Advances in Engineering, Management and Sciences
J. Mech. Cont.& Math. Sci., Special Issue, No.-3, September (2019) pp 183-193

Evaluate the Performance of the Clustering Algorithms by Using Data Discrepancy Factor

S Govinda Rao[1]*, N V Ganapathi Raju[2], A Sai Hanuman[3], P Varaprasada Rao[4]

[1,2,3,4] Department of Computer Science and Engineering, GRIET, Hyderabad, India.

*Corresponding author: S Govinda Rao

Email: govind.griet@gmail.com

https://doi.org/10.26782/jmcms.spl.3/2019.09.00014

Abstract

DDF is the most valuable measure among various cluster performance techniques to evaluate the perfectness of any cluster mechanism. Normally, best clusters are evaluated by computing the number of data points within a cluster. When this count is equivalent to the number of required data points then this cluster is considered to be perfect. The excellence of the cluster methodology is essential not only to find the data count inside a cluster but also to examine it by totaling the data points these are (i) present within a cluster where it should not be and vice versa and (ii) not clustered i.e. outliers (OL). The main functionality of DDF is that all cluster points can be grouped in similar clusters without outliers, the present paper highlights on how compared to DDF more efficient Clusters can be formed through the Modern DDF. Further, we evaluate the performance of some clustering algorithms, K-Means. Recently we developed the Modified K-Means Algorithm and Hierarchical Algorithm by using the Data Discrepancy Factor (DDF).

Keywords : K-Means, Modified K-Means, Hierarchical Clustering, DDF, and Modern DDF.

I. Introduction

DDF is the most important measuring technique for the cluster performance The excellence of the cluster methodology is not to be adjudicate depended on only the data quantity inner cluster, therefore the effectiveness of a clusters need to be verified by totalling the data points these are (i) currently within a cluster where it should not be and vice versa and (ii) not clustered i.e. outliers (OL). The main functionality of DDF is all cluster points can be grouped in similar clusters without outliers, but Cluster performance computation is the DDF calculation. It is computed by using standard formula given below. DDF is the very important calculation among all other measures to evaluate the performance of any clustering in this paper compared To DDF the Modern DDF is more efficient Clusters can be formed. The

DDF can be calculate the how the clustering data points will be arranged in each cluster in K Means and Modified K Means and similarly in hierarchical clustering algorithms[II].

II. Related Work

Liu Xumin and Guan Yong developed new approach of k-means algorithm is required to keep few data in all iterations and it is to be supplied in the next iteration. This algorithm escaping calculation of the distance of each data objects to the cluster centers again and again, compressing the execution time. Experimental results gave the improved algorithm can efficiently increase the speed of clustering accuracy and decreasing the computational time of the k-means [III].

JuanyingXie introduced a new approach of K-means clustering algorithm, this clustering algorithm giving effective results in point of the square mean clustering error. It does not bet on any initial values by experimenting the standard K-means algorithm as a local search procedure. It executes in an ordered way to pursue to nominal add one new cluster center at each stage, but it also caused its heavy computational load. The most important step in this algorithm is to finding the initial center for the next new cluster center at each stage. This algorithm also reduced its computational time [X]. In this paper we introduced new kind of K Means Clustering algorithm, compared to above algorithms it run time is very low and also it improves cluster accuracy.

III. Data Descripency Factor (DDF)

It is a new approach to observe the how the cluster data points are occurred in each inside of the cluster and it evaluate the performance e of modified algorithm of K-means, while accomplish the clustering on benchmark data set of IRIS and h-g indices. Both IRIS, as like g-h indices datasets, exquisitely clustered into their relevant groups. Their outputs are shown in tables 1 and 2.

$$DDF = \frac{1}{C_k}[(WI + WO + OL)] \times 100 \tag{1}$$

It was calculated by summing the total count of (i) "wrong data points" occured inside of cluster (WI), (ii) the 'Right' data points occurred outside of the cluster (WO) of any one k^{th} cluster and (iii) total count of data points, which are not to be clustered i.e. the outliers (OL) when found with the courier data (Ck). In the final computation, it is given as a fraction of the final count of data points (N). Perfection of the DDF equal to 0%, i.e. it means all the data points are clustered effectively it should a nil outlier. Similarly other cluster performance is also measured same way using the DDF calculation. It is computed by using the equation given above. The DDF is very effective scope among all other quantum to assess the accomplishment of any clustering technique. Properly, the best clustering is determined by calculating the number of objects within a cluster. If the total objects matched to the number of desired data points, so those cluster are identified to be perfect. The effectiveness of the clustering techniques need not be evaluated depending on only the data count

inside a cluster, but rather the effectiveness of a cluster must be examined by aggregating up the data points which are (i) current within a cluster where it should not be and vice versa and (ii) not clustered i.e. outliers (OL). From tables 3 and 4, it signifies that the modified algorithm of K-means conferred in the present paper is performed well than normal [IV].

Table 1 DDF computation on IRIS data using algorithm K-mean

#Cluster	Data Points	Target	Observed	Wrong data points	OL	Modern-DDF (%)	Conventional DDF (%)
1	1-50	50	61	14	0	{14+0+3+1/150}*100=12%	11+1+11+1/150}*100=16%
2	51-100	50	49	0	1		
3	101-150	50	39	3	0		

Table 2 DDF calculation on IRIS dataset using modified K-Means algorithm

#Cluster	Data Points	Target	Observed	Wrong data points	OL	Modern-DDF (%)	Conventional DDF (%)
1	1-50	50	49	0	1	{0+14+2+1/150}*100 = 11.3%	{1+12+12+1/150}*100=17.3%
2	51-100	50	62	14	0		
3	101-150	50	38	2	0		

IV. Simulation Results and Analysis

K-Means Algorithm

This Algorithm is also an iterative clustering process, but it predefines the figure of cluster that will be in your dataset. The algorithm begins by significant "*centroids*", which are point that will finally expedite to the center of each cluster. The figure of centroids chosen therefore, determines the number of clusters in the dataset [V]. The centroids are placed at possible spots in the dataset, then choose a detachment metric to decide how far away each centroid is from each of the data objects. The distances of the substance that is closest to each of the centroids are then averaged, and the centroid is enthused to the center of the relative data points. This process recurs by identifying new distances from each of data points to the centroids. The algorithm ends when centroids no longer progress within a confident threshold of aloofness. The neighboring data objects to each of the centroids are the ensuing clusters. This process is similar to hierarchical clustering. In that, it uses a distance metric to form clusters. It is dissimilar in that the total clusters for K-Means is predefined, whereas hierarchical clustering creates levels of clusters [I].

Hierarchical clustering

This type of hierarchical algorithms works by linking the data one by one on the preference of the adjacent distance computed of all the pairwise distance between the data points. In this way, the process is, grouping the data until one cluster is built. Now on the basis of dendogram graph, we can measure the total number of clusters that are actually present.

Modified K Means Clustering Algorithm

1. Initialization: choose k initial centroids arbitrarily (or randomly).
   Current centroid = randomly generate values for
   Each. Done = False
2. Nominate each objects to the centroid that is closer to it
   All instances cluster = none
   WHILE not done
   Total distance = 0
   Done = true
3. Calculate the distance among the centroids and objects using the Euclidean Distance equation.
4. For each instance's previous cluster = instance's cluster.
5. measure Euclidean distance to each centroid
6. Find smallest distance and assign instance to that cluster.
7. Restore centroids. The cluster new centroids are consisting of mean distance of all the points within that Cluster.
8. If new cluster !=previous cluster
9. Done=False

10. add smallest distance to total distance
11. Report total distance
12. Compute the distance by using sum of squares with in-cluster.
13. Value among the centroid of its cluster and each data Point.
14. Execute the K Means Algorithm n times (n=5) and return the Clustering with the smallest sum of squares within-Cluster.
15. Update the centroids
16. This process will stop when new centroids are unchanged.otherwise move to step 3.

V. Results and Discussions

Existing algorithm of K-means has been modified by introducing clustering and then repeating k-means program to attain more nearer attributes of convergence. In the below chart is shown the conventional Data Discrepancy factor for Clustering algorithms on RIS, ZOO and h-g indices datasets [VI].The below chart shows the Modern-DDF.Chart for these clustering algorithms on same datasets. Compared to conventional DDF, the Modern- DDF is best

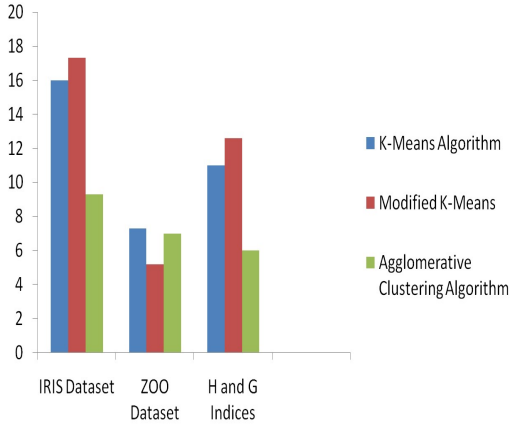

Fig1. Conventional DDF Chart

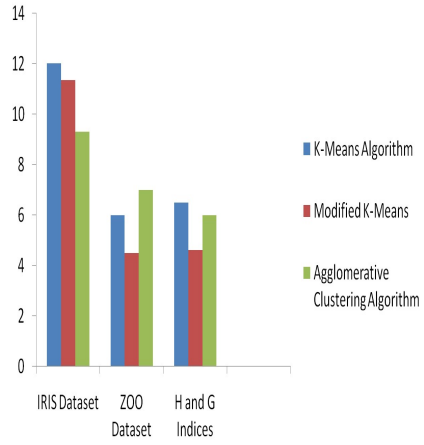

Fig2. Modern- DDF Chart

Accuracy of the K Means Clustering

The below table shows the accuracy of the K Means algorithm. This Algorithm is applied on Benchmark datasets ZOO and IRIS and also my own data set H and G Indices of the various author's indices values up to 150 data records in each dataset[V].

Table 3. Accuracy of K-Means algorithm

Database	Modern-DDF	Conventional DDF	Algorithm Used
IRIS Dataset	12%	16%	K-Means Algorithm
ZOO Dataset	6%	7.3%	
H- G- Indices	6.5%	11%	
IRIS Dataset	11.33%	17.33	Modified K-Means Algorithm
ZOO Dataset	4.5%	5.2%	
H- G- Indices	4.6%	12.6%	
IRIS Dataset	9.3%	9.3%	Hierarchical Agglomerative Algorithm
ZOO Dataset	7%	7%	
H- G- Indices	0.66%	0.6%	

Table 4. DDF calculation for below algorithms

Dataset	No of records	Correctly clustered records	K-means Accuracy	K- Means run time
IRIS	150	133	88.66	5.82
ZOO	150	137	91.66	6.20
H –G Indices	150	131	87.33	8.23

The below chart shown the accuracy of both the algorithms

Fig3. Clustering algorithm accuracy chart

Accuracy of the Modified K Means Algorithm

Compared to existing clustering algorithms, modified k-means algorithm accuracy is the best. The Modified K Means algorithm is also applied on two benchmark datasets and my own dataset where each data set contains 150 data records. As per my observation, the modified k means algorithm runtime is very optimal compared to existing algorithms and it can be shown below

Table5. Accuracy of Modified K-Means algorithm

Dataset	# Clusters	K Means Algorithm (Execution Time)	Modified K Means Algorithm (Execution Time)	Hierarchical Agglomerative Algorithm
IRIS Dataset	1	2.44	2.81	16.12
	2	3.52	3.06	
	3	6.63	3.30	
	4	7.70	3.72	
	5	8.84	4.11	
ZOO Dataset	1	2.16	2.01	15.87
	2	3.12	2.16	
	3	3.98	2.76	
	4	4.6	2.87	
	5	6.2	3.0	
H-G indices	1	2.10	0.12	15.33
	2	6.16	0.16	
	3	15.21	0.37	
	4	16.20	0.39	
	5	16.22	0.36	

Table6. Computational time of above algorithms

Dataset	No. of records	Correctly clustered records	Modified K-means Accuracy	Modified K- Means run time
IRIS	150	134	89.33%	3.49
ZOO	150	136	90.66%	5.61
HG index	150	143	95.33%	0.50

Computational Time

Accomplishment of the proposed algorithm was evaluated by calculating the computational time taken to done the run time using IRIS and h-g indices datasets. The results are noted in tables below.

Table 7 Execution times Varies Based on Number of Runs.

Dataset	K Means Algorithm (Execution Time)	Modified K Means Algorithm (Execution Time)
IRIS	First Run : 5.28 Second Run: 5.66 Third Run: 5.64	First Run: 3.49 Second Run: 3.30 Third Run : 3.47
h-g indices	First Run: 15.21 Second Run: 14.25 Third Run: 15.22	First Run: 0.40 Second Run: 0.37 Third Run :0.37

The contents in table6 show the number of clusters formed in three datasets by using clustering algorithms. Comparisons of the existing K Means and Hierarchical agglomerative algorithms and proposed K-Means algorithm takes very less time to form the various clusters in given datasets [VII]. The main underlying principle near clustering analysis is to organize and splitting data set base on inherent information pertained within the result of such clustering process is grouping of data points in a data set, where the objects within a group has a large set of similarity and a low degree of similarity with objects in other groups An significant field within literature is citations of articles published in various journals. An author could receive citation to his paper once it contains information unrelenting to the query of interest. As a rule of thumb, the more citations a particular paper receives, the greater benefit is expected to the publisher in terms of validity and technical quality content of the journal. This can be perceived with the fact that the topic chosen by authors to publish in one journal specifies the importance of spectators to the journal [IX]. The chosen algorithms rely upon the data need and the purpose of clustering algorithm. Some of the algorithms output in clusters of same quantity while others make clusters of dissimilar size. Some algorithms generate spherical clusters while other algorithms form lengthened clusters, and a few clustering algorithms are sensitive to outliers and

so on. However, the output is dependent on how the input data is pre-processed and represented [VIII].

h- and g-index values data set of authors who have published papers of scientific fineness in journals of repute. When mining and during information retrieval the question of quality data with scientific up-to-date information and journal source is crucial. The retrieved information from a publisher source should be error free and hold a minimum number of citations to the specific paper. Hence, data equivalence with respect to information and corresponding authors along with number of citations are deemed to be important and hence the contextual clustering analysis is presented at this time with validity metrics on the modified algorithm of K-means reported elsewhere by our group. Several validation measures have been reported since many years, with recent method anticipated every year, however, several of the initial algorithms has shown to be the largely efficient Validity measure is calculated to determine which is the most excellent clustering by finding the minimum value for authors measure. Therefore, the effective validations possible with Davies-Bouldin index, Silhouette index and quantization error are presented.

The initial step in algorithm of K-means is to divide the given data set into user defined number of clusters. The initial option of k in k-means is an interpretive decision and successive runs should be performing to obtain an optimized division of data for any chosen k value. A prior knowledge on the data structure would result in more appropriate clusters. However, as the data dimensionality enhances, it becomes ever more difficult to decide a proper K value. Therefore, considerable attention has been given to the topic of cluster validation, a procedure which tries to estimate a particular splitting up of data into clusters In order to compare validity metrics for modified k-means clustering; in this paper we used h- and g-indices data set. The size and attributes of the information set are varied from one another. The number of clusters ranged from 3 to 8. The data was clustered using the modified algorithm of K-means [V].

Table: 8 Results of cluster validity matrices

Data set h-g indices	Davies-Bouldin index	Silhouette index	quantization error
k=3	1.66	0.65	1.30
k=4	2.89	0.55	1.31
k=5	1.75	0.45	1.30
k=6	2.89	0.7	1.30
k=7	1.66	0.52	1.30

For the authors set of data, k-means runs consecutively to create the best clustering for k values of the data between 3 and 8. These best clustering's were practiced by the three validity methods, resulting in a set of values for each validity measure, one each for k = 2 through 8. These scores were then compared against each other to find the best k value for the clustering according to the validity measure. The results are given

in the table [V]. From the results it is observed that the customized algorithm of K-means is able to generate clusters based on user input and the validation metrics reported here suggest that the scores of each metric with admiration to k values are significant. The Davies-Bouldin index resulted in a reasonable number of clusters and k=3 represents best cluster.

VI. Conclusion

Analysis of clustering tells the proposed algorithm of K-means is much more speed compared to the existing algorithm in terms of computational time and accuracy, cluster performance measured by using the data discrepancy factor. The work shall be extended to include cluster plots of varying significance. It also tells that the modified Algorithm of K-means is much faster and exceeds the conventional algorithm both in terms of measured time and clustering achievement. The Davies-Bouldin index and Silhouette index detected the correct number of clusters for all the k values in the data set. Further, work is in progress to study the effect of dimensionality on cluster validation metrics.

References

I. B.Giovanni, "AClAP, Autonomous hierarchical agglomerative Cluster Analysis based protocol to partition conformational datasets." Bioinformatics Vol: 22, Issue: 14, pp: e58-e65, 2006.

II. M.Ujjwal, S.Bandyopadhyay. "Performance evaluation of some clustering algorithms and validity indices." IEEE Transactions on Pattern Analysis and Machine Intelligence Vol:24, Issue: 12, pp: 1650-1654, 2002.

III. N.Shi, L.Xumin, G.Yong. "Research on k-means clustering algorithm: An improved k-means clustering algorithm."Intelligent Information Technology and Security Informatics (IITSI), Third International Symposium on. IEEE, 2010.

IV. O.J.Oyelade, , O.Oladipupo, I.C.Obagbuwa. "Application of k Means Clustering AlgorithmFor prediction of Students Academic Performance." arXiv preprint arXiv:1002.2425, 2010.

V. R.P.Vaishali, R.G.Mehta. "Modified k-means clustering algorithm." Computational Intelligence and Information Technology. Springer, Berlin, Heidelberg, pp: 307-312, 2011.

VI. S.E.Brian, "Hierarchical clustering." Cluster Analysis, 5th Edition, pp: 71-110, 2011.

VII. S.G.Rao, A.Govardhan. "Assessing h-and g-Indices of Scientific Papers using k-MeansClustering." International Journal of Computer Applications Vol: 100, Issue: 11, 2014.

VIII. S.G.Rao, A.Govardhan. "Investigation of Validity Metrics for Modified K-Means Clustering Algorithm." i-Manager's Journal on Computer Science Vol: 3, Issue: 2, pp: 33, 2015.

IX. S.G.Rao, A.Govardhan. "Performance Validation of the Modified K-Means Clustering Algorithm Clusters Data." International Journal of Scientific & Engineering Research Vol: 6, Issue: 10, pp: 726-730, 2015.

X. X.Juanying, "An Efficient Global K-means Clustering Algorithm." JCP Vol:6, Issue: 2, pp:271-279, 2011.