Asian Journal of Computer Science and Technology
ISSN: 2249-0701 (P) Vol.8 No.2, 2019, pp.55-58
© The Research Publication, www.trp.org.in
DOI: https://doi.org/10.51983/ajcst-2019.8.2.2140

Document Classification Using Artificial Neural Network

Kshitij Tripathi[1], Rajendra G. Vyas[2] and Anil K. Gupta[3]
[1]Department of Computer Applications, [2]Department of Mathematics
[1&2]The Maharaja Sayajirao University of Baroda, Vadodara, Gujarat, India
[3]Department of Computer Applications, Barkatullah University, Bhopal, Madhya Pradesh, India
E-Mail: kshitij.tripathi-compapp@msubaroda.ac.in, vyas.rajendra-math@msubaroda.ac.in, akgupta_bu@yahoo.co.in

(Received 12 March 2019; Revised 23 March 2019; Accepted 21 April 2019; Available online 28 April 2019)

Abstract **- The Document classification system is the field of data mining in which the format of data is based on bag of words (BoW) or document vector model and the task is to build a machine which after successfully learn the characteristic of given data set, predicts the category of the document to which the word vector belongs. In this approach document is represented by BoW where every single word is used as feature which occurs in a document. The proposed article presents artificial neural network approach which is hybrid of n-fold cross validation and training-validation-test approach for classification of data.**

Keywords: **N-Fold Cross Validation, Validation, Classification, Neural Network, Bag of Words**

I. INTRODUCTION

The document classification is the technique by which digital documents are categorized into different classes according to their content. The classes are already defined on the basis of their content and technique attempts to build classifier through the training data. After training is done the classifier tries to classify a previously unseen document into different categories. There are various approaches present in document classification like rule based methods, probabilistic methods and machine learning methods are few of them. This article presents artificial neural network (ANN) for identifying the document through n-fold TVT (Training-Validation-Test) approach [17]. The text document may be ASCII, HTML or may be XML. Text classification is a supervised learning technique in which document always belong to some class and after building a model through training it has to classify the document for categorization on test data-set.

II. ARTIFICIAL NEURAL NETWORK

The Artificial Neural Networks (ANNs) are input-output processing systems [14] [18] [19] inspired from human brain that is they consist of neurons which is fundamental unit of the human brain.

In the given fig.1ANN [18] [19] contains multiple layers and these layers contain multiple neurons. Each neuron is associated with a transfer function. Layers are connected with each other's through the edges. Each edge is associated with a numeric value called as weight. The ANNs are configured properly before applying the dataset to it. The working of ANN is based on back-propagation algorithm [6].

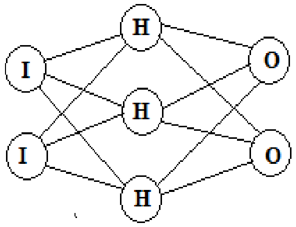


Fig. 1 The architecture of artificial neural network

The working of artificial neural networks for classification is as follows:

Step1: The data is prepared for applying it to ANN i.e. it must be numeric.
Step2: Before processing or applying dataset the ANN is initialized and configured
Step3: The ANN is trained by applying input-output dataset by employing back-propagation algorithm.
Step4: The testing is done on the network obtained in Step3.

III. EXISTING TECHNIQUES

Text classification can be performed by employing so many techniques like term graph , support vector method, nearest neighbour method and Bayesian classification method to have a few of them. The present article presents a technique based on ANN which is a refinement of training-validation-test approach.

IV. LITERATURE REVIEW

There are various literatures found in the field of text classification like [13] proposed neural network approach by representing text with its ASCII format. [4] Represents problem as information selection in which the process is based on spreading activation methods. Merkl and Rauber [11] article argue in favour of establishing a hierarchical organization of the document space based on unsupervised neural network using hierarchical feature map for text archive organization, a type of self organizing map. Kakade *et al.,* [2] create a term document matrix that helps in conversion of the text document into a quantitative format. This paper presents a new idea which allows the use of term factor (tf) and tf-inverse document factor (tf-idf) vectors to represent a text document. Hsieh *et al.,* [8] proposed a novel usage of word and document embedding for emotion classification. Pavan Kumar *et al.,* [19] have a very nice presentation on sentiment analysis and classification.

V. PROPOSED METHODOLOGY

The proposed methodology is based on n-fold training-validation-test approach which is combination of n-fold cross-validation with validation [21]. In n-fold cross-validation the dataset is initially divided into 'n' folds. Than during training each 'n' folds one by one acts as test set and remaining sets are treated as training set. Finally the mean error or accuracy (as the case may be) obtained from each of 'n' test sets are the final result of classification. In the TVT (training-validation-test) approach the data is first divided in to three parts (generally ratio is 70:15:15) and each time during experiment the training is performed and before testing, the built model is first validated again and again with validation set till best validation performance is achieved.

After obtaining best validation performance they obtained ANN is employed for testing on test set. The proposed approach is hybrid approach of n-fold cross-validation and training-validation-test approach. That is after dataset is divided in to 'n' folds, each time during testing 'nth' fold is reserved for testing and one of the remaining folds one by one are treated as validation fold and remaining n-2 folds are used for training as shown in fig 3. As each of the remaining n-1 folds one by one are used as validation set till the best friend of test set is discovered it is called an exhaustive approach.

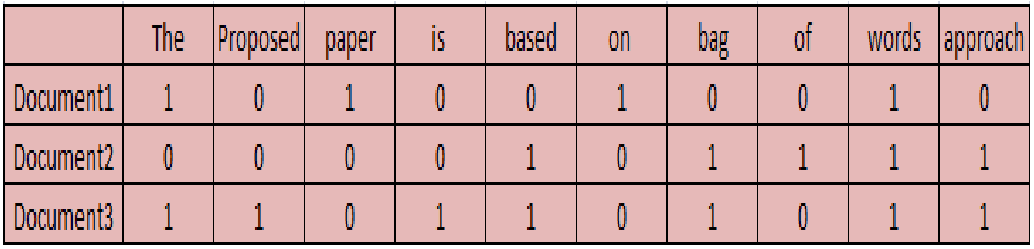
Further in the proposed approach we use exhaustive weight initialization as it is observed that weight initialization in ANN affects the accuracy of results at the end so we select the initial weight configuration ANN which gives best result at the end and hence the proposed method is called exhaustive validation and weight initialization. The most distinguishing feature of n-fold TVT approach is that it discovers the best ANN for the given dataset by exhaustive validation and weight initialization and also avoids over-fitting.

VI. DATA SETS IN THE EXPERIMENT

In the present article we used five corpuses of data-sets whose details are given in table I. Characteristics of datasets (corpus) on which experiments are performed.

TABLE I CHARACTERISTICS OF DATASETS

Datasets	Number of documents	Number of attributes (terms)	Number of classes (categories)
CNAE	1080	856	9
Db world (bodies) (stemmed)	64	3721	2
Db world (subjects) (stemmed)	64	229	2
Gender(Female)	3232	100	2
Gender(Male)	3232	100	2
Amazon	1500	50	50
Reuters 21578(Acq)	12897	100	2
Reuters21578(Earn)	12897	100	2
Reuters21578(Grain)	12897	100	2
Reuters21578(Corn)	12897	100	2
Reuters21578(Money)	12897	100	2



	The	Proposed	paper	is	based	on	bag	of	words	approach
Document1	1	0	1	0	0	1	0	0	1	0
Document2	0	0	0	0	1	0	1	1	1	1
Document3	1	1	0	1	1	0	1	0	1	1

Fig. 2 The document vector model

VII. ANALYSIS AND RESULTS

The results obtained on five datasets through N-fold TVT approach are given in table II.

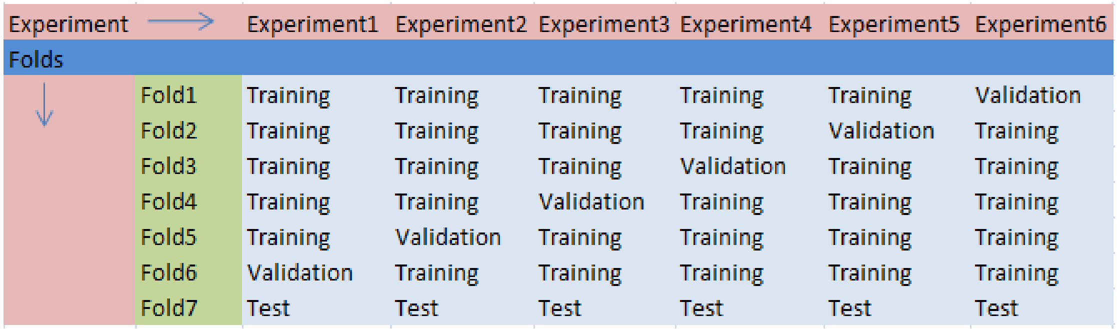
Accuracy= (Number of correctly classified instances / total instances) * 100.

TABLE II CLASSIFICATION ACCURACIES OBTAINED ON CORPUSES

Datasets	Results based on taking term frequency (tf) (%)	Results based on taking term frequency-inverse document frequency (tf- idf) (%)
CNAE	93.4	92.68
Db world (bodies) (stemmed)	98.57	98.58
Db world (subjects) (stemmed)	97.00	97
Gender(All)	69.00	68.9
Amazon commerce reviews	100.00	100.00
Reuters 21578(Acq)	94.31	94.31
Reuters21578(Earn)	97.51	97.51
Reuters21578(Grain)	99.04	99.04
Reuters21578(Corn)	99.67	99.67
Reuters21578(Money)	97.18	97.18

TABLE III CROSS-VALIDATION WITH VALIDATION APPLIED FOR PROPOSED APPROACH

Datasets	folds	Details of ANN
CNAE	4 fold TVT	4 hidden layers with 25 neurons in each hidden layer
Db world (bodies) (stemmed)	7 fold TVT	4 hidden layers with 25 neurons in each hidden layer.
Db world (subjects) (stemmed)	7 fold TVT	3 hidden layers with 20 neurons in each hidden layer.
Gender(Female)	4 fold TVT	4 hidden layers with 25 neurons in each hidden layer
Gender(Male)	4 fold TVT	4 hidden layers with 25 neurons in each hidden layer
Amazon	4 fold TVT	4 hidden layers with 25 neurons in each hidden layer
Reuters 21578(Acq)	4 fold TVT	4 hidden layers with 25 neurons in each hidden layer
Reuters21578(Earn)	4 fold TVT	4 hidden layers with 25 neurons in each hidden layer
Reuters21578(Grain)	4 fold TVT	4 hidden layers with 25 neurons in each hidden layer
Reuters21578(Corn)	4 fold TVT	4 hidden layers with 25 neurons in each hidden layer
	4 fold TVT	4 hidden layers with 25 neurons in each hidden layer



Experiment →		Experiment1	Experiment2	Experiment3	Experiment4	Experiment5	Experiment6
Folds							
↓	Fold1	Training	Training	Training	Training	Training	Validation
	Fold2	Training	Training	Training	Training	Validation	Training
	Fold3	Training	Training	Training	Validation	Training	Training
	Fold4	Training	Training	Validation	Training	Training	Training
	Fold5	Training	Validation	Training	Training	Training	Training
	Fold6	Validation	Training	Training	Training	Training	Training
	Fold7	Test	Test	Test	Test	Test	Test

Fig. 3 Example of 7 folds TVT approach

In the experiments performed and results obtained there is no feature extraction and reduction done in CNAE and Amazon datasets. All the datasets are available from UCI [10] and Keel [9] repository in term-frequency matrix and we further converted into tf-idf format. One example of proposed approach is given in Figure 3 in which 7^{th} fold is treated as test set and we discover the best friend to this test set, that is validation set, and respective network which gives minimum error on validation set is the final ANN for given test set. All the experiments are performed in matlab using 'trainscg' as training function. The formula for finding tf-idf weight value is given below.

$w=\log(1+ tf_{t,d}) \times \log_{10} (N / df_t)$

Where w= tf-idf weight vector, $tf_{i,j}$= term frequency (number of documents of i in j), N= total number of documents, df_t = number of documents containing t.

TABLE IV FINDINGS OF OTHER RESEARCHERS ON CORPUS

Dataset	Method	Accuracy		Description	Reference
CNAE-9	Feature selection technique	Decision tree	65%	Feature selection Based on Information gain Mutual information Chi-squared Symmetric uncertainty	Subhajit Dey Sarkar & Saptarsi Goswami , 2013[16]
		SVM	89%		
		Naïve Bayes	60%		
		KNN	89%		
CNAE-9	Naïve Bayes Chi-square		80%	--	Subhajit Dey Sarkar, Saptarsi Goswami, 2014[17]
Reuters 21578	Clustering word embeddings	87.74		Bag of super word embeddings,7768 (training) 3011(test)	Andrei M. Butnarua, Radu Tudor Ionescua ,2017 [3]
Reuters4	Data reduction	SVM	93.56%	--	Maria Luiza C. Passini, Katiusca B. Estébanez, 2013[12]
		Naïve Bayes	92.89%		
Reuters10	Data Reduction	SVM	93.53%	--	Maria Luiza C. Passini, Katiusca B. Estébanez ,2013 [12]
		Naïve Bayes	92.92%		
Reuters21578 R8	Naïve Bayes	90.23%		--	Ali Allahverdipoor, Farhad Soleimanian Gharehchopogh, 2016[1]
Amazon commerce reviews	Feature selection & synergetic neural networks	80%		--	Sanya Liu, Zhi Liu, Jianwen Sun, Lin Liu, 2011[20]
Db world (bodies) stemmed	Feature selection	96.87%		--	Michele Filannino [13]
Db world (subjects)stemmed	Feature selection	98.43		--	Michele Filannino [13]
Gender(All)	Naïve Bayes Chi-square	69%		--	Subhajit Dey Sarkar, Saptarsi Goswami ,2014
Reuters	Naïve Bayes Chi-square	76%		--	Subhajit Dey Sarkar, Saptarsi Goswami, 2014 [17]
Db world (All)	Naïve Bayes Chi-square	90%		--	Subhajit Dey Sarkar, Saptarsi Goswami 2014 [17]

VIII. CONCLUSION

It is revealed that n-fold TVT approach is giving better results (Table II & Table IV) than all other techniques used for classification of data as it has a capability to discover most optimum ANN for the given dataset.

REFERENCES

[1] Allahverdipoor and F. S. Gharehchopogh, "A new hybrid model of K-means and Naïve Bayes algorithms for feature selection in text documents categorization", *Journal of advances in computer research*, Vol.8, No.4, 2017.

[2] A. Kakade and K. Dhumal, S. Das, S. Jain and N. M. Ranjan, "A neural network approach for text document classification and semantic text analytics", *Journal of data mining and management*, Vol. 2, No. 2, pp.1-6. 2017. [9].

[3] A.M. Butnarua and RaduTudorIonescua, "From Image to Text Classification: A Novel Approach based on Clustering Word Embeddings", *Procedia Computer Science* Vol.112, pp.1783-92, 2017.

[4] C. Brouard, "Document classification by computing an echo in a very simple neural network", *IEEE 24th international conference on tools with artificial intelligence*, 2012.

[5] C. Naik, V. Kothari and Z. Rana, "Document classification using neural networks based on words". *International journal of advanced research computer science*. Vol. 6, No. 2, 2015.

[6] E. Rumelhart, G. E. Hinton and R.J. Williams, "Learning internal representation by error propagation", *Parallel distributed processing: Explorations in the microstructure of cognition*, Vol. 1, *Bradford books, Cambridge, MA*, 1986.

[7] G. Liu "The Semantic Vector Space Model: implementation and evaluation", *Journal of American Society for Information Science*, Vol. 48, No. 5, pp. 395–417, 1997.

[8] Hsieh Yu-Lun, Liu, Shih-Hung, Chang Yung-Chun and Hsu Wen-Lian, "Neural network-based vector representation of documents for reader emotion categorization", *IEEE 16th International conference on information reuse and integration*, 2015.

[9] J. Alcal´a-Fdez, A. Fern´andez and J. Luengo *et al.*, "KEEL data mining software tool: data set repository, integration of algorithms and experimental analysis framework", *Journal of Multiple-Valued Logic and Soft Computing*, Vol. 17, No. 2-3, pp.255–287, 2011.

[10] K. Bacheand and M. Lichman, "UCI Machine Learning Repository, University of California", *School of Information and Computer Science*, Irvine, California, USA, [Online] Available at: http://archive .ics.uci.edu/ml/, 2013.

[11] M. Dieter and R. Andreas, "Uncovering the hierarchical structure of text archives by using an unsupervised neural network with adaptive structure", *Proceedings of the 4th Pacific Asia conference on knowledge discovery and data mining, Current issues and new applications*, 2000.

[12] M. L. C. Passini, Katiusca B. Estébanez, Grazziela P. Figueredo and Nelson F. F. Ebecken, "A Strategy for Training Set Selection in Text Classification Problems", *IJACSA*, Vol. 4, No. 6, 2013.

[13] Michele Filannino, "DB World e-mail classification using a very small corpus".

[14] O. Awodele and O. Jegede, "Neural networks and its application in engineering", *Proceeding of Informing Science & IT education conference (InSITE)*, pp.83-95, 2009.

[15] P. Kumar, M. Ra and J. B. Prabhu, "Role of sentiment classification in sentiment analysis: a survey", *Annals of Library and Information Studies*, Vol. 65, pp.196-209, 2018.

[16] S. D. Sarkar and S. Goswami, "Empirical study on filter based feature selection methods for text classification", *IJCA*, Vol. 81, No.6, 2013.

[17] S. D. Sarkar, S. Goswami, A. Agarwal and J. Akhtar, "A novel feature selection technique for text classification using Naive Bayes", *Hindawi Publishing Corporation International Scholarly Research Notices* Volume, Article ID 717092, 10 pages, 2014.

[18] S. Haykin, *Neural Networks: A Comprehensive Foundation,* 2nd Edition, *1998*.

[19] S. Kumar, *Neural Networks A Classroom Approach,* Tata McGraw Hill, 2013.

[20] S. Liu, Z. Liu, J. Sun and Lin Liu, "Application of synergetic neural network in online write print identification", *International Journal of Digital Content Technology and its Applications*, Vol. 5, No. 3, 2011.

[21] Tripathi K. Tripathi, R. G. Vyas and A. K. Gupta, "The classification of data: A novel artificial neural network (ANN) approach through exhaustive validation and weight initialization", *International Journal of Computer Sciences and Engineering*, Vol. 6, No. 5, pp.241-254, 2018.