Neural Network Approach for Text Classification using Relevance Factor as Term Weighing Method

Anuradha Patra
Barkatulallah University Institute of Technology
Barkatulaah University Bhopal,
MP,India

Divakar Singh
Head CSE Deptt
Barkatullah University
Institute Of Technology Barkatulaah University
Bhopal MP,India

ABSTRACT

With the rapid growth of online information there is growing need for tools that help in finding filtering and managing the high dimensional data .text classification is a supervised learning task whose goal is to classify document into the predefined categories. Phases involved in text classification are collecting data set, preprocessing, stemming, and implementing the classifier and performance measure. There are several learning method for Text classification such as Naïve bayes, k-nearest neighbor decision tree, SVM, BPNN etc. algorithm is applied to multilayer feed forward networks consisting of processing element with continuous differentiable activation function. The network associated with back propagation learning algorithm called BPNN. This paper demonstrates the result of text classification using BPNN and relevance factor (rf) as term weighing method.

General Terms

Neural network, Data mining, Text classification.

Keywords

Relevance factor, performance measure, BPNN

1. INTRODUCTION

Text classification is the process of deriving high quality information from text due to increased availability of document in digital form and rapid growth of online information. text classification has become one of the key techniques for handling and organizing text data[1].there are many method for constructing classification methods such as naïve Bayesian [2], support vector machine [3], and k-nearest neighbor[4] classification, neural network[5] etc. A neural network typically a collection of neuron like processing unit with weighted connection between the units. . Neural network technique have been promising tool for text classification various architecture of neural network have been used for the application of text classification e.g. Error Back Propagation algorithm, ADALINE and MADALINE network etc. Phases involved in text classification are collecting the dataset, preprocessing, dimensionality reduction, implementing the classifier.

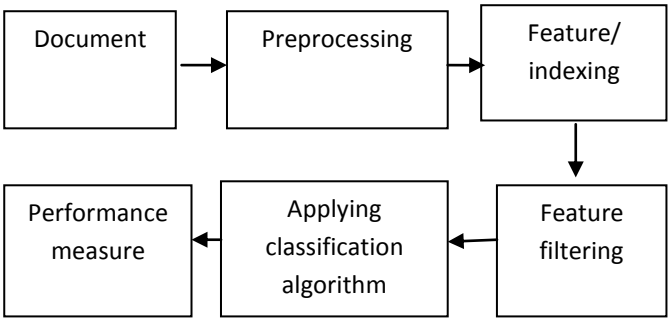

Fig 1: Text Classification Framework

2. PREPROCESSING

Data mining is the process of extracting hidden pattern in a large dataset .real world data is often incomplete inconsistent and lacking in certain behavior and is likely to contain many error[6]s. Data goes through a series of steps:

1) Data cleaning: data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in data.

2) Data integration: data with different representation are put together and conflicts within the data are resolved.

3) Data transformation: data is normalized aggregated and generalized

4) Data reduction: this step aims to present a reduce representation of data

5) Data discretization: involves the reduction of a number of values of a continuous attribute by dividing the range of attribute intervals

2.1 Stop word removal:

It is well recognized among the information retrieval experts that a set of functional English words (eg. “the”,”a”,”and”,”that”) is useless as indexing terms. These words have very low Discrimination value, since they occur in every English document [7]. Hence they do not help in distinguishing between documents with contents that are about different topics. The process of removing the set of non content- bearing functional words from the set of words produced by word extraction is known as stop words removal. In order to remove the stop words, this involves first creating a list of stop words to be removed, which is also called the

stop word list. After this, the set of words produced by word extraction is then scanned so that every word appearing in the stop list is removed.

2.2 Porter stemming algorithm:

Porter stemming algorithm is a process for removing the commoner morphological ending words in English [8]. Rules in porter stemming algorithm are separated into five distinct steps:

1) Gets rid of plurals and -ed or -ing. eg-> caress ponies -> ponities -> ti caress -> caress cats -> cat
2) Turns terminal y to i when there is another vowel in the stem. eg happy->happi
3) Maps double suffices to single ones. so -ization (= -ize plus -ation) maps to -ize etc.
4) Deals with -ic-, -full, -ness etc. similar strategy to step3.
5) Takes off -ant, -ence etc.

3. Term weighing method

According to the importance of the terms in the documents; measurements must be made for each term in the vocabulary for producing the set of initial features from preprocessed term [9]. This involves assigning each term a weight indicating the relative importance of the term in a document. This process of assigning a weight to each term is commonly known as term weighting method.

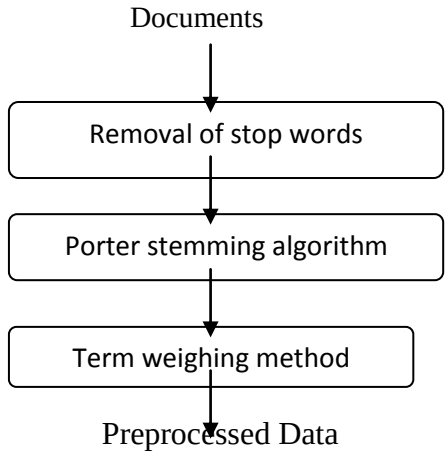

Fig 2: preprocessing the data

3.1 Term frequency factors:

Table1 represent four commonly used term frequency factors, including a binary weight term frequency alone (tf), logarithm of term frequency log (tf) [10], and inverse term frequency (ITF). The simplest term frequency factor Binary represent the occurrence of term as 1, and absence as 0. tf means number of times a term occurs in a document. Moreover, different variants of term frequency have been existing, log (1+tf) used to scale the effect of unfavorable high term frequency in one document, inverse term frequency is inspired by inverse document frequency which emphasize discriminating factor of term [11].

TABLE 1
Different Term Frequency Factors.

Term frequency factor	Represented by	Description
1.0	Binary	1 represent term present in document
Term frequency	tf	Number of times a term occurs in frequency
Log(1+t.f)	log tf	Logarithm of the term frequency
1-1/r+tf	ITF	Inverse term frequency generally r=1

4. Supervised term weighing method Relevance factor

In this approach a chosen category is tagged in a positive category and other categories in the same document are pooled together as a negative category. For e.g. a document contain four term t1, t2, t3, t4, t5, t6 given one chosen positive category on one data collection [10]. In this figure, each column represents the document, and horizontal line divide column into two categories, first is positive i.e. above and second is negative i.e. below

use a,b,c,d to donate the different document, as listed below

a is the number of document in the positive category that have that term.

b is the number of document in the positive category that do not have this term.

c is the number of documents in the negative category that have this term.

d is the number of document in the negative category that do not have this term.

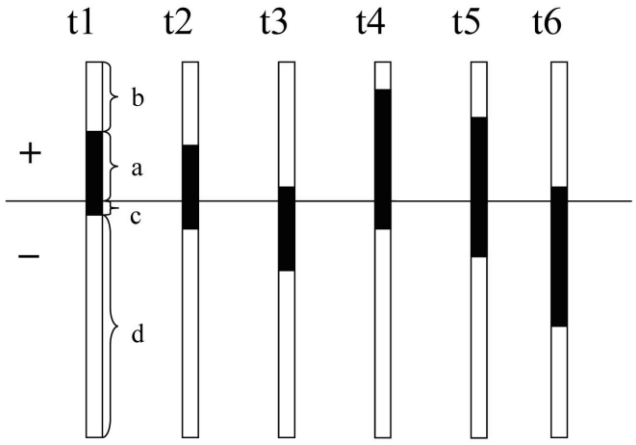

Fig 3: Different distribution of document contain six terms in the whole collection

Relevance factor is:

$$rf = \log(2 + \frac{a}{c})$$

and when combined tf with rf by multiplication operation term can be:

$$tf.rf = tf * \log(2 + \frac{a}{c})$$

Let us consider two extreme case if a=0 then rf =1 then the weight of term is equal to term frequency.

In case 2: if c=0, to avoid zero divisor, set minimal denominator as 1 thus rf formula can be:

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

Then the final term weight is replaced by

t.f.r.f=tf*log (2+a/max (1, c))

There are different supervised term weighing as shown in table 2

TABLE 2

Different Supervised Term Weighing Method

Approach	Represented by	Description
Supervised term weighing method	tf.rf	t.f.r.f=tf*log(2+a/max(1,c))
	$tf.X^2$	Based on information theory
	tf.ig	t.f information gain
	tf.logOR	Tf.log(odd ratio)

5. Text classification learning algorithm

Typical machine learning algorithms are NB (naïve Bayes), KNN (k-nearest neighbor), SVM (support vector machine) and BPNN (back propagation algorithm). NB classify document based on prior probabilities of category and probabilities that attribute that values belong to categories [12]. its performance is feasible in text categorization. Another popular approach is SVM it is an algorithm that work as follows it uses a non linear mapping to translate and transform the original data into higher dimension[13]. Within the new dimension it searches for linear optimal separating hyperplane.with an appropriate non linear mapping to sufficiently separate high dimension, data from two classes can always be separated by a hyper plane. Druckeral adopted SVM for implementing a spam filtering system and compared it with NB in implementing the system in 1999[14] and they conclude that SVM was the superior approach than NB. In KNN when the discrete set of cases take the instance x to be classified find K nearest neighbors of x in the training data , determine the class of the majority of instances among the nearest neighbors ,return the class as the classification of x[15].KNN is evaluated as easy and competitive algorithm as SVM. Its disadvantage is that KNN cost a lot for classifying the object and it is inefficient in the case of high dimensional and large scale data set .One more traditional approach is neural network. It is a popular classification method, set of connected input/output units in which each connection has a weight associated with it[16] . During the learning phase, the network learns by adjusting the weights so as to be able to predict the correct class label of the input tuples. There are many ANN models available namely BPNN (back propagation algorithm [17]), CPNN (counter propagation networks)[1] .in this paper we use BPNN algorithm for classification.

The structure of the three layered back propagation neural network [18] is shown in fig

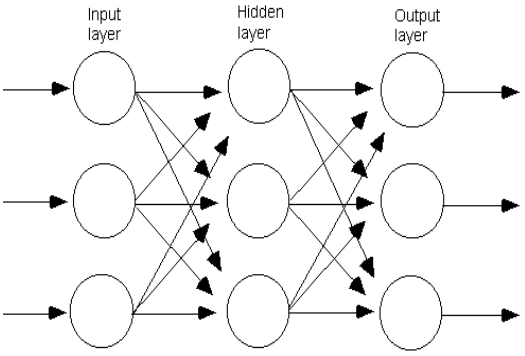

Fig 3: Three layered back propagation neural network

5.1 BPNN Algorithm:

A back propagation neural network [19]is a multilayer, feed-forward neural network consisting of input layer, a hidden layer and an output layer. The neurons present in hidden layer and output layer have biases, which are connection from units whose activation function is always 1.the bias term is also act as weights. The inputs are sent to the BPNN and the output obtained from the net could be 0 0r 1or bipolar (-1, +1) [20] .terminologies used in BPNN is as follows:

X=input training vector(x_1,…….x_i……x_n)

t=target output vector (t_1……..t_k……t_m)

α =learning rate parameters

Xi=input unit i

V_{0j}=bias on j_{th} hidden unit

W_{0k}=bias on k_{th} output unit

Z_j=hidden unit j. the net input to z_j is

$$Z_{inj} = V_{oj} + \sum_{i=1}^{n} x_i v_{ij}$$

And the output is

$Zj=f(Z_{inj})$

y_k=output unit k. the net output to y_k is

$$y_{ink}=w_{0k}+\sum_{j=1}^{p} z_j w_{jk}$$

And the output is

$y_k=f(y_{ink})$

Compute error correction

$$\delta_k = (t_k - y_k) f'(y_{ink})$$

Update the change in weight and bias

$$\Delta w_{jk} = \alpha \delta_k z_j ; \Delta w_{0k} = \alpha \delta_k$$

Each hidden unit (z_j , j=1 to p)

$$\delta_{inj} = \sum_{k=1}^{m} \delta_k w_{jk}$$

Calculate error term

$$\delta_j = \delta_{inj} f'(z_{inj})$$

Compute change in weight & bias based

$$\Delta v_{ij} = \alpha \delta_j x_i ; \Delta v_{0j} = \alpha \delta_j$$

Update weight & bias on output unit

$$w_{jk}(new) = w_{jk}(old) + \Delta w_{jk}$$
$$w_{0k}(new) = w_{0k}(old) + \Delta w_{0k}$$

Update weight and bias on hidden unit

$$v_{ij}(new) = v_{ij}(old) + \Delta v_{ij}$$
$$v_{0j}(new) = v_{0j}(old) + \Delta v_{0j}$$

The stopping condition may be certain number of epochs reached.

6. Proposed work:

In this paper take the data set from newsgroup 20 .the text document must pass through set of steps: elimination of stop words, numerals, porter stemming. After applying preprocessing apply term weighing method. use tf.rf as term weighing method .finally, the classifier is constructed by learning the characteristics of every category from a training set of document.80% of data is for training set and 20% of data is for testing purpose. Once a classifier has been built, its effectiveness may be tested by applying it to test set and checking the degree of accuracy of the classifier on text document

6.1. Data set

In experiment use newsgroup20.this data is a unstructured data in the form of text represent the number of document for each category; In this paper 20% data set is used testing and 80% for training:

TABLE 3

Category	Number of training document	Number of testing document
Computer science	40	10
Sports	40	10
Medicine	40	10

6.2. Preprocessing the data:

Aim of work is to evaluate performance of using tf.rf term weighing method. After the preprocessing in each document, represent document term frequency tf i.e. the number of how many times the word occur in document. Calculation of term frequency weighing method is calculated by=

$$tf.rf = tf * \log(2 + \tfrac{a}{c})$$

In this paper, compared the performance by evaluating the term weighing method without using relevance factor.

6.3. 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 $\bigcap$ retrieved, precision is the percentage of retrieved documents that are in fact relevant to the query (i.e. "correct" responses).it is defined as

$$precision = \frac{|\{\text{Relevant}\} \cap \{\text{Retrieved}\}|}{|\{Retrieved\}|}$$

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

$$\text{Recall} = \frac{|\{\text{Relevant}\} \cap \{\text{Retrieved}\}|}{|\{\text{Relevant}\}|}$$

And F measure [13] is

$$\text{F measure} = \frac{\text{Recall x } precision}{(\text{Recall} + precision)/2}$$

6.4 Characterization of neural network:

TABLE 4

Neural network parameters	Value
Hidden layer	20
Training function	Traingda
Learning Rate	0.3
Momentum	0.6
Epoch	20000

Comparison between term frequency and term frequency with relevance factor:

TABLE 5

S.No	Data Set	Term Frequency factor	F measure
1	Computer science	Tf	0.72
	Sports		
	Medicine		
2	Computer science	tf.rf	0.92
	Sports		
	medicine		

6.5Cassification results:

This paper presented work on the effective neural network algorithm BPNN for the task of text classification, 150 text document belonging to 3 categories i.e. computer science, sports, and medicine of Newsgroup20 have been preprocessed by stop words removal, porter stemming algorithm followed by two supervised term weighting method (tf) and (tf.rf), experiment result demonstrates that (tf.rf) performs better (tf) the performance is evaluated by F-measure term. Without using relevance factor value of f measure is 0.72 while using tf .rf gives f measure=0.92 .Preprocessing time is very high, future work on quick preprocessing methods can be applied

7. REFERENCES

[1] Fouzi Harrag,Eyas,EI-Qawasmah, Abdul Malik S.AI Salman(2010) "Comparing dimension reduction techniques for Arabic text classification using BPNN algorithm".international journal conference on integrated intelligence computing.2010

[2] Martinez-Arroya, M(2006). "learning on optimal Naïve Bayes classifier" ,p(1236-1239)

[3] Sapan Kevych,N(24-38)."Time Series prediction using support vector mahines A survey" ,p(24-38)

[4] Xianfei Zhang, Zhengzhou , Zhengzhor,Bichengli,Xianzhu Sun. "A K-nearest neighbor text classification algorithm based on fuzzy integral",p(2228-2231)

[5] Zhihang chen, Chengwe Ni and Yil. "Neural network approaches for text document categorization" p(1050-1060)

[6] Jasdeep Singh Malik,Prachi Goyal, Akhilesh K Sharma." A Comprehensive approach towards data preprocessing techniques &association rules"

[7] S.Ramasundram, S.P.Victor, "text categorization by BackPropagation" ,Proc.Int'l journal of computer application pp.(0975-8887).2010

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

[9] V.Srividhya, Anitha "Evaluating Preprocessing Techniques in Text Categorization" Proc.Int'l journal of computer science and application Issue .2010

[10] Man Lan,Chew Lim Tan,Jain Su,Yue Lu."Supervised and Traditional Term Weighing Methods For Automatic Text Categorization"Proc.IEEE Transactions on Pattern Analysis and Machine Intelligence pp. (721-735) 2009

[11] E.Leopold and J.Kindermann,"Text Categorization with Support Vector Machines.How to Represent Texts in Input Space .Machine Learning" vol.46,nos.1-3,pp.423-444,2002.

[12] Tacho Jo "Neural Text Categorizer for exclusive Text categorization" proc journal of information processing system vol 4, no.2,june 2008

[13] Data Pre-processing & Mining Algorithm, Knowledge & Data Mining & Preprocessing, 3rd edition Han & Kamber.

[14] H. Drucker, D. Wu, and V. N. Vapnik, "Support Vector Machines for Spam Categorization", IEEE Transaction on Neural Networks, Vol.10, No.5, pp.1048-1054, 1999.

[15] M.E. Ruiz, and P. Srinivasan, "Hierarchical Text Categorization Using Neural Networks", Information Retrieval, Vol.5, No.1, pp.87-118, 2002.

[16] Miguel E.Ruiz, Padmini Srinivasan "Automatic Text Categorization using neural networks "Advance in classification research vol III.

[17] Nita Mathuriya, Ashish Bansal "Comparision of k-means and back propagation algorithm' proc.Int'l journal of computer technology and electronics engineering vol 2, Issue 2

[18] YashPal Singh, Alok Singh Chouhan "Neural Networks in Data Mining" Journal of theoretical and applied information technology.

[19] S. N.Sivanandam, S. N. Deepa "Principles of Soft Computing "

[20] Nitin Mathuriya, Ashish Bansal "Applicability of Backpropagation neural network for recruitment data mining vol 1. Issue 3,2012.