International Journal of Computer Trends and Technology
ISSN: 2231 – 2803 / https://doi.org/10.14445/22312803/IJCTT-V68I11P109

Volume 68 Issue 11, 64-68, November 2020

Original Article

Hybrid Combination of Error Back Propagation and Genetic Algorithm for Text Document Clustering

Ashwani Mathur

IT Consultant, Senior Manager, Capgemini America, Dallas, TX

Received Date: 14 October 2020
Revised Date: 22 November 2020
Accepted Date: 24 November 2020

Abstract *- High dimensional test data need clustering. So clustering is an important and difficult task to perform when automation is required. Many scholars are working in this field to reduce manual operation or background information passing. This paper has proposed a model for documents clustering without having back-ground information. Document term features were extracted and collect in a matrix as per term frequency value. A genetic algorithm was applied to cluster each term in a cluster as per the similarity of content. Term frequency distance was a measuring evaluation parameter for finding the fitness of the chromosome. Cluster centers representing document terms were obtained from genetic algorithms. The output of the genetic algorithm was used as a training vector for the document cluster class identification. The experiment was done on a real dataset of research articles from various fields of engineering. The result shows that the proposed model has increased the precision, recall, and accuracy parameter of document clustering.*

Keywords *- Clustering, Document Clustering, Genetic Algorithm, Text Mining, Pattern Feature.*

I. INTRODUCTION

The term data mining has the common definition that the process in which useful, factful, accurate, and previously unknown data is extracted from the large volume of data or information in various forms that can be useful for better decision making. The data mining is mostly linked with the knowledge discovery in database (KDD), "a nontrivial process of identifying a novel, valid, potentially useful, and understandable pattern in data". The correspondent definition of textual data mining is the process of extracting valid, accurate, and useful data from the large text collection volume data.

Handling and exploring such huge data and text documents is a thing of concern in the field of text data mining and information retrieval. Data clustering is one of the methods in data mining that helps the users to navigate, summarize, and organize the text documents. After the organizing of the data text documents in to organize manner, data clustering can do the browsing of the collection of documents and results in search engines by the user's query. This method causes improvement in the precision and retrieval of the information system, also an efficient way to find the nearest neighbor of a document.

But there is the problem with this method of data clustering that when the given sets of documents, like to partition in a predetermined way or automatically defined clusters, as documents assigned to the same cluster are similar to assigned to a different one. Thus, in brief, clusters having the same documents have one topic and others have a different topic. In most of the existing clustering algorithms, the documents are group as the vector space model which treats the model as the bag of the words. The high dimensionality of the feature space can be obtained from it, which leaves the big challenge for the data clustering algorithm. Due to the inherent sparseness of data, the high dimensional cannot be worked. As all features are not important for data clustering, some are irrelevant and redundant. Because some of them misguide the results of it and the selection of better clustering performance is led by reddening.

The selection of features reduces the high dimensionality of the algorithm and also provide better data understanding. The selected feature set contains the non-reliable data set of the original information. In data clustering, thus is regarded as the problem of identifying the informative words within the data sets for clustering.

II. RELATED WORK

In [7] proposes in this paper the application of data mining through the help of developing an automatic tool which is a grouping and representing tool. Which concerns the analysis and visualization of cognitive information that supports the collaborative work of learning in the classroom. The tool development is made with the combination of the

vector space model and LDA, which is validated in an experimental case study. So, the conclusion is made by case study result, a significant effect of the discussion on student learning was observed.

In [8] the authors proposed the analysis of the topic modeling concept, as in software engineering the topic modeling is made to identify which topic modeling to applied to have more software repositories. They put efforts into articles between December 1999 to December 2014 in number 167 and focus on its usage in software engineering. They studied and research data mining with the topic modeling trends in the unstructured repositories. They conclude that the majority of software engineering focused on limited access.

In [9] the author presented the concepts of the optimization problems created through the automatic classification of scientific text, which allows the groups from data sets. The usage of an evolutionary algorithm for the solving of classification problems is a common method. Though, there are only a few approaches, in which classification problems are solved and data attributes to be classified as text types. In such a way, the association of computing machinery taxonomy to get the similarity between documents, each consisting of a set of keywords.

In [10] author proposes the Mods up-based frequent itemset and rider-based optimization Moth search algorithm (Rn-MSA) for the clustering of documents. The process is started with giving of the input documents to pre-processing and then extraction is done based on TF-IDF, and wordnet features. After the extraction is done, the features selection is done with the help of frequent itemset by knowledge establishment of features. The clustering is done with the Rn-MSA, just be a mixer of rider optimization algorithm (ROA) and Moth search algorithm (MSA).

In [11] the author presented the algorithm on clustering Arabic document on bond energy with the abbreviation CADBE. This gets to try to identify and display the natural variable clusters with huge sized data. CADBE proposes in three steps to cluster Arabic document: in initial step the instantiates the cluster affinity matrix using the BEA, the second is the new method to detached the cluster matrix automatically into small coherent pieces, the third step uses the fuzzy method to merge similar clusters based on association and interrelation between the resulted clusters.

III. PROPOSED METHODOLOGY

This section briefs the proposed document clustering hybrid model of a genetic algorithm with a neural network. The proposed model block diagram was shown in Fig. 1. The explanation of each block diagram important and steps was done using some notations shown in table 1.

A. Document Preprocessing

Words are arranged in a text file for an explanation, discussion of any topic. But words have their importance in a document so few of them were used to clear the meaning and others were used for framing a sentence. In this pre-processing step words that were used for framing are removed from the content. Such a set of sentence framing words are known as Stopword (SW). Filter words that are not stopword are put I a Bag of Word file. So for the D document, each term was compared with SW and the result of this pre-processing step was BOW.

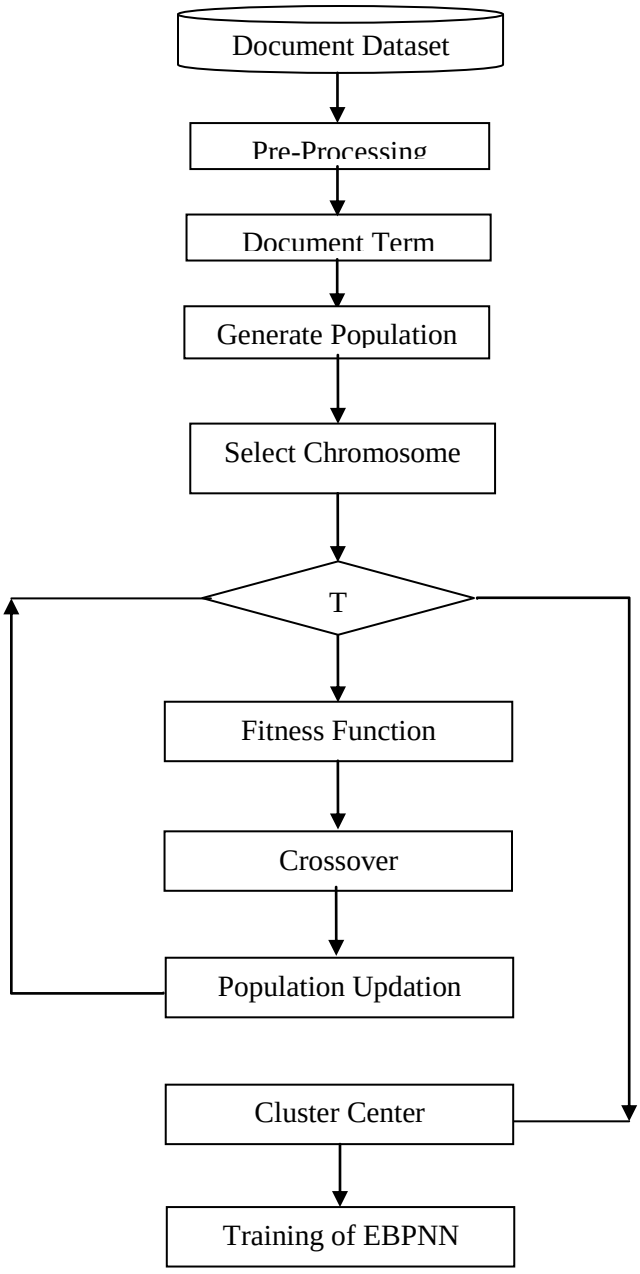

Fig. 1 Proposed hybrid model Block diagram.

BOW←Pre-Processing(D, SW)----------Eq. 1

Table 1 Notation used in the proposed model.

Notation	Meaning
D	Document
SW	Stop Word Dictionary
BOW	Bag of Word
Tf	Term Frequency
P	Population
ch	Chromosome
c	Number of Clusters
n	Number of chromosomes
F	Fitness Value

B. Document Term Frequency

Each word present in Bag of Word was the term of a document and the number of times a term repeat or found in a single document is term frequency. Words having a minimum number of presence were term as frequent. Let a term t is 5 times present in a document so this word is important and acts as a keyword for a document to decide a category. Term frequency of a document having BOW was determined by:

Term Frequency Algorithm

Input: BOW

Output: Tf

```
Loop x start 1 to BOW
        If BOW[x] has the word
                Loop y start 1 to BOW
                        If BOW[x] equals BOW[y]
                                BOW[y] is Null
                                Tf[x]← Tf[x]+1
                        Endif
                end loop
        EndIf
end loop
```

Term frequency value was a further process to normalize the term value with other document features. So each selected term in the document (having a minimum number of count) was divided by the maximum term count of the same document. This can be understand by let document Tf set is [{t1, 4}, {t2, 6},…..{tn, 3}] then if t2 have maximum term count 6 and new normalize values for each term is [{t1, 0.33}, {t2, 1},…..{tn, 0.5}].

C. Genetic Algorithm

In this step of the proposed model cluster representing documents was identified. A possible set increases exponentially in work if document number increase in the dataset so the genetic algorithm finds a feasible solution in less time. The substep of the genetic algorithm is population generation, fitness function, crossover, and population update.

D. Generate population:

Clustering needs a cluster center and each documented acts as a candidate for becoming a cluster center, so by Gaussian random function, few documents Tf were select as cluster centers. This set of cluster center Tf was termed as population or chromosome. Each chromosome has the possibility that it becomes a cluster center for the current set of documents. So if documents assigned in c number of class and n number of chromosomes (ch) were generated then population P have cXn number of Tf document set.

E. Selection of chromosome

Population P has n number of clusters but it might possible that few chromosomes have the same set of Tf for different clusters. Hence in this substep of genetic algorithm selection of a unique set of Tf for each cluster in a chromosome was passed for further substep of the algorithm.

F. Fitness Function

The chromosome has its own set of cluster centers so the fitness of a cluster center set in chromosomes was evaluated by the second algorithm. Each document Tfvalues of a dataset were parsed with chromosome Tf set and minimum distance value was summed. Summation of this difference, the value was fitness function output for a chromosome.

Fitness Function Algorithm

Input: P, Tf

Output: F

```
Loop x start 1 to P
        Loop y start 1 to D // For each document
                Loop z start 1 to c
                        d←Tf_P[x,c] – Tf_D[y]
                end loop
                F[x]←F[x]+Minimum(d)
        end loop
end loop
```

Crossover

The population has a random Tf set for cluster center representation this can be optimized by changing the current set of solutions. So this change was obtained by crossover operation. As per fitness value, the best chromosome modifies other sets of chromosomes by replacing its Tf for a cluster center with another. Crossover operation randomly selects a position c and adopt the best chromosome element value. This operation generates a new child chromosome in the population.

G. Update population

Each new child chromosome fitness value was estimated and compared with the parent chromosome if the child fitness value was better then the parent chromosome was removed from the population. If the parent chromosome fitness value was better than the child chromosome was not involved in the population. After a sufficient number of iteration, the algorithm gives an output of cluster center representative.

H. Error Back Propagation Neural Network

Cluster center feature set cluster all document Tf values as per distance between them. Now each document of a cluster was used for the training of neural network where input training vector was Tf vales of the document and desired output was cluster number. So document Tf set was passed one by one in EBPNN. The neural network has two hidden layers one input and the output layer. Weight between layers was adjusted by the backpropagation step whereby partial differentiation small change in weights was done. The sigmoidal activation function was used for triggering the neuron in the network. After training the neural network takes the Tf vector for identifying the cluster class.

VI. EXPERIMENT and RESULTS

The experiment was done on a real dataset of research papers where different branch papers were taken to cluster the input dataset. Implementation of the proposed hybrid model was done on MATLAB software. Results were compared to the following evaluation parameters:

$$\Pr ecision = \frac{True\_Positive}{True\_Positive + False\_Positive}$$

$$\mathrm{Re}call = \frac{True\_Positive}{True\_Positive + False\_Negative}$$

$$F\_Score = \frac{2 * \Pr ecision * \mathrm{Re}call}{\Pr ecision + \mathrm{Re}call}$$

Results

Table 2 Precision value comparison of text document clustering techniques.

Document Type	Proposed Hybrid Model	Previous Method
Computers	0.8333	0.7500
Electrical	0.9167	0.6667
Electronics	0.6667	0.5625

Table 2 shows precision values of the proposed hybrid model and it was obtained that the use of a genetic algorithm with an error backpropagation neural network has improved the precision by 18.1% as compared to the previously existing algorithm. Term feature transformation to normalize frequency value has increased the accuracy of work and reduced confusion as well.

Table 3 Recall value comparison of text document clustering techniques.

Document Type	Proposed Hybrid Model	Previous Method
Computers	1	0.5000
Electrical	0.7857	0.6667
Electronics	1	0.75

Table 3 shows recall values of the proposed hybrid model show that the previous model recall value was increased by 31.19%. Training of neural network by normalizing Term frequency as per cluster center obtained by the genetic algorithm has increased the performance of document clustering. The proposed model works better with highly keyword set documents.

Table 4 F-Measure value comparison of text document clustering techniques.

Document Type	Proposed Hybrid Model	Previous Method
Computers	0.909	0.6000
Electrical	0.8462	0.6667
Electronics	0.8	0.6429

Table 4 shows precision values of the proposed hybrid model and it was obtained that the use of genetic algorithm with error backpropagation neural network has improved the precision by 25.26% as compared to the previously existing algorithm. Term feature transformation to normalize frequency value has increased the accuracy of work and reduce confusion as well.

Table 5 Accuracy value comparison of text document clustering techniques.

Document Type	Proposed Hybrid Model	Previous Method
Computers	0.7778	0.6667
Electrical	0.7778	0.7222
Electronics	0.8333	0.7778

Table 5 shows the accuracy values of the proposed hybrid model shows that the previous model's accuracy value was increased by 9.3%. Training of neural network by normalizing Term frequency as per cluster center obtained by the genetic algorithm has increased the performance of document clustering. The proposed model works better with highly keyword set documents.

VI. CONCLUSION

Automatic text clustering reduced manual involvement and increased the utilization of work. This paper has proposed a clustering model where a genetic algorithm was used for text terms as per their content relevance and neural network learn those terms feature. Error back propagation neural network has increased the accuracy of document clustering by work involving the term feature as a training and testing feature. Paper has experimented on a real dataset of research articles where result shows that proposed work has increased the precision value by 18.1% as compared to other existing methods. In the future researcher can apply the same model to other language text data.

REFERENCES

[1] Abroyan N. Convolutional and recurrent neural networks for real-time data classification. Seventh International Conference on innovative Computing Technology (INTECH) : 2017 : 42-45. IEEE.

[2] Zhang Y, Er MJ, Venkatesan R, Wang N, Pratama M. Sentiment classification using comprehensive attention recurrent models. International Joint Conference on neural Networks (IJCNN) : 2016 : 1562-1569. IEEE.

[3] B. Gourav& R. Jindal, Similarity Measures of Research Papers and Patents using Adaptive and Parameter Free Threshold, International Journal of Computer Applications. 33(5) (2011).

[4] B.P.Yudha, and R. Sarrno. Personality classification based on Twitter text using Naive Bayes, KNN and SVM, In Data and Software Engineering (ICoDSE), in proceedings of International Conference 170-174. (2015) IEEE.

[5] B.Tang, H. He, et al., A Bayesian classification approach using class-specific features for text categorization. IEEE Transactions on Knowledge and Data Engineering 28(6) (2016) 1602-1606.

[6] X. Wang, J. Wang, et al., Labelled LDA-Kernel SVM: A Short Chinese Text Supervised Classification Based on Sina Weibo. In 2017 4th International Conference on Information Science and Control Engineering (ICISCE) : 2017 : 428-432. IEEE.

[7] T.-H. Chen, S. W. Thomas, and A. E. Hassan, A survey on the use of topic models when mining software repositories, Empirical Softw. Eng., 21(5) (2015) 1843–1919.

[8] M. Erkens, D. Bodemer, and H. U. Hoppe, Improving collaborative learning in the classroom: Text mining based grouping and representing, Int. J. Comput.-Supported Collaborative Learn., 11(4) (2016) 387–415.

[9] Alan Díaz-Manríquez , Ana Bertha Ríos-Alvarado, José Hugo Barrón-Zambrano, Tania Yukary Guerrero-Melendez, And Juan Carlos Elizondo-Leal. An Automatic Document Classifier System Based on Genetic Algorithm and Taxonomy. (2018).

[10] Madhulika Yarlagadda, K.Gangadhara Rao, A.Srikrishna. Frequent itemset-based feature selection and Rider Moth Search Algorithm for document clustering. Journal of King Saud University - Computer and Information Sciences (2019).

[11] Rana Husni AlMahmoud, Bassam Hammo, Hossam Faris. A modified bond energy algorithm with fuzzy merging and its application to Arabic text document clustering. Expert Systems with Applications 159 (2020).