neurons in hidden layer, training function used is gradient descent adaptive training function, tansig as activation function for hidden layer and linear function for output layer, learning rate used is 0.3, momentum of 0.6 .

## 5.2. Evaluation criteria

To evaluate performance of text classifier first calculates precision and recall. let the document relevant to a query is denoted as retrieved. The set of documents that are both relevant and retrieved is denoted as relevant ⋂ retrieved, precision is the percentage of retrieved documents that are in fact relevant to the query(i.e. ”correct” responses).it is defined as

Precision= Relevant ⋂ retrieved/retrieved

Recall: this is the percentage of document that are relevant to the documents that are relevant to the query and were in fact, retrieved. it is formally defined as

Recall= relevant ⋂ retrieved/relevant

And f measure is

F measure=2* (recall*precision)/ (recall +precision)

## 5.3 Result

It has been observed that with the increase in the number of epochs the training time also increases.

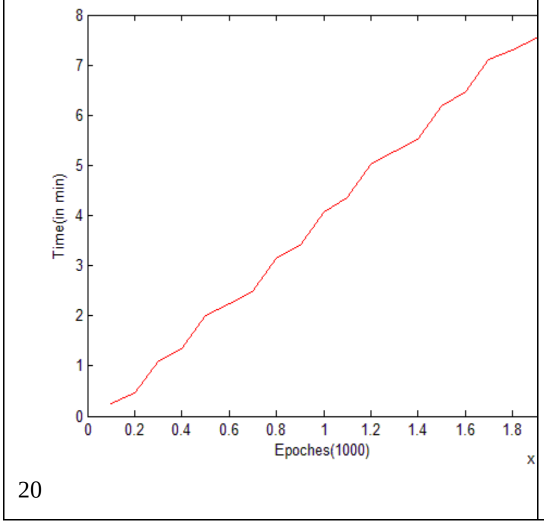
| Epoches(1000) | Training Time(in min) |
|---|---|
| 1 | 00.24 |
| 2 | 00.46 |
| 3 | 1.10 |
| 4 | 1.34 |
| 5 | 2.01 |
| 6 | 2.24 |
| 7 | 2.49 |
| 8 | 3.14 |
| 9 | 3.40 |
| 10 | 4.08 |
| 11 | 4.35 |
| 12 | 5.02 |
| 13 | 5.28 |
| 14 | 5.53 |
| 15 | 6.18 |
| 16 | 6.46 |
| 17 | 7.11 |
| 18 | 7.30 |
| 19 | 7.53 |
|  <br>20 | 7.58 |

**Figure1 shows the graph between time and number of epoches**

The result on the basis of F measure for Mini newsgroup data With Term frequency and Relevance frequency and Term frequency is as shown in the graph. The dotted line shows F measure for (tf.rf) and dashed line shows F measure for tf. The parameters used for BPNN are

| Neural network parameters | Value |
|---|---|
| Hidden layer | 20 |
| Training function | Traingda |
| Learning Rate | 0.03 |
| Momentum | 0.06 |
| Epoch | 1000-20000 |

**Table 2**

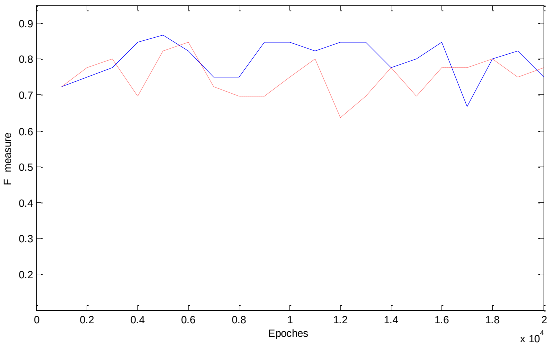


**Figure 2**

The Result shows that Term Weighting method (tf.rf) (blue line) performs better than (tf)(dashed line)in most of the cases.and performance of BPNN is best with F measure 0.8679 at 5000 epoches.

## 6. Conclusion

In the paper term frequency (tf) and product of term frequency and relevance frequency (tf.rf) has been applied as term weighting method on dataset mini newsgroup for three classes electronics, space, medicine on 50 text files each, results demonstrate that term weighting method (tf.rf) is better than (tf) with text classification method BPNN ,the results demonstrate that BPNN gives Fmeasure .8679 for tf.rf at 5000 epoch . The supervised term weighting method gives better result ,further some variants of relevance factor can also be applied ,time taken for preprocessing is high future work on quick preprocessing can be done. Training time of BPNN can also be reduced by modifying the algorithm.

## 7. References

[1] Goyal R. D. 2007," Knowledge based neural network for text classification". In proceedings of the IEEE international conference on Granular Computing, pp. 542 – 547.

[2] .T.Joachims,"Text Categorization with Support Vector Machine:Learning with many relevant features" Machine learning:ECML-98.10th European Conference on Machine Learning,p. 137-42,Proceeding 1998

[3]. B. Svingen, "Using Genetic programming for document classification",FLAIRS-98,Proceeding of eleventh Florida Artificial Intelligence Research,p 63-67,1998.

[4] M.Benkhalifa,A Bensaid and A Mouradi"Text Categorization using Fuzzy C means Algorithm," 18th international conference of the north American Fuzzy Information Proceeding Society-NAFIPS,p.561-5,1999

[5] J.Farkas"Generating Document Clusters using Thesauri and NeuralNetworks"Canadian Conference on Electrical and Computer Engineering, Vol 2,p. 710-713,1994

[6] M A Wajeed,T Vijayalaxmi,"Different Similarity Measure for Text Classification using KNN" International Conference on computer Communication Technology at NIT Allahabad Sept.2011

[7] P. Rothman. "Syntactic Pattern Recognition ." *AI Expert,* Vol. 7 . pages 41-51, 1992

[8] Zhihang Chen, chengwen Ni,Murphey Y. L,"Neural network approaches for text document categorization",Neural Network 2006 IJCNN,

[9] Wang,Z, He,Y, Jiang M"A Cmparison among three Neural Networks for text classification"Internation Conference on Signal Processing,2006 volume 3 p 16-20

[10] Wei Wang, Bo Yu" Text categorization based on combination of modified back propagation neural network and latent semantic analysis" Neural Computing & Application (2009) p: 875–881

[11] Combination of modified BPNN algorithms and an efficient feature selection method for text categorization

[12] M. Lan, C.L. Tan, H.B. Low, and S.Y. Sung, "A Comprehensive Comparative Study on Term Weighting Schemes for Text Categorization with Support Vector Machines," Special Interest Tracks and Posters of the www, pp. 1032-1033, 2005.

13] Kim S., Han K., Rim H., and Myaeng S. H. 2006. "Some effective techniques for naïve bayes text classification".IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 11, pp. 1457-1466.

[14] Zhang W., Yoshida T., and Tang X. 2007. "Text classification using multi-word features". In proceedings of the IEEE international conference on Systems, Man and Cybernetics, pp. 3519 – 3524.

[15] Hao Lili., and Hao Lizhu. 2008. "Automatic identification of stopwords in Chinese text classification". In proceedings of the IEEE international conference on Computer Science and Software Engineering, pp. 718 – 722.

[16] Porter M. F. 1980. "An algorithm for suffix stripping". Program, 14 (3), pp. 130-137.

[17] Gerard Salton , Christopher Buckley "Term-weighting approaches in automatic text retrieval "(1988) in Information Processing And Management,p 1214-9

[18] Harry Wu, Gerard Salton" The Estimation Of Term Relevance Weights Using Relevance Feedback" Journal of Documentation, Vol. 37 Iss: 4, pp.194 - 214

[19] Combination of modified BPNN algorithms and an efficient feature selection method for text categorization Cheng Hua Li *, Soon Cheol Park, Information Processing and Management 45 (2009) 329–340