Journal of Engineering and Applied Sciences 13 (5): 1304-1309, 2018
ISSN: 1816-949X

Dimension Reduction Techniques for Document Categorization with Back Propagation Neural Network

Yaqeen Saad and Khalid Shaker
Department of Computer Science, University of Anbar, Al-Anbar, Iraq

Abstract: Text classification refers to the problem of classifying text documents into one class or more from a set of predefined categories. Text classification is significant part of "text mining". In addition, the text classification problem has become the focus of researchers because of its great importance applications in organizing large input data. Many successful algorithms applied to the text categorization. In this research, we are trying to improve performance and increase the accuracy of the results by applying the "Singular Value Decomposition" (SVD) mechanism in order to minimize the dimension of input attributes and "Feature selection" approach to choose the features that hold enough information to help in the classification task. This classification has been done by using back propagation neural network.

Key words: Text classification, feature extraction, feature selection and back propagation neural network, mechanism, predefined

INTRODUCTION

Text document classification is the procedure of assigning one or two classes to the document. Text classification problem has become the focus of researchers because of its great importance application such as 'Indexing', 'Filtering', controlling vocabulary. The studies of "Text mining" are obtaining more importance, recently because of the accessibility of increasing the numbers of the electronic documents from kinds of sources that contain 'unstructured' and 'semi-structured' information. Major goals of "text-mining" is to let users to taking out the input from textual resources and treating with the operations such as "Information retrieval." Categorization (supervised, unsupervised and semi-supervised) and "Natural Language Processing." (NLP), 'Data mining' and Machine-Learning (ML) methods work together to mechanically categorize and determine patterns from the various kinds of the documents. Text classification is an essential part of "text mining". Text Classification (TC) can be defined as assigning or structuring documents into to a defined set of classes known in advance. Text categorization deal with sorting documents based on their content while text classification is used to classify documents based on any kind of assignment to classes by content, researcher, publisher or by language (Hijazi *et al*., 2016). In the few years that have passed, the text compartmentalization method distributing into two approaches: the first is "Knowledge Engineering" (KE) and another is "Machine-Learning" (ML). Actually, machine learning is a sophisticated method within sorting of the text that is advantageous for several domains. Many important approaches have been applied to the operation of the text classification like "Naive Bayes, support vector machine and neural network". The main disadvantage of neural network is the time spent for execution is height (Kim *et al*., 2005; Zelaia *et al*., 2011).

Text categorization procedures: Figure 1 displays the most essential steps of text classification process.

Text gathering: The first operation in the document categorization process is the gathering of text document. This data can be in different formats such as 'doc', 'web-content', 'pdf'. These data collections are separated into phases which are training/phase and testing/phase. The initial stage (training) It is also called as 'learning' phase or "model construction", denotes to the group of documents whose class labels are already recognized and is used to construct the classification model. In the second phases (Testing). Also, called mode usage or classification phase. "Test-data' set denotes to groups of records whose category labels are identified but when given as an input to construct categorization model should return the accurate class labels of the records (Kim *et al*., 2005; Goyal, 2007).

Dimension reduction: The issues of "dimension reduction" is introduce by Richard E. Bellman in 1961. DR

Corresponding Author: Yaqeen Saad, Departmentof Computer Science, University of Anbar, Al-Anbar, Iraq

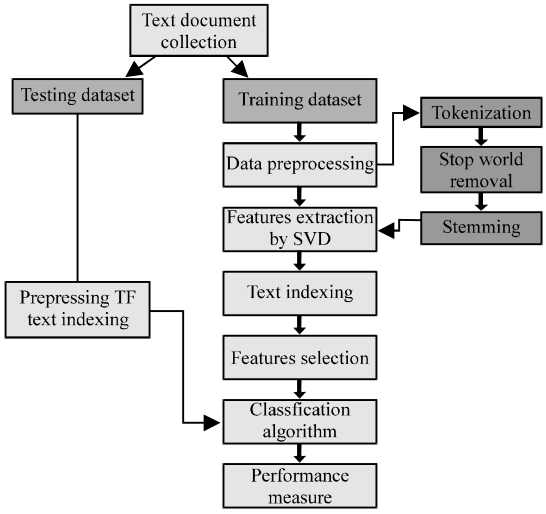

Fig. 1: Text classification process

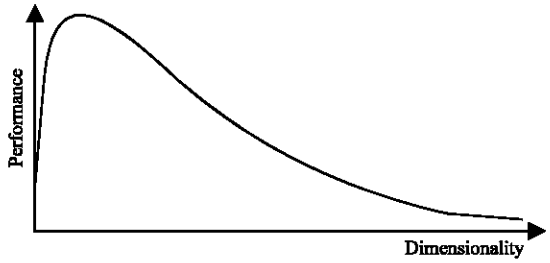

Fig. 2: Connection between the size of attribute space and the performance of the classification model

Table 1: Filter approach method

Parameters	Methods
Document frequency	$df = \sum_{t=1}^{M}(Ai)$
Information gain	Gain (A) = Info (D)-Info $_A$(D)
Mutual information	MI (t, c) $\log \frac{A+N}{(A+c)\times(A+b)}$

is the problem of input data with a large dimensional feature space. Since, many problem like "pattern Recognition", "text classification", cannot treat with high dimensional feature space. There are many approaches for DR are developed by the researcher in order to reduce the dimension of input data set. There are two approaches for dimension reduction these are (Saeys *et al*., 2007).

First approach is feature extraction: This stage is very important, especially if the input data is too large (high dimension input space). FE additionally-referred to as "feature transformation". There are many approaches applied for the extraction of feature like "principle component analysis", "linear discriminate analysis". All these approaches aim to reduce the size of the feature space in another meaning FE approach aims to transform input data set into lower feature space in order to reduce the complexity of the classification task without any trade-off in accuracy (Sebastiani, 1999).

Singular Value Decomposition (SVD): "Singular value decomposition" is well-developed technique applied to the operation of reducing the dimension feature space. SVD is an important technique because it is select the term that contain most sufficient information and remove noisy and irrelevant terms (Munindar, 2004).

Second approach is feature selection: Feature/selection have a large influence within "data-mining." in generic and "text mining" in specific. FS is most essential step in the problem of text classification. Feature selection aims to improve the accuracy and efficiency of text classifier. FS is well technology applied to choose the attribute that hold enough data and delete noisy and irrelevant features. In the operation of feature selection, there are two methods or approaches "wrapper and filter" approach (Saeys *et al*., 2007; Li *et al*., 2009). Figure 2 shows relevance, among the quantity of datum and reliability of the classifier.

Within filter approach sub-group of attributes will be elected depending on "Scoring metrics" of each feature via. applying several techniques such as 'document-frequency', 'information-gain', 'chi-square' Table 1 shows filter approach.

Preprocessing: The second important step within the Text classification is "text preprocessing". The chief objective of the preprocessing is to get the basic features or keywords within the text stored. Preprocessing step consists of three main parts preprocessing attempt to improve text classification by removing worthless information (Ramasundaram and Victor, 2013; Erlin *et al*., 2014; Hijazi *et al*., 2016).

Tokenization: It is a procedure of separating the text into smaller parts such as word, symbol and phrase.

Stop word removal: As well known as "function term" there are common words in the text but does not carry any meaning useful to the classification process. So, this words must be removed from the text. Example of these words: "of", "the", "a" and so on.

Stemming: It is the procedure for returning terms to them trunk or origin, style where structural datum is utilized to

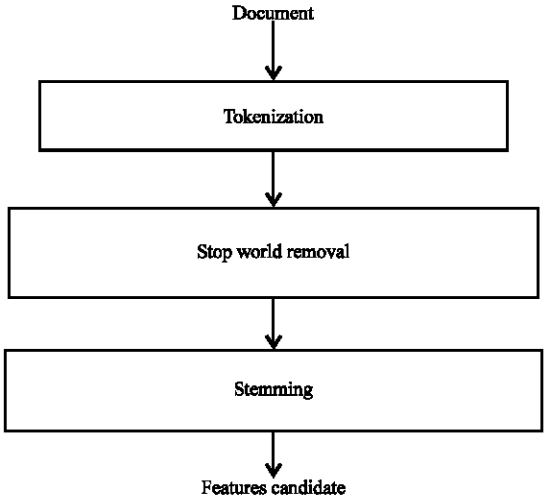

Fig. 3: Preprocessing steps

Table 2: Document word matriox with frequencies

Document	Word 1	Word 2	World 3	World m
1	3	1	2	...
2	2	4	5	...
3	0	0	1	...

correspond different diverse in terms. For instance, the verb write, writing and wrote may be returned to the origin verb write.

Text representation: After the choosing of common terms of the text document in the preprocessing step (Fig. 3). The next step is to represent these terms in proper format. Text indexing is one of the "Preprocessing" methods that are applied to minimize the text document complication and make them easier for handled. A text document is indexed as "vector space model", so, the text is performed as vectors of word. But this representation will be linked with some limitations such as high dimension input space, so, to beat this trouble. Term weighting methods are applied to allocate suitable weights to the term. Term weight is the most important part that is very necessary in the text classification problem. The most common method used for term weight is the "Term Frequency" (TF). This term refers to the number of times the term appear within the document, Table 2 represents a document word frequency (Erlin *et al*., 2014; Androutsopoulos *et al*., 2000).

Back propagation neural network: Among many algorithms offered by Artificial "Neural Network" System (ANS), 'Back propagation' is so, popular algorithm and so, helpful within recognition hard patterns and implementing of nonlinear function, Fig. 4 displays straightforward BP graph. The rounded objects carry out neurons or processing elements of a neural network. The weights are the directed lines that are connecting the

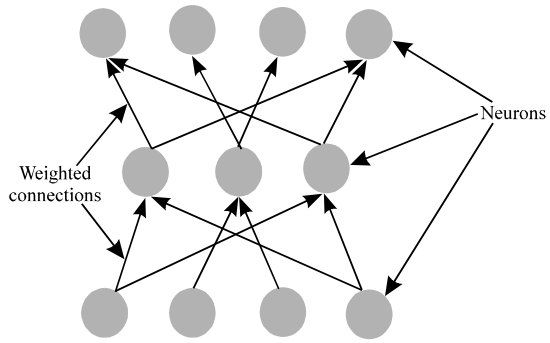

Fig. 4: Back propagation neural network

neurons. As well, every line of processing elements is layer of the network. Thus, in this figure there are "three layers" available. However, there can be multiple layers present in ANS generally there will be three layers present in the Back Propagation Network (BPNN) (Cilimkovic, 2015).

Activation function: The operation within the ANN is for gathering the product of linked weights and input signal and produce the output or activation function. For input layer this activation function is unit function. The neuron for the particular layer get the same form of activation function. In most whole states, non-linear activation functions are used. Each neuron sample is composed of handling element with synaptic input connections and single output (Zurada, 2006) The neuron output signal is given by the relationship $o = f(\Sigma)$ which is illustrated in Fig. 5.

The linear activation function: "Linear activation function" would just produce affirmative numbers over full real number range (Sibi *et al*., 2013).

Sigmoid activation function: "Sigmoid function" will just reproduction affirmative numbers among zero and one that activation task is helpful for coaching datum which is as well among 0 and 1 It is one of common applied activation function (Sibi *et al*., 2013):

$$Pan: 0>y>1, y = 1/(1+\exp(-2\times x\times s)), d = 2\times s\times y\times(1-y)$$

Literature review: In this study, the researchers operate on new dimension reduction procedures in order to minimize the size of the document vectors. They also produced "decision functions" for centroid-based categorization algorithm and support vector classifiers to handle the sorting trouble where text may belong to multiple classes. They applied three methods, centroid, orthogonal centroid and LDA/GSVD which are prepared for decreasing the dimension of clustered data (Kim *et al*., 2005).

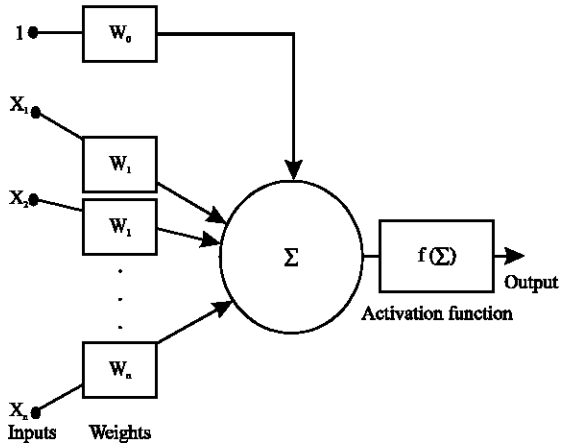

Fig. 5: Activation function

Zhang *et al.* (2007) in this study applied support vector machine and back propagation neural network for task of textual sorting of Xiang Shan Science Conference (XSSC) web documents, they are making rapprochement on the execution of the many object compartmentalization of these two learning methods. The result of an experiment demonstrated that SVM substantially outperformed the one by BPNN in prediction accuracy and recall.

By Ali and Venkateswaran (2012), feature selection which focuses on identifying relevant and informative features can help reduce the computational cost of processing voluminous amounts of data as well as increase the effectiveness for the subsequent text categorization tasks. In this study, the researchers propose a new evident theoretic feature selection approach for text categorization based on Transferable Belief Model (TBM). An evaluation on the performance of the proposed evident theoretic feature selection approach on benchmark dataset is also presented. They empirically show the effectiveness of the approach in outperforming the traditional feature selection methods using two standard benchmark datasets.

MATERIALS AND METHODS

Data set: In order to complete the classification process, we have collected data from"Reuters-21578, distribution 1.0 test collection, there are a total of 674 categories in Reuters 21578 collections". They are totally divided in to 5 fields. Each field has several categories of document collection. Table 3 shows the number of fields and categories in each field for Reuters-21578 collections.

This research fundamentally concentrates on the domain of the topic. We select 10 classes out of 135 available. They are listed as follows with the number of document for each class (Table 4).

Table 3: "Reuters-21578" set classes

Domains	Classes
Topics	135
Organization	56
Exchange	39
Place	176
People	269

Table 4: Categories with their document

Categories	Document numbers
Earn	500
Acqu	500
Money-supply	500
Trade	500
Crude	500
Coffee	500
Grain	500
Interest	500
Ship	500
Corn	500

Table 5: Contingency table for evalution measures; classfication confusion matrix

Classes	Categories	--Classfication confusion matrix--	
Predicated class	Belong	TP	FP
Actual class	Not belong	FN	TN

Performance measurements: Classification performance is recognized or measured by: 'Recall' and 'Precision' where the recall is the number of text document that are correctly categorized among all text documents belonging to that category. However, the precision is the proportion of properly labeled documents of text among all text documents that are assigned to the category by the classifier. Recall and precision used to evaluate the performance accuracy.

True Positive (TPi): It is the number of text records properly distributing in class ci.

True Negatives (TNi): The numbers of document of text properly classified as not belong for class ci.

False Positive (FPi): It is the numeral of text records falsely labeled in category ci.

False Negative (Fni): The number of text documents falsely distributing as not belonging to class ci as shown in Table 5:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Precesion} = \frac{TPi}{TPi+FPi}$$

$$\text{Recall} = \frac{TPi}{TPi+FNi}$$

Table 6: The results of BPNN for training data with IQ: 400

Class names	Recall (%)	F1 measure (%)	Percision (%)	Accuracy (%)
Earn	9.68	9.82	9.95	9.82
Acq	9.91	9.91	9.91	9.91
Trade	9.91	99.50	99.90	99.50
Money-fx	9.95	99.40	9.93	99.40
Interest	9.66	99.60	9.95	99.60
Ship	9.98	99.50	9.90	99.50
Sugar	9.99	99.80	9.97	99.80
Cofee	9.99	99.40	9.90	99.40
Gold	9.98	99.60	9.92	99.60
Crude	9.97	99.80	9.98	99.80
Average	99.30	99.40	99.40	99.40

Table 7: The results of BPP with DR when IQ: 72

Class names	Recall (%)	F1 measure (%)	Percision (%)	Accuracy (%)
Earn	9.81	9.860	9.90	9.860
Acq	9.87	9.920	9.98	9.920
Trade	9.91	9.910	9.90	9.910
Money-fx	9.96	9.940	9.93	9.940
Interest	9.95	9.970	9.99	9.970
Ship	9.99	9.985	9.98	9.985
Sugar	9.99	9.955	9.90	9.955
Cofee	9.99	9.990	9.99	9.990
Gold	9.99	9.970	9.95	9.970
Crude	9.99	9.970	9.94	9.970
Average	9.95	9.950	9.95	9.950

RESULTS AND DISCUSSION

This results show the classification of feed forward networks by applying back propagation neural network using Reute's data set. BP is procedure of training multilayer artificial neural networks that use the procedure of supervised learning. The algorithm implemented using an activation function with three hidden layer layers for BPNN algorithm. Table 6 shows the result of BPNN with all the terms 400 also before applying the techniques of dimension reduction.

Table 7 display the result of BPNN on the training data with group of features consist of 100 terms chosen by MI this group chosen according to the scoring metric for each feature. In addition, this size will be reduced to 72 terms by Principle Component Analysis (PCA) technique. Table 6 show how the techniques of dimension reduction which are (Feature selection using Mutual information) and (Feature extraction using singular value decomposition) will be effect on the performance of the results (Fig. 6).

As is clear between the results of the two tables there is a clear improvement when applied the technique of dimension reduction. In addition, back propagation neural network introduces good result for this type of data set but BPNN has drawback in the long time it takes during implementation (Fig. 7).

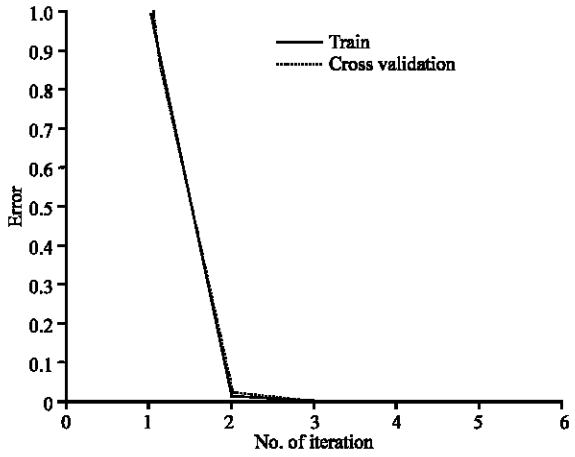

Fig. 6: Training vs. valid curve (testing curve λ = 0.010000)

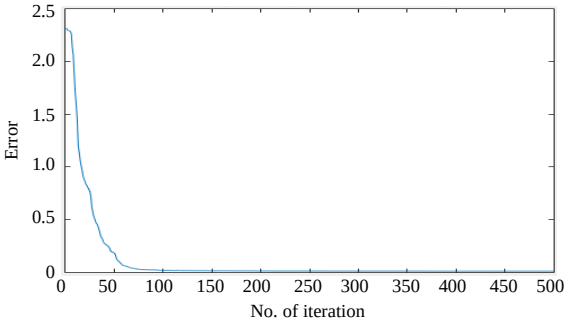

Fig. 7: Testing curve for BPNN valid vs. train curve (λ = 0.01000)

CONCLUSION

The categorization problem is one of the main issues in the text classification in specific and machine learning in generic. TC refers to the problem of assign one or more class label to the text document. Several techniques for dimension reduction such as feature extraction and feature selection are applied in the operation of document in order to improve results efficiency of text classification. Back propagation neural network algorithm is used to make the classification task with data set collecting form "Reuters-21.587". The performance of this research is measured by evaluates recall, precision, F1-measure and accuracy.

SUGGESTION

Within future research the classification of the text can be extended by implementing the categorization engines on the entire Reuters-21578 collection of documents.

REFERENCES

Ali, U. and J. Venkateswaran, 2012. An evident theoretic feature selection approach for text categorization. Intl. J. Comput. Sci. Eng., 4: 1193-1193.

Androutsopoulos, I., J. Koutsias, K.V. Chandrinos and C.D. Spyropoulos, 2000. An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages. Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. July 24-28, 2000, ACM, Athens, Greece, pp: 160-167.

Cilimkovic, M., 2015. Neural networks and back propagation algorithm. BA Thesis, Institute of Technology Blanchardstown, Dublin, Republic of Ireland.

Erlin, E., R. Rahmiati and U. Rio, 2014. Two text classifiers in online discussion: Support vector machine vs back-propagation neural network. Telecommun. Comput. Electron. Control, 12: 189-200.

Goyal, R.D., 2007. Knowledge based neural network for text classification. Proceedings of the IEEE International Conference on Granular Computing GRC, November 2-4, 2007, IEEE, Fremont, California, ISBN:0-7695-3032-X, pp: 542-542.

Hijazi, M.M., A.M. Zeki and A.R. Ismail, 2016. Arabic text classification: Review study. J. Eng. Appl. Sci., 11: 528-536.

Kim, H., P. Howland and H. Park, 2005. Dimension reduction in text classification with support vector machines. J. Mach. Learn. Res., 6: 37-53.

Li, Y., D.F. Hsu and S.M. Chung, 2009. Combining multiple feature selection methods for text categorization by using rank-score characteristics. Proceedings of the 21st International Conference on Tools with Artificial Intelligence ICTAI'09, November 2-4, 2009, IEEE, Newark, New Jersey, ISBN:978-1-4244-5619-2, pp: 508-517.

Munindar, P.S., 2004. The Practical Handbook of Internet Computing. CRC Press, Boca Raton, Florida,.

Ramasundaram, S. and S.P. Victor, 2013. Algorithms for text categorization: A comparative study. World Appl. Sci. J., 22: 1232-1240.

Saeys, Y., I. Inzaand P.Larranaga, 2007. A review of feature selection techniques in bioinformatics. Bioinformatics, 23: 2507-2517.

Sebastiani, F., 1999. A tutorial on automated text categorisation. Proceedings of the ASAI 1999, 1st Argentinian Symposium on Artificial Intelligence, pp: 7-35.

Sibi, P., S.A. Jones and P. Siddarth, 2013. Analysis of different activation functions using back propagation neural networks. J. Theor. Appl. Inf. Technol., 47: 1264-1268.

Zelaia, A., I. Alegria, O. Arregi and B. Sierra, 2011. A multiclass/multilabel document categorization system: Combining multiple classifiers in a reduced dimension. Appl. Soft Comput., 11: 4981-4990.

Zhang, W., X. Tang and T. Yoshida, 2007. Text classification with support vector machine and back propagation neural network. Proceedings of the International Conference on Computational Science, May 27-30, 2007, Springer, Beijing, China, ISBN:978-3-540-72589-3, pp: 150-157.

Zurada, J.M., 2006. Introduction to Artificial Neural Systems. Vol. 8, West Publisher, St. Paul, Minnesota.