Enliven Archive
Intensifying Thoughts

Research Article www.enlivenarchive.org Enliven: Bioinformatics

ISSN: 2376-9416

Clustering Enhancement Using Similarity Indexing to Reduce Entropy

Swathi Muppalaneni

Master of Science in Medical Sciences, Long Island University, Brooklyn & Brookville, New York, USA

*Corresponding author: Swathi Muppalaneni,Master of Science in Medical Sciences, Long Island University, Brooklyn & Brookville, New York, USA, E-mail: Smuppalaneni@gmail.com

Received Date: 15th December 2016
Accepted Date: 18th January 2017
Published Date: 23rd January 2017

Citation: Swathi M (2017) Clustering Enhancement Using Similarity Indexing to Reduce Entropy. Enliven: Bioinform 4(1): 001.

Abstract

Clustering is a helpful cheme that organizes automatically a large amount of unstructured data into clusters that comprises of structured data so that it has better similarity as compared to the unstructured document in other clusters. The use of document clustering increasing day by day due to its applicability in different area like web mining, search engine and to retrieve information. It this paper, we present K-mean clustering technique which is used to organize the unstructured data into structured data known as clusters. Then Cosine similarity index is used in order to determine the similarity measures among clusters. At last the metrics like FAR, accuracy, FRR and entropy is evaluated in MATLAB simulation tool. We made use of K-mean clustering to affectively process the segmentations and categorizations. By using the K-mean clustering, we are going to proceed with behavioral segmentation on entropies and categorize the clustering records to be evaluated with simulation tools. Making the centroid update in this study is very accurate and supported the simulations very well.

Keywords Cluster; Similarity index; Entropy; FAR; FRR; Accuracy

Introduction

Clustering splits the items 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 classified into the objects that are same but different in the group.

The above figure represents the three category of data clustering represented by different shapes and different colors. The three category of data are represented by diamond, heart and circle shapes, when clustering algorithm like "K-mean clustering is applied the data is grouped into clustered with a centroids of clusters represented by black color [4] (Table1).

The clustering is performed as per the similarity index, which is explained below.

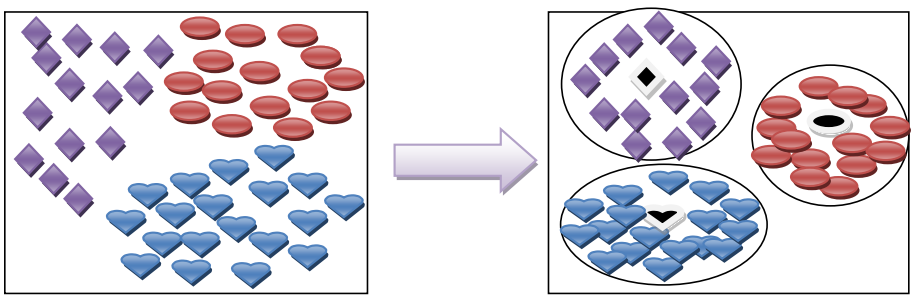

Figure 1: Scattered and clustering of data

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 category of clustering each data group is structured in the form of tree.
	It does not need an accurate 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 according to the mean values obtained for dissimilar group.
	Depends upon the similarity index k-means splits the statistics into dissimilar clusters

Table 1: Types of clustering [5-8]

Similarity Index

A similarity index is a real valued function, which is used to determine the similarity among two categories. For similar data they take larger value whereas for dissimilar data they take zero or negative value [9]. There are a number of similarity index that are used for clustering, some of them are explained below (Table 2).

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).

Distance Measure	Description	Formula
Sq-Euclidean	It is known as Squared Euclidean distance.	d(x,y)=(x-y)(x-y)
	Every centroid in the cluster specifies the average of the data.	
cosine	It is defined cosine of the incorporated angle between data points in terms of vectors minus one.	d(x,y) =1- G(xx')(yy') xy'
	Every centroid signifies the average of the points in the given cluster, after organizing the points to element's Euclidean length.	
hamming	This is used only for binary data.	$D(x,y)=\frac{1}{P}\sum_{i=1}^{p} x_i \neq y_i$
	It is the amount of bits that vary.	
	Every centre point is the item wise midpoint of data in that cluster.	
Jaccard Coefficient	It is also named as Tanimoto co-efficient used to measure the similarity. Its value is fall between 0 &1.	$D(x^{\rightarrow}, y^{\rightarrow}) = \frac{x^{\rightarrow}, y^{\rightarrow}}{x^{\rightarrow 2} + y^{\rightarrow 2} - x^{\rightarrow} \times y^{\rightarrow}}$

Table2: Types of similarity index [10, 11, 12]

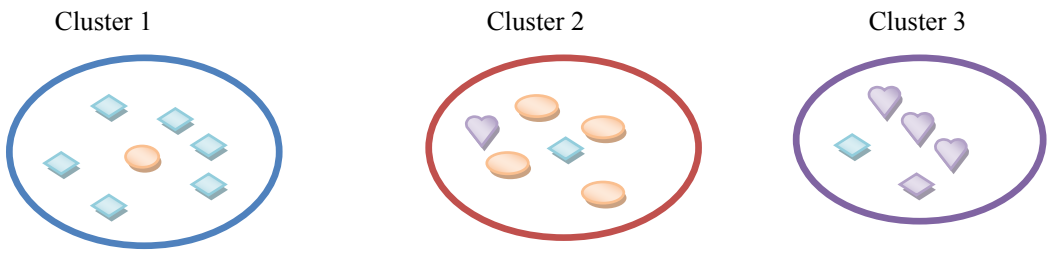

Figure 2: An example to determined Entropy

The above figure shows the example of cluster from which we calculate the entropy. Cluster 1comprises 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 1st 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 we know that entropy is used to measure the amount of disorder of a system. In clustering, the objects in the identical clusters are look similar. Therefore, we must utilize some form of entropy in the objective function of clustering [14]. A simple minimum entropy clustering measure can be defined as below:

$$H_m = \sum_{j=1}^{d} H\left(\frac{q_i}{c}\right)$$

Here, $H\left(\frac{q_i}{c}\right)$ represents the cluster's entropy C.

Algorithm 1: Minimum entropy clustering algorithm

Input: n objects, c number of clusters, initial partition performed by K-mean clustering

Output: Minimized entropy clusters

Method:

Repeat

For each data x in the dataset do
If the cluster C comprises of maximum neighbors of x is varied from the present cluster C of x then

$$h \leftarrow \sum_{y} (H_y'(C) - H_y(C)$$

Where, y is the neighbor of x and $H_y'(C)$ and $H_y(C)$ are the entropy related to y after and before assigning x to the cluster C respectively.

If h<0 **then**
Assign x to the C cluster
end if
End if
End for
Until no change

Related Work

In this section, the work done by numerous researchers in the research area of clustering algorithms like K mean, hierarchical clustering, similarity index such as Gaussian similarity, CS (Cosine similarity), Jaccard (JC) and Pearson coefficient (PC) along with effect of entropy is discussed (Table 3).

References	Proposed work	Techniques used/simulator	Outcomes
[15]	Used Gaussian similarity index for identifying the intrusion detection.	K-mean algorithm	The space complexity has been reduced and thus increases the time efficiency.
[16]	A minimum entropy criterion has been proposed. Shannon's entropy along with Havrda-Charvat's structure has been used.	h k-means/medians, hierarchical clustering.	The performed parameters shows that the presented algorithm worked better as compared to K-mean, hierarchical clustering algorithms.
			The proposed technique also efficiently identify outliers that is very helpful in practice
[17]	An Quadratic distance measure algorithm has utilized to minimize the entropy in clustering.	Gaussian kernel function is utilized to determine the first eigenvector related to the largest eigenvalue of gausian function.	It has been observed that Gaussian function provide better results than k-mean clustering because k-mean clustering is done by using Euclidean distance between data pint and the mean vector.
[18]	The experimental results of three similarity index clustering algorithms named as cosine similarity (CS), Jaccard (JC) and Pearson coefficient (PC) has been proposed.	Cosine similarity (CS), JC (Jaccard) and PC(Pearson coefficient) have been used.	The Jaccard and Person co-efficient algorithm generate more coherent clusters.
			The performance of the proposed algorithm like accuracy ad clustering quality has been improved.
[19]	A simple and proficient execution of k-means clustering method has been presented.	K mean clustering, filtering algorithm	Higher effectiveness has been achieved as data position is not varied throughout the calculation thus the data structure is not evaluated at each stage.
			The clusters are divided more accurately and thus enhance the speed of the algorithm.
			Filtering algorithm is more accurate than k-mean.
			By using K-means as a post processing step the distortion increases more powerfully.
[20]	Bisecting K-mean clustering has been used.	K-mean clustering Bisecting K-mean clustering	It is concluded that Bisecting K-mean clustering perform well in terms of efficiency, quality, susceptibility when compared with the K-mean clustering.

Table 3: Comparative study of existing techniques in clustering

Result Analysis

In this section, the results obtained for the proposed research work are explained. The performance parameter such as false acceptance rate, FRR, accuracy and entropy are measured. The values obtained after simulating the code in MATLAB is listed below (Table 4) (Figure 3):

No. of samples	FAR	FRR	Accuracy	Entropy
1	0.785	0.775	0.924	0.235
2	0.688	0.768	0.945	0.247
3	0.568	0.745	0.935	0.296
4	0.795	0.812	0.973	0.286
5	0.882	0.782	0.917	0.276
6	0.846	0.796	0.905	0.312

Table 4: Performance parameters

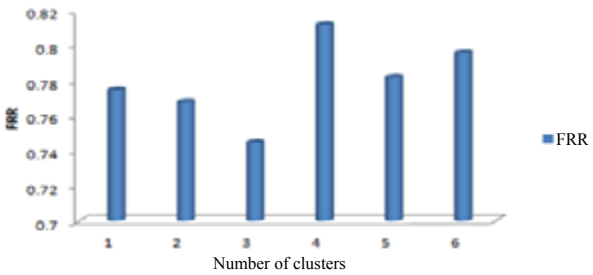

Figure 3: FAR of proposed work

The above figure shows the FAR of proposed work. Here FAR represents the false acceptance rate. From the above figure we observed that when six numbers of clusters are equated with the true data, the FAR values observed are shown in figure above. From the above figure it is observed that cluster 5 has maximum FAR value and cluster 3 has less FAR value. The average value obtained for the FAR is 0.760 (Figure 4).

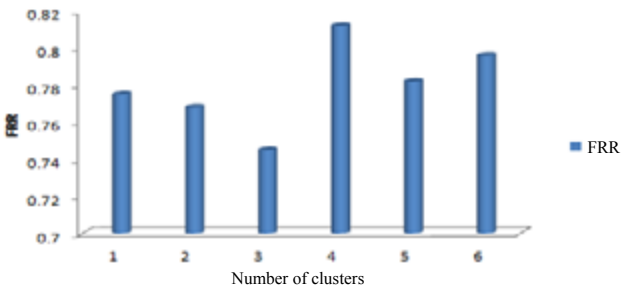

Figure 4: FRR of proposed work

The above figure represents the FRR value of the proposed research work. X-axis signifies the clusters value whereas y-axis represents the FRR values obtained for six numbers of clusters. The average value of FRR observed for the proposed research work is 0.779 (Figure 5).

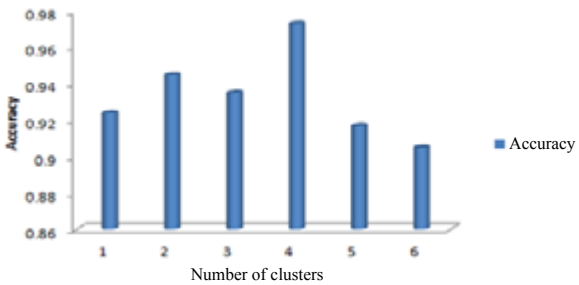

Figure 5: Accuracy of proposed work

The accuracy obtained for the simulated code is displayed in figure 5. 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 (Figure 6). The above figure represents the entropy determined for the simulated work. As we know that entropy represents the disorder or dissimilarity between the clusters. From the above figure we observed that the average entropy observed for six numbers of clusters is 0.275.

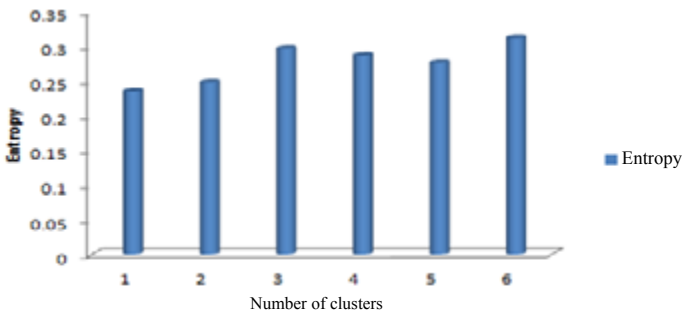

Figure 6: Entropy of proposed work

Conclusion

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 FAR, FRR, accuracy and entropy has been analyzed and it is observed that the average values of FAR, FRR, accuracy and entropy measured are 0.760,0.779,0.933 and 0.275 respectively.

References

1. Steinbach M, Karypis G, Kumar V (2000) A comparison of document clustering techniques. In KDD workshop on text mining 400: 525-526.
2. de Hoon M, Imoto S, Nolan J, Miyano S (2004) Open source clustering software. Bioinformatics 20: 1453-1454.
3. https://academic.oup.com/bioinformatics/article/20/9/1453/195075.
4. Zhao Y, Karypis G (2002) Criterion functions for document clustering: Experiments and analysis, Technical report 1: 40.
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: 2973-2987.
6. Suzuki R, Shimodaira H (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics 22: 1540-1542.
7. Kriegel HP, Kroger P, Sander J, Zimek A (2011) Density-based clustering. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1: 231-240.
8. Selvakumar J, Lakshmi A, Arivoli T (2012) 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 186-190.
9. Huang A (2008) Similarity measures for text document clustering. In Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand 49-56.
10. Kim E, Lam J, Han J (2000) Aim: Approximate intelligent matching for time series data. Data Warehousing and Knowledge Discovery 1874: 347-357.
11. Hamers L, Hemeryck Y, Herweyers G, Janssen M, Keters H, et al. (1989) Similarity measures in scientometric research: the Jaccard index versus Salton's cosine formula. Information Processing & Management 25: 315-318.
12. Norouzi M, Punjani A, Fleet DJ (2012) Fast search in hamming space with multi-index hashing. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on pp: 3108-3115.
13. Hermenier F, Lorca X, Menaud JM, Muller G, Lawall J (2009) Entropy: a consolidation manager for clusters. In Proceedings of the 2009 ACM SIGPLAN/SIGOPS international conference on Virtual execution environments 41-50.
14. Jing L, Ng MK, Huang JZ (2007) An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. IEEE Transactions on knowledge and data engineering 19: 1026-1041.
15. Kumar GR, Mangathayaru N, Narsimha G (2016) A novel similarity measure for intrusion detection using gaussian function.
16. Li H, Zhang K, Jiang T (2004) Minimum entropy clustering and applications to gene expression analysis. Proceedings. In Computational Systems Bioinformatics Conference, 2004. 142-151.
17. Lee Y, Choi S (2004) Minimum entropy, k-means, spectral clustering. In Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference 1: 117-122.
18. Narayanan N, Judith JE, Jayakumari J (2013) Enhanced distributed document clustering algorithm using different similarity measures. In Information & Communication Technologies (ICT), 2013 IEEE Conference 545-550.

19. Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R, et al. (2002) An efficient k-means clustering algorithm: Analysis and implementation. IEEE transactions on pattern analysis and machine intelligence 24: 881-892.

20. Zhuang Y, Mao Y, Chen X (2016) A Limited-Iteration Bisecting K-Means for Fast Clustering Large Datasets. In Trustcom/BigDataSE/ISPA IEEE 2257-2262.