Indian Journal of Science and Technology, *Vol 9(40), DOI: 10.17485/ijst/2016/v9i40/97722, October 2016*

ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Text Document Clustering and Classification using K-Means Algorithm and Neural Networks

Ramanpreet Kaur* and Amandeep Kaur

Department of CSE, Chandigarh University, Gharuan, Mohali - 140413, Punjab, India;
ramanpreetsidhu9@gamil.com, amancse.cu@gmail.com

Abstract

This paper demonstrated the outcomes of the research of a number of general document clustering and classification methods. **Objectives:** This research improves the clustering. Its objective is to create a system which reduces the retrieval time of text documents from clusters. **Method:** In this paper, we propose a new method supporting clustering and classification, using k-means with feed forward neural networks using MATLAB. We use k-mean for the clustering of text documents and neural networks for classification of text documents. **Findings:** Earlier various techniques have come up like semi supervised models for labelled text, namely Partially Labeled Dirichlet Allocation and the Partially Labeled Dirichlet Process, genetic algorithm, Guassian distribution, hybrid genetic algorithm, fast k means global, k-means clustering. But all these techniques have their merits as well as demerits and the common thing is that these techniques are very time consuming. That is why the main aim of the work is to develop the model based on supervised as well as unsupervised techniques to achieve the similarity between documents. **Improvements:** To remove that time consuming problem we used neural networks for classification and k-means for clustering. We developed a model based on supervised as well as unsupervised technique to achieve the similarity between documents.

Keywords: Artificial Neural Network, Cosine Similarity and Data Mining, K-mean Algorithm, Similarity Measure Function, Text Document Clustering

1. Introduction

Clustering[1] is a most important data mining technique for organizing data. In the clustering process, the data objects are grouped into number of groups and clusters for more similarity of the objects, except some are unlike to objects in the another clusters. Clustering problem can be written as:
Given

a. Dataset= A1, A2,A3……AN
b. Desired no. of clusters C
c. Function FN to find the clusters.
We need = B (1, 2……N) → (1, 2…..C)

The similarity measure is the key point to the clustering problem.

Text Clustering[2] as shown in Figure 1 is an unsupervised problem[3]; its main aim is to find the structure from collected data. In other words, it can be said that clustering[4] in the dataset of the objects collect those objects in subsets on the basis of similarities matched. That is why cluster is the collection of the objects that are similar but clearly dissimilar in the group.

Collection of the data on which clustering[5] has to be done is very crucial process of the document clustering. It is not merely a single process in fact it is the series of the sub processes. The sub process may be filtering, weighting etc. In the following figure, we will consider the various stages in the process of the clustering. It is the method of the clustering of documents[6] on similar basis. The goal of document clustering[7] is to store the documents in db of similar features. To extract the relevant features of the

**Author for correspondence*

document it is very important to represent the documents into vector space model.

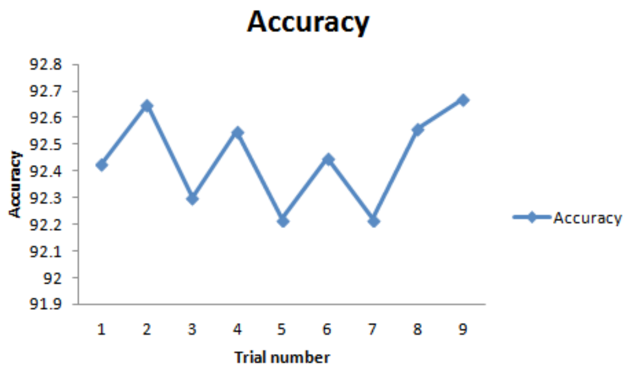


Figure 1. Accuracy graph.

K-Means algorithm[8,9] is an algorithm for clustering algorithms. It takes input and then partitions it into k clusters. Partition is done in such a way that inter cluster distance between similar clusters is very less. It is calculated using "centre of gravity". The process of k- means algorithm can be written as below;

1. Random selection of k objects from large no. of clusters.
2. Each object o is assigned as cluster.
3. Iteration will be done no. of times to get clusters.
4 .Map function will be used for assigning of the clusters.

In proposed work various methods like cosine similarity and k- means has been utilized to achieve the goal of this work.

Due to enormous increase in the use of internet, there is striking increase in the digital information. This digital information is characterized by different form of information, same information in different form, unrelated information and also there is lot of redundant information. Another next important thing to note is that most of the time we require textual information. To search or retrieve small information one has to go through thousands of documents, read all the retrieved documents irrespective whether they contain useful information or no. It becomes very difficult to read all the retrieved documents and prepare exact summary out of it within time. Earlier various techniques have come up like semi supervised models for labelled text, namely Partially labelled Dirichlet Allocation and the Partially Labelled Dirichlet Process, genetic algorithm, Gaussian distribution, hybrid genetic algorithm, fast k means global, k-means clustering. But all these techniques have their merits as well as demerits and the common thing is that these techniques are very time consuming.

Therefore, the goal of this research is to build a hybrid model which applies the supervised and the unsupervised learning approaches to reduce the gap time consumption of document clustering and classification. Based on this hybrid model, we examine various techniques like neural network[10], clustering algorithm[11] and cosine similarity measure[12].

2. A Glance of Existing Techniques

Has discussed clustering[13] technique k-mean which is partition-based clustering method. K-mean firstly initializes the center and then calculates the distance of another element.[14] proposed algorithm uses standard deviation that reduces the total time to formulate the cluster by simple k-mean. The proposed method divides the square root distance with standard deviation. Zaghoul[15] proposed the classification of documents written in Arabic language using Artificial Neural Network (ANN). This technique has been used in limited version in previous years. Arabic documents have been collected from Arabic text corpus.[15] presented a technique for classification of English text documents. The document classification has been done on the basis of content present in documents. Dataset for text classification has been obtained from various newspapers, games and sports data. It has been concluded that proposed technique works well for classification of three types of content[16–18] discussed a Hybrid Genetic k-means Algorithm (HGKA). HGKA collaborates the advantages of FGKA with IGKA and give pretty good when value of mutation probability is small or large.

3. Simulation Model

The objective of the work is to develop a hybrid model that applies the supervised with unsupervised learning techniques for the reduction of gap time consumption of document clustering with classification.

In the proposed work, utilization of vector space document is done to reduce the large document set. Neural network is used to retrieve the text document. Text document space dimension reduction is done using following two methods;

a) Parallel k means method for getting clusters and then initializes random weights between Input Layer (IL) and hidden Layer (hL). Initialize weights between output Layer (oL) and hidden Layer (hL) using neural network.

b) Then, use a clustering or classification algorithm, a similarity measure between two documents must be defined. In proposed work derive number of clusters from documents using cosine similarity measure.

c) To ensure that the system works correctly, the retrieval parameters are computed using various parameters like Precision rate, recall rate and accuracy.

In the proposed IR System feature extraction of keywords is done using similarity method and then classification is done using neural network. On the basis of similarity functions, features are extracted[19–21] and then NN network is utilized for classification. This nature is attained by changing the weights according to the output of the system. Iteration of weights is done to get the desired output according to the input. This weight adjustment is called "learning" model of neural network. Architecture of neural network is based on n inputs, m outputs and weight is assigned as w. So, mathematically it can be written:

1. Initialize weights

2. Initialize inputs and each input unit is represented as (Ci, i = 1 . . . m).

3. Each hidden unit is represented as (Xj, j = 1, . . . , a) and weighted input Signals are represented below,

$$\mathbf{X_{in} = B_{oj} + \Sigma^{m}_{i=1} C_i B_{ij}}$$

Where B0j: hidden unit j bias.

Bij: Weight between output and input unit.

$$\mathbf{X_j = g(Xim_j)}$$

4. Each output unit gets its output as show below;

$$\mathbf{Uim_k = E_{ok} + \Sigma^{a}_{j=1} X_j E_{jk}}$$

5. Output unit having activation function is shown belwo:

$$\mathbf{U_k = g(uim_k)}$$

6. Finding of back propagation error.

7. Each output unit gets its output by changing error values.

$$\mathbf{\delta_k = (y_k - l_k) g\,(uim_k\,)}$$

8. Calculation of weight function.

$$\mathbf{\Delta Ejk = \partial \delta_k X_j}$$

9. Calculation of bias correction term

$$\mathbf{\Delta E_{ok} = \partial \delta_k}$$

10. Update weights and biases:

11. Every output unit (Yk, k = 1, . . . , m) update its bias and weights (j =. . . p) :

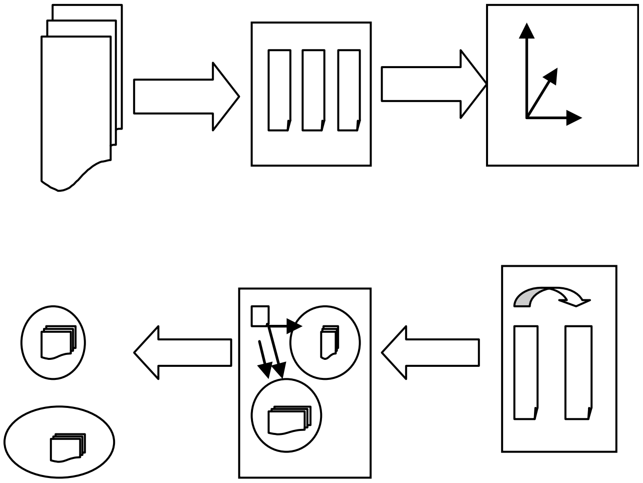
$$\mathbf{Ejk(new1) = Ejk(old1) + \Delta Ejk}$$

12. Test stopping condition.

3.1 Simulation Results

Table 1 shows the comparison of parameters like Precision rate, recall rate and Accuracy. According to this table, the graph is being drawn. Figure 2 shows the comparison graph of precision rate and recall rate. Blue line is for the precision rate and red line is for the recall rate, from the figure it is clear that the value of precision rate is less and the value of recall rate is more. The value of precision rate is varying from 0.00181 to 0.00193. The average of precision rate is come out to be 0.001924. Similarly, the value of recall rate is varying from 0.0204 to 0.0209. The average of recall rate is 0.020422.

Figure 3 shows the accuracy graph that varies from 92.4 to 92.7. Average of accuracy is come out to be 92.45. From the accuracy graph, it is clear that in the proposed work, the accuracy is enhanced.



Document cluster Mapping documents Similarity computation

Figure 2. Text clustering framework.

Table 1. Comparison Graph

Trial no.	Precision Rate	Recall Rate	Accuracy
	.00181	.0204	92.43
	.00193	.0205	92.65

	.00194	.0206	92.3
	.00195	.0207	92.55
	.00196	.0202	92.22
	.00197	.0203	92.45
	.00191	.0201	92.22
	.00192	.0209	92.56
	.00193	.0201	92.67

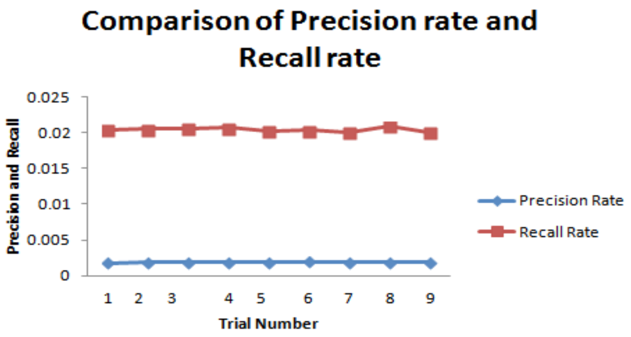


Figure 3. Comparison graph of precision and recall rate.

4. Conclusion

In IR systems the proposed model is performed on test collections Recall rate and precision values has been used for finding the system performance. Also, precision values try to get the relevant documents from non-relevant documents. Test collection of the IR experiments has been done as following:

- Set of documents for training.
- Set of queries submitted by users to get relevant documents.

So, this proposed work presented an approach to classify documents using neural network. From simulation result it has been concluded that using proposed method, the obtained results are very acceptable having good recall, precision rate having graph values of precision rate =.00191, .00193, .00194, .00195, .00196, .00197, .00191, .00192, .00193 and recall rate .0204, .0205, .0206, .0207, .0202, .0203, .0201, .0209, .0204.

5. References

1. Wenliang C, Xingzhi C, Huizhen W. Automatic word clustering for text categorization using global information. ACM Digital Library. 2004; 3411:1–11.
2. Al-Mubaid H, Syed A, Umair U. A new text categorization technique using distributional clustering and learning logic. IEEE Transactions on Knowledge and Data Engineering. 2006; 18(9):1156–65.
3. Tan CL, Chen J, Ji D, Niu Z. Unsupervised feature selection for relation extraction [Internet]. 2005. Available from: http://www.aclweb.org/anthology/I05-2045.
4. Liu H, Yu L. Toward integrating feature selection algorithms for classification and clustering. Department of Computer Science and Engineering. 2005; 17(4):491–502.
5. Martin HCL, Mario ATF, Jain AK. Simultaneous feature selection and clustering using mixture models. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2004; 26(9):1154–66.
6. Liu T, Liu S. An evaluation on feature selection for text clustering. Proceedings of the Twentieth International Conference on Machine Learning (ICML), Washington DC; 2003. p. 1–8.
7. Luo YLC. Text clustering with feature selection by using statistical data. IEEE Transactions on Knowledge and Data Engineering. 2008; 20(5):641–52
8. Vora P et al. A survey on k-mean clustering and particle swarm optimization. International Journal of Science and Modern Engineering. 2013; 3.
9. Anuradha A. Neural network approach for text classification using relevance factor as term weighing method. International Journal of Computer Applications. 2013; 68(17):37–41.
10. Ramakrishnan M. Modified k-Means algorithm for effective clustering of categorical data sets. International Journal of Computer Applications. 2014; 89(7):39–42.
11. Sruthi K, Reddy BV. Document Clustering on Various Similarity Measures. International Journal of Advanced Research in Computer Science and Software Engineering. 2013; 8(3):1269–73.
12. Verma A. Performance enhancement of k-means clustering algorithms for high dimensional data sets. International Journal of Advanced Research in Computer Science and Software Engineering. 2014; 31(4):264–323
13. Chowdhury N, Saha D, Gupta KD. English and Bengali text document classification using MST of data points. Researchgate. 2008.
14. Alckmin D, Varejao FM. Hybrid genetic algorithm applied to the clustering problem. Revista Investigacion Operacional. 2012; 33(2):141–51.
15. Zaghoul Z. Arabic text classification based on features reduction using artificial neural networks, UK Sim 15th International Conference on Computer Modelling and Simulation (UKSim), USA; 2013. p. 485–90.
16. Mohammad O. Approximate K-nearest neighbour based spatial clustering using K-D tree. International Journal of Database Management Systems. 2013; 5(1):1–25.

17. Abdullah N, Yee TC, Mohamed A, Mustafa MM, Osman MH, Mohamad AB. Control of continous stirred tank reactor using neural networks. Indian Journal of Science and Technology. 2016 Jun; 9(21):1–7.
18. Naveen A, Velmurugan T. Identification of calcification in MRI brain images by K-Mean algorithm. Indian Journal of Science and Technology. 2015 Nov; 8(29):1–7.
19. Shanmugasundaram TA, Nachiappan A. Multi-layer support based clustering for energy-hole prevention and routing in wireless sensor networks. Indian Journal of Science and Technology. 2015 Apr; 8(S7):1–11.
20. Latha A, Reddy KVK, Rao JCS, Raju AVSR. Performance analysis on modeling of loop heat pipes using artificial neural networks. Indian Journal of Science and Technology. 2010 Apr; 3(4):1–5.
21. Li Z, Yang C, Zhao K. The existence of anti-periodic soluyion for a class of cellular neural networks. Indian Journal of Science and Technology. 2010 Jan; 3(1):1–5.