iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** IV **Month of publication:** April 2019

DOI: https://doi.org/10.22214/ijraset.2019.4289

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue IV, Apr 2019- Available at www.ijraset.com

Entropy Reduction using K-Mean Clustering Algorithm

Priyanka Dayal
ECE Department, CTIEMT, Punjab

***Abstract:** This Clustering is a technique which is used in different organization which is used to organize a large number of unordered text documents into different clusters and obtained high similarity value as compare to the documents that are presents in other clusters. The utilization of document clustering is used in different area such as web mining, large organizations, in health organization and to extract information. In this research work, we are applying K-mean clustering algorithm to organize the un-structured document into structured document. To determine the similarity among them Cosine similarity algorithm is used. At last, the metrics such as entropy and accuracy is computed with the help of the MATLAB simulator.*

***Keywords:** Cluster, similarity index, entropy, accuracy, MATLAB.*

I. INTRODUCTION

Clustering divides the data into similar groups in which each group is known as a cluster, which consists of objects Similar and dissimilar with other objects group. Data can be represented in fewer clusters in order to improve efficiency. Text file clustering is used to build large documents, automatic collection. This is especially useful in mainly distributed environment, large-scale operation Document collection, such as distributed digital Library [1] and peer-to-peer (P2P) information management System, scattered on the network. A lot of Existing P2P systems use text clustering Improve the efficiency and effectiveness of information retrieval [2]. Therefore, efficient distributed clustering algorithm is needed to improve accuracy, Cluster quality when applicable to huge networks along with large text collection [3]. By sing clustering, the document with the same topics is grouped together. The main aim of the document clustering is to reduce the intra-cluster distance between the documents whereas enhancing the inter cluster distance. In figure1 the clusters are grouped into the objects that are same but different in the group. The clustering is performed on the basis of similarity index, which is explained in Figure 1.

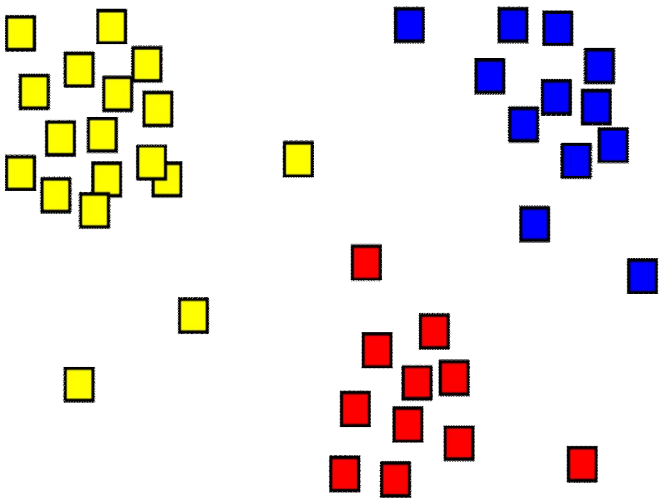


Figure 1: Scattered and clustering of data

II. SIMILARITY INDEX

A similarity index is a real valued function, which is used to determine the similarity between two groups. For similar data they take larger value whereas for dissimilar data they take zero or negative value [9]. There are different types of similarity index that are used for clustering such as Cosine similarity etc.

Table 1: Types of clustering [4, 5, 6, 7, 8]

Clustering types	Description
Partitional clustering	It is used to fragment the data into a class of dissimilar clusters.
	The group of data cannot be overlapped to each other.
Hierarchical clustering	In this type of clustering each data group is structured in the form of tree.
	It does not need an exact value of K as that of k-mean clustering technique.
	A distance matrix is used to form cluster.
	It might be top-down or bottom up scheme
Density based clustering	It is used to set the data into a class, which are strongly associated to each other.
	Generally it needs two parameters comprises of minimum number of points from the intense area.
	Due to its minimum points it can find the cluster covered by a number of clusters.
	Ecudian distance is used to measure the distance
K-mean clustering	This technique is used to divide 'n' number of interpretation into 'k' clusters by using the approximate means.
	Centroids are prepared on the basis of mean values obtained for dissimilar group.
	Depends upon the similarity index k-means divides the data into dissimilar clusters

III.ENTROPY

The concept of entropy is introduced in this paper as it represents the probability of distribution among two different documents or categories. To understand entropy, we are considering an example of entropy discussed below.

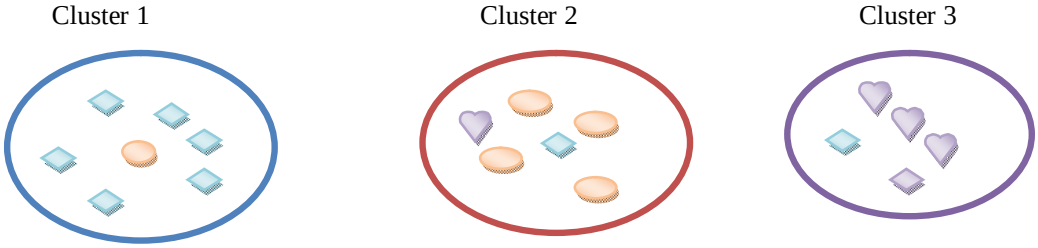


Figure 2: An example to determined Entropy

The Figure 2 shows the example of cluster from which we calculate the entropy. Cluster 1 comprises of 6 diamond shape, cluster 2 comprises of 4 circles and cluster 3 comprises of three heart shapes [13].

Then the entropy can be determined by using the formula written below:

$$H = -\sum_{i} q_i(\log_2 q_i)$$

Entropy for first cluster is given by:

$$\left(\left(\frac{5}{6}\right) \times \log\left(\frac{5}{6}\right) + \left(\frac{1}{6}\right) \times \log\left(\frac{1}{6}\right)\right)$$

Entropy for 2nd cluster is given by:

$$\left(\left(\frac{1}{6}\right) \times \log\left(\frac{1}{6}\right) + \left(\frac{1}{6}\right) \times \log\left(\frac{1}{6}\right) + \left(\frac{4}{6}\right) \times \log\left(\frac{4}{6}\right)\right)$$

Entropy for 3rd cluster is given by:

$$\left(\left(\frac{2}{5}\right) \times \log\left(\frac{2}{5}\right) + \left(\frac{3}{5}\right) \times \log\left(\frac{3}{5}\right)\right)$$

At last the total entropy of cluster is determined by:

Entropy of 1st cluster+ entropy of 2nd cluster+ entropy of 3rd cluster

As K mean, hierarchical clustering, similarity index such as Gaussian similarity, Cosine similarity, Jaccard and Pearson coefficient along with effect of entropy is discussed.

Table 2: Performance parameters of the proposed work

No. of samples	Accuracy	Entropy
1	0.924	0.235
2	0.945	0.247
3	0.935	0.296
4	0.973	0.286
5	0.917	0.276
6	0.905	0.312

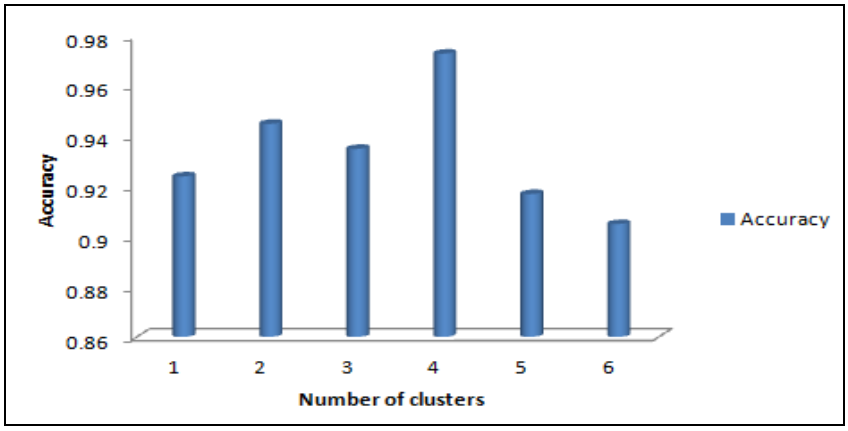


Figure 3: Accuracy of proposed work

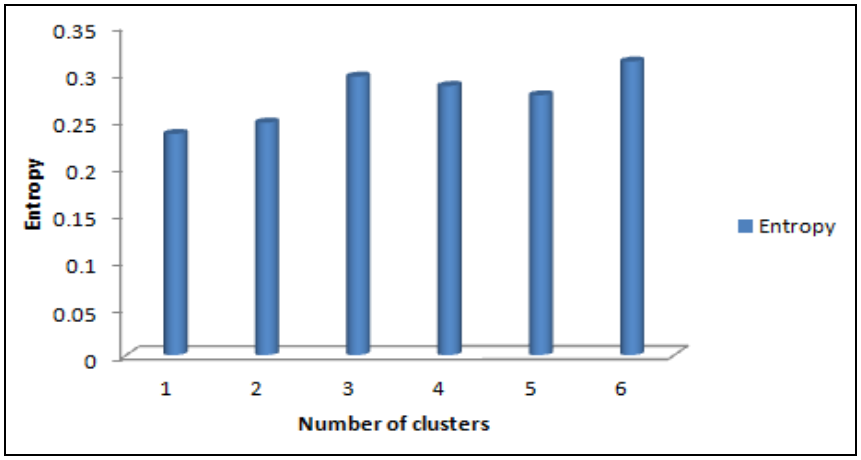


Figure 4: Entropy of proposed work

IV. RESULT ANALYSIS

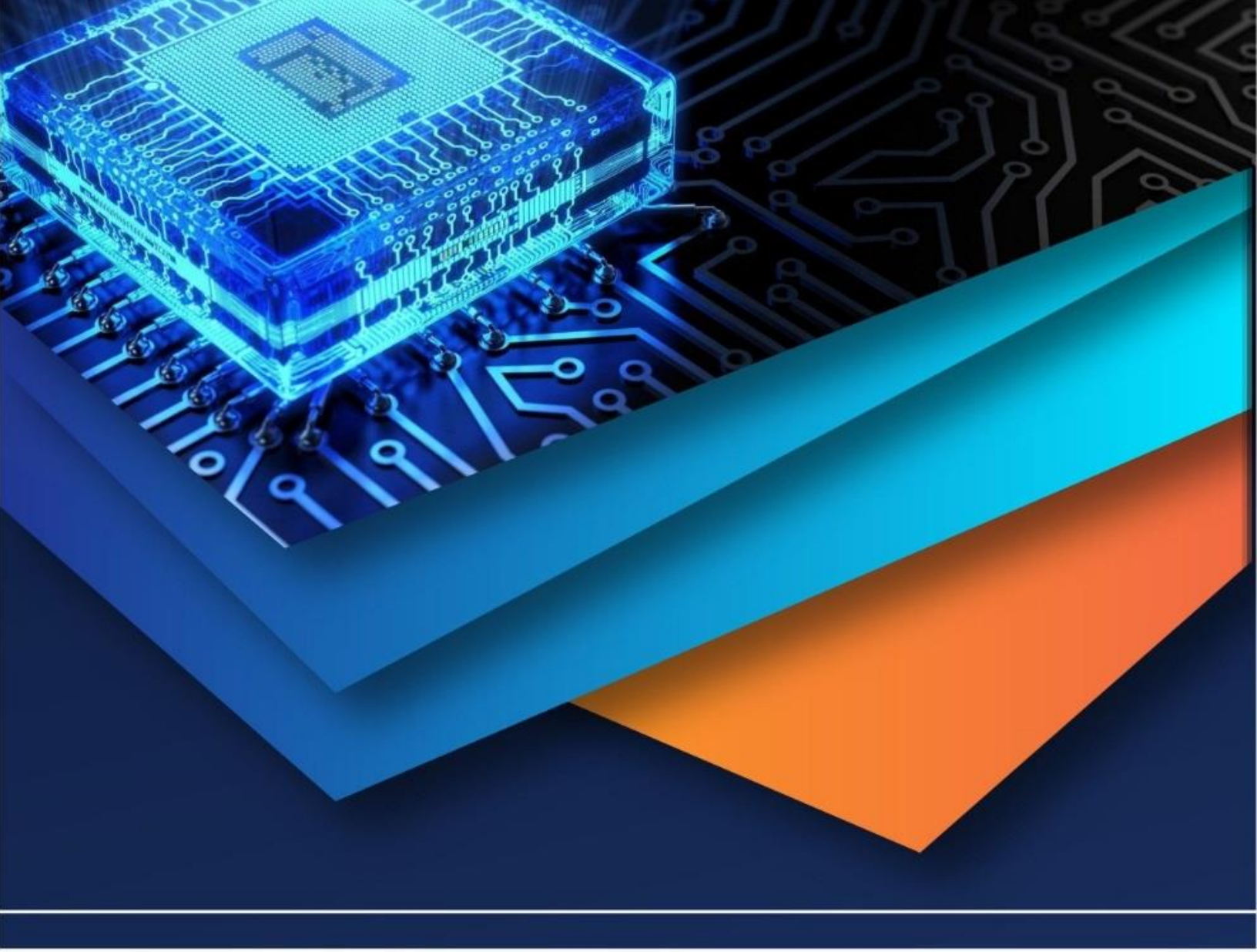
In this section, the results obtained for the proposed research work are explained. The performance parameter such as accuracy and entropy are measured. The values obtained after simulating the code in MATLAB is listed in Table 2. The accuracy obtained for the proposed work is represented in Figure 3. It is clear from the above figure that cluster 4 has maximum accuracy, which is around 0.973 and the average accuracy value obtained is 0.933. The Figure 4 represents the entropy of the proposed work. As we know that entropy represents the disorder or dissimilarity between the clusters. From the above figure it is concluded that the average entropy observed for six numbers of clusters is 0.275.

V. CONCLUSIONS

Clustering is not only the primary tool for revealing the infrastructure of a given data set, but also a promising tool for revealing the local input-output relationships of complex systems. The information entropy based clustering method is very efficient in characterizing the performance of a number of clusters. It is concluded that performance of clustering technique has been enhanced by using Cosine similarity and also the entropy maintains constant. The performance parameters like accuracy and entropy have been analyzed and it is concluded that the average values of accuracy and entropy measured are 0.933 and 0.275 respectively.

REFERENCES

[1] Steinbach, M., Karypis, G., & Kumar, V. (2000, August). A comparison of document clustering techniques. In KDD workshop on text mining (Vol. 400, No. 1, pp. 525-526).
[2] de Hoon, M. J., Imoto, S., Nolan, J., & Miyano, S. (2004). Open source clustering software. Bioinformatics, 20(9), 1453-1454.
[3] Xu, W., Liu, X., & Gong, Y. (2003, July). Document clustering based on non-negative matrix factorization. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (pp. 267-273). ACM.
[4] Zhao, Y., & Karypis, G. (2001). Criterion functions for document clustering: Experiments and analysis (Vol. 1, p. 40). Technical report.
[5] Tarabalka, Y., Benediktsson, J. A., & Chanussot, J. (2009). Spectral–spatial classification of hyperspectral imagery based on partitional clustering techniques. IEEE Transactions on Geoscience and Remote Sensing, 47(8), 2973-2987.
[6] Suzuki, R., & Shimodaira, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics, 22(12), 1540-1542.
[7] Kriegel, H. P., Kröger, P., Sander, J., & Zimek, A. (2011). Density□based clustering. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(3), 231-240.
[8] Selvakumar, J., Lakshmi, A., & Arivoli, T. (2012, March). Brain tumor segmentation and its area calculation in brain MR images using K-mean clustering and Fuzzy C-mean algorithm. In Advances in Engineering, Science and Management (ICAESM), 2012 International Conference on(pp. 186-190). IEEE.
[9] Huang, A. (2008, April). Similarity measures for text document clustering. In Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand (pp. 49-56).
[10] Kim, E., Lam, J., & Han, J. (2000). Aim: Approximate intelligent matching for time series data. Data Warehousing and Knowledge Discovery, 347-357.
[11] Hamers, L., Hemeryck, Y., Herweyers, G., Janssen, M., Keters, H., Rousseau, R., & Vanhoutte, A. (1989). Similarity measures in scientometric research: the Jaccard index versus Salton's cosine formula. Information Processing & Management, 25(3), 315-318.
[12] Norouzi, M., Punjani, A., & Fleet, D. J. (2012, June). Fast search in hamming space with multi-index hashing. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on (pp. 3108-3115). IEEE.
[13] Hermenier, F., Lorca, X., Menaud, J. M., Muller, G., & Lawall, J. (2009, March). Entropy: a consolidation manager for clusters. In Proceedings of the 2009 ACM SIGPLAN/SIGOPS international conference on Virtual execution environments (pp. 41-50). ACM.
[14] Jing, L., Ng, M. K., & Huang, J. Z. (2007). An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. IEEE Transactions on knowledge and data engineering, 19(8).
[15] Kumar, G. R., Mangathayaru, N., & Narsimha, G. (2016). A novel similarity measure for intrusion detection using gaussian function. arXiv preprint arXiv:1604.07510.
[16] Li, H., Zhang, K., & Jiang, T. (2004, August). Minimum entropy clustering and applications to gene expression analysis. In Computational Systems Bioinformatics Conference, 2004. CSB 2004. Proceedings. 2004 IEEE (pp. 142-151). IEEE.
[17] Lee, Y., & Choi, S. (2004, July). Minimum entropy, k-means, spectral clustering. In Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on (Vol. 1, pp. 117-122). IEEE.
[18] Narayanan, N., Judith, J. E., & Jayakumari, J. (2013, April). Enhanced distributed document clustering algorithm using different similarity measures. In Information & Communication Technologies (ICT), 2013 IEEE Conference on (pp. 545-550). IEEE.
[19] Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. IEEE transactions on pattern analysis and machine intelligence, 24(7), 881-892.
[20] Zhuang, Y., Mao, Y., & Chen, X. (2016, August). A Limited-Iteration Bisecting K-Means for Fast Clustering Large Datasets. In Trustcom/BigDataSE/ISPA, 2016 IEEE (pp. 2257-2262). IEEE.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY