International Journal of Engineering & Technology, 7 (1.1) (2018) 603-606

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Text classification using artificial neural networks

P. Lakshmi Prasanna[1]*, Dr. D.Rajeswara Rao[2]

[1]Research Scholar,[2]Professor·Department of Computer Science and Engineering,*
Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India

Abstract

Text Categorization is the process of classifying the text or documents into its corresponding categories which are defined previously. As the text or data is increasing enormously now-a-days it's not possible to classify all the text documents manually hence its necessary to use some techniques or methods to classify the text automatically. In this paper we are using the ANN technique to classify the text, the purpose of choosing it, its advantages and its process is described in the further sections.

1. Introduction

With the increase in amount of data stored text categorization became one of the important tasks now-a-days for web searching, datamining, spam filtering, sentiment analysis, decision making etc[2][18]. Document or text classification is the main task in text mining, natural language processing(NLP) to organize the information in a supervised manner into some categories[16][6]. Text categorization is a pattern classification technique for text mining and necessary for effective management of textual data[8]. It is also required in the medical text classification, health professionals spend large amount of time scanning the notes to identify the key problems and understand the status of the patient[4]. The need for this classification is successive computerization of file processing reliant on the file type[3]. There are many automatic text categorization techniques like Naïve bayes, SVM, KNN, decision trees, ANN etc. Text categorization of Ann involves the following stages, after the required data is collected each document is preprocessed and in turn the preprocessing stage involves document conversion, stemming, indexing.

2. Preprocessing

Once preprocessing is done then the neural network will be trained to categorize the documents into the categories which were predefined. When we go deep into the preprocessing stage in the first step that is document conversion the document is converted into plain text and the stop words like prepositions, participles are removed from the documents for example the, a, an etc because they are considered as unimportant. After the document conversion stemming takes place in the stemming affixes (which is a letter or series of letters added to a root word that can change its meaning) of the words are removed and the root word is obtained. Once the stemming is done document is gone through the indexing, creation of internal representation will be done for the document, the first step of indexing involves construction of super vector which holds all terms that appear in all documents of corpus after that in the second step of indexing for dimensionality reduction is done by extracting subset of the super vector based on some criteria this is going to work more effectively [1]. After that in the third step term weighting is done for every term and for every document in the corpus.

3. Neural network

Neural network is an assemblage of neurons with weightages which connects them, they process records one at a time and learn by comparing their classification with the actual classification. Neural network has the properties like robustness, self learning and adaptiveness[17]. Neural network is one of the most advanced classifiers in the testing category[7]. Neural network can be defined in three parts or layers they are input layer, hidden/ intermediate layer and output layer. The duty of the input layer is to receive the input signals from the outer system. Coming to the hidden layer it is comprised of neurons. The learning of the neural network is fully supervised hence for the input provides to the neural network has an answer or output[14]. The neural network takes input values and weights from the input layer as input and then it goes to the hidden layer in which a function sums the weights and maps the results to the corresponding output layer units. We can have 'n' number of hidden layers in between the input and output layers. Depending on the number of hidden layers the network will be named as single layer neural network or multi layered neural network(for more than one hidden layers).

4. Neural network training

Training a neural network involves arranging all the weightages by replaying two major steps, forward and backward propagations. In the forward we give a collection of weights to the input and then figure out the output. For the initial forward propagation the inputs are chosen randomly. On the otherhand in the backward we ration the margin of error of the figured out output and then modify the weights accordingly to reduce the error. Neural networks replay both forward and backward propagations til the weights are balanced to predict the output accurately. A function is used in the hidden layers of the neural network that sums the inputs with the weights and maps the accurate output. Some of the functions are linear, sigmoid, hyperbolic tanget etc.

When we go in more detail about the training of the neural networks in the first stage that is the feedforward propagation each input (xi) receives an input signal after receiving the input it sends that signal to each of the hidden units(z1,z2,z3,…..) and then the concealed layer calculates its
activation signal and sends that signal to the output unit and then each output unit uses the activation signal to compute the output or response for the given input signal or pattern. Once the output value is generated then each output unit checks its activation(yk) with the target value(tk) that is the value need to be obtained after the input pattern is processed, where k=1,2,

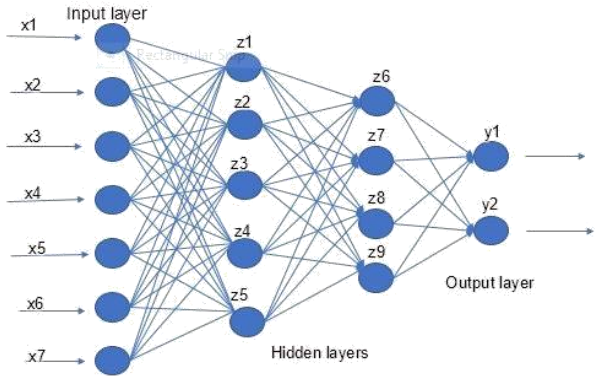


Fig. 2: Neural Network Architecture

There are five types of neural networks they are feedforward neural network, radial basis function neural network, kohonen self-organizing neural network, recurrent neural network, modular neural networks.

4.1 .Feed forward neural network:

In this network data can move in one direction only, from the input layer to the output layer passing over the hidden layers if they are present. As the presence of the hidden layers is not compulsory. This network doesn't have any loops or cycles in it, this is the simplest network of a ll the neural networks.

4. 2. Radial basis function (RBF):

In this network every neuron saves an example from training set as a template. Linearity which is present in the network for its functioning prevents it from local minima. When interpolating in the multidimensional space this network is the best option.

4.3.Kohonen's neural network:

It can describe hidden elements from the data which is un-labeled with the usage of functions. This network can arrange by itself for the presentation of the low dimensional views of high-dimensional data. It uses competitive learning on the input that is denied by fault correction learning by another network.

4.4.Recurrent neural network:

Unlike the feedback NN it allows bi-directional flow of data. The connected units form a directed cycle. It has the capacity to use its internal memory for processing arbitrary inputs. It is popular in the handwriting and speech recognitions.

4.5.Modular neural networks:

In this network the independent NN's don't interact with each other. This network consists of a sequence of autonomous neural networks that are balanced by intermediary. Each of them work with different inputs executing sub tasks that finally produces the task that the network as a whole wants to obtain. The intermediary welcomes the inputs of each of these autonomous networks processes them and generates the final output values.

5. Existing Work

Many researchers have been working on the text categorization as the necessity is increasing day-by-day. Many researches have been conducted on the text categorization after analyzing all of them it is found that ANN gives the best results. In one of the research it is found that neural network along with SVD(Singular value decomposition) gives the best result in this SVD is used to decrease data in terms of both extent and magnitude. In other research analysis was done on the two types of neural networks that is recurrent NN(neural network), recursive NN and to overcome the problems that still arise even after the usage of those techniques the author proposed a new technique that is recurrent convolution neural network which is the combination of recurrent neural network and the convolution neural network. Coming to the other research CNN was found to give the best results in which 'word2vec' algorithm is used. In the three other researches it was found that back propagation NN is efficient it is a ANN technique. BPNN along with SVM(support vector machine) was found to give best result. Even though we have the problem of huge dimensionality and sparse distribution that was addressed with the help NTC(Neural text categorizer) it is represented in the form of string vector which is an ordered finite set of words. Very few people have studied neural networks compared to the traditional algorithms and also it has many advantages and can handle data with high dimensionality and more advantages are listed in the following section.

6. Proposed System

After analyzing all the text classification techniques their advantages and disadvantages we came up to classify the text using ANN because of its advantages and researches done on it which gave better results than other classification techniques as described in the previous section. One of the main reason for choosing the ANN is it can provide better solution for the problems which cannot be solved linearly or by using linear statistical classification techniques[6]. Some other features of ANN are it can even learn in the presence of noise[5], it teaches to accomplish the task instead of programming computational system to perform required tasks. ANN has the capacity to learn and build non-linear and complex relationships. After learning from the inputs and their connections it can deduce unseen connections on unseen data also. Ann doesn't force restrictions on the input variables like other prediction techniques.

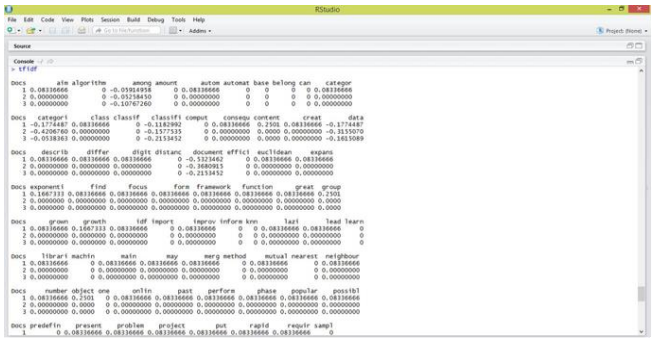
7. Literature Survey

S.NO	Author	Algorithm	Dataset Taken	Results
1	Fouzi Harrag, Eyas Al-Qawasmah	Artificial neural network	Hadiths	SVD is 52.4s times faster than MLP NN, MLPNN gets best performance of 52%
2	Siwei Lai, Liheng Xu, Kang Liu, Jun Zhao	RCNN was proposed	20Newsgroups, Fudan Set, ACL Anthology Network,and Sentiment Treebank	Compared to traditional methods neural Networks performs well and comparing. The CNN, Recursive NN convolution based algorithms performs well
3	Jaromir Veber	Information gain and Neural network	Reuters-22173 and OHSUMED collection	Train set accuracy was 95% and it is expected 90% accuracy on the real data
4	Mark HUGHES, Irene LI , Spyros KOTOULAS and Toyotaro SUZUMURA	Word2vec	PubMed collection	Outperforms several natural language Processing algorithms by 15%
5	SaravananK and S. Sasithra	Backpropagation neural network	Multi-dimensional datasets From medical background	84% correct prediction for control subjects and 89% for the WAD
6	Gertrud jeschke, mounia lalmas	Dempster shafer theory of Evidence and neural networks	OSHUMED(subset of MEDLINE collection) AND MeSH thesaurus	It is observed that the usage of demster shafer theory on the top Of neural network performs well
7	Taeho Jo	Neural text categorizer	Newspage.com, 20NewsGroups, and Reuter 21578	NTC is more practical than other classification techniques in the real world

Rie Johnson, Tong Zhang [8] came up with deep pyramid convolution neural networks for text categorization used datasets namely AG, Sogou, Dbpedia, Yelp.p, Yelp.f, Yahoo, Ama.f, Ama.p and found that DPCNN outperforms the previous best models on six benchmark datasets. Miguel E. Rniz [9] Padmini Srinivasan came up with automatic text categorization using neural networks, taken 2,344 MEDLINE documents after experimenting it was found that BPNN performs well than counter propagation network.[10]Anthony Rios, Ramakanth Kavuluru done experiments on biomedical text using convolution neural network and it was found that CNN's with convolution filters(CF's) are best for biomedical text classification rather than other techniques and achieved a macro F-score improvement of 3%.

8. Results

This is the TFIDF matrix construction of the text documents it is implemented in R -tool. This is obtained after removing stop words from the documents and after the stemming process that is after the preprocessing as discussed earlier. In this matrix the rows include the values of that document and the columns include the words.



9. Deduction and Future Scope-

In this research we have constructed a TFIDF matrix which will be given to the artificial neural network from which the neural networks learns and then categorizes the documents into prede fined categories that will be implemented and it will be described in the further papers. As it is given in the existing system that very few people have studied about the neural networks and its working with respect to the text categorization we wanted to implement and show how the text categorized easily using the ANN techniques as it is easy and has many advantages compared to the traditional methods like Naïve bayes, KNN, SVM etc

References

[1] Fouzi Harrag1 , Eyas Al-Qawasmah2 "Improving Arabic text categorization using Neural network with SVD ,Journal of digital information management, 2010.

[2] Siwei Lai, Liheng Xu, Kang Liu, Jun Zhao ,Recurrent convolution neural networks for text classification" in: proceedings of twenty ninth AAAI conference on artificial intelligence,2015.

[3] Jaromir Veber ,Classifying plain source files using neural network" in: Journal of systems integration, 2010.

[4] Mark HUGHESa , Irene L, Spyros KOTOULASa and Toyotaro SUZUMURAb, Medical Text Classification using Convolutional Neural Networks,2015.

[5] Saravanan K and S. Sasisthra Review On Classification Based On Artificial Neural Networks", in: International Journal of Ambient Systems and Applications (IJASA) Vol.2, No.4, December 2014.

[6] Chaitanya Naik, Vallari Kothari, Zankhana Rana, "Document Classification using Neural Networks Based on Words", In: International Journal of Advanced Research in Computer Science,2015.

[7] Raghvan Gachli Online news text classification using Neural network and SVM,2015.

[8] Taeho Jo NTC (Neural Text Categorizer): Neural Network for Text Categorization,2015.

[9] Miguel E. Ruiz And Padmini Srinivasan Hierarchical Text Categorization Using Neural Networks , 2001.

[10] Rie Johnson, Tong Zhang "Deep Pyramid Convolutional Neural Networks for Text Categorization ,2012.

[11] Miguel . E ruiz and Padmini srinivasan Automatic text categorization using neural networks 2012.

[12] Anthony Rios, Ramakanth Kavuluru Convolutional Neural Networks for Biomedical Text Classification: Application in Indexing Biomedical Articles.

[13] Gertrud Jeschke and Mounia Lalmas Hierarchical Text Categorization based on Neural Networks and Dempster-Shafer Theory of Evidence,2013.

[14] Rahul Bansod, R. S. Mangrulkar, V. G. Bhujade Spam classification using artificial neural networks with weight measures, in: International journal of advanced computer technology (IJACT)

2012.
[15] Revathi N, Anjana Peter, S.J.K.Jagadeesh Kumar Web Text Classification Using Genetic Algorithm and a Dynamic Neural Network Model,International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) , 2013.
[16] Upendra Singh, Saqib Hasan, Survey Paper on Document Classification and Classifiers, International Journal of Computer Science Trends and Technology (IJCST),2015.
[17] Stefan Wermter ,Neural Network Agents for Learning Semantic Text Classification.
[18] Gurmeet Kaur, Karan Bajaj News Classification using Neural Networks, 2016.
[19] Nerijus Remeikis, Ignas Skucas, Vida Melninkait E Text categorization using neural networks initialized with decisiontrees,2004