Turkish Journal of Computer and Mathematics Education *Vol.12 No.3(2021), 3269-3275*

Research Article

Improving the Performance of Multidimensional Clinical Data for OLAP using an Optimized Data Clustering approach

Anjana Yadav[a], Anand Kumar Tripathi[b]

[a,b]Department of Computer Science, P. K. University, Shivpuri, India
[a]yadavanjana3485@gmail.com
[b]dr.aktripathi@gmail.com

Article History: Received: 10 November 2020; Revised 12 January 2021 Accepted: 27 January 2021; Published online: 5 April 2021

Abstract: Medicine is a fresh way to utilize for curing, analyzing and detecting the diseases through data clustering with OLAP (Online Analytical Processing). The large amount of multidimensional clinical data is reduced the efficiency of OLAP query processing by enhancing the query accessing time. Hence, the performance of OLAP model is improved by using data clustering in which huge data is divided into several groups (clusters) with cluster heads to achieve fast query processing in least time. In this paper, a Dragon Fly Optimization based Clustering (DFOC) approach is proposed to enhance the efficiency of data clustering by generating optimal clusters from multidimensional clinical data for OLAP. The results are evaluated on MATLAB 2019a tool and shown the better performance of DFOC against other clustering methods ACO, GA and K-Means in terms of intra-cluster distance, purity index, F-measure, and standard deviation

Keywords: OLAP, Cluster, DFOC, Multidimensional, Purity Index

1. Introduction

The huge amount of data is collected in the form of data warehouse [1, 2, 3] to combine all the information about organisations. This data information is very difficult to access in minimum time due the big data for OLAP. to improve the performance of OLAP, the data is organised in several groups to save the accessing time and query processing cost. This organisation of data into groups is known as data clustering. KPI (Key Performance Indicator) [4] is also merged with OLAP [5] to perform fast query processing.

OLAP is also used for power cost examination in marketable areas to diminish the expenses with increasing the influence performance [6, 7]. A multidimensional data is used for calculating the influence expenses at various stages of simplification. Hence the rule association is developed with OLAP to obtain efficient results on numerous building data using UML (Unified Modeling Language) [8] and SQL (Structured Query Language) [9, 10]. The decision support system is also developed with data clustering for fast accessing the huge data [11, 12] with maximum accuracy of information with respect to future aspects [13, 14].

Here, several researchers introduced data clustering techniques for improving the efficiency of multidimensional data model [15, 16]. K-Means is one of the widely useful clustering techniques for simple and easy development for huge amount of data. But, there is still some drawback in K-Means like highly dependable on initial cluster. So, here we utilized the optimization for data clustering on huge multidimensional data sets to obtain optimal results by removing the limitation of K-Means. The GA (Genetic Algorithm) and ACO (Ant Colony Optimization) are two most popular optimization approaches are used with data clustering to improve the quality of clustering. In this work, we implemented a DOFC (Dragon Fly Optimization based Clustering) approach on clinical multidimensional datasets to generate optimal clusters with cluster centroids and compared the results with ACO, GA and K-Means in terms of several parameters.

2. Dragon Fly Optimization based Clustering (DFOC) approach

A. DFOC approach

Dragon Fly Optimization (DFO) approach is a nature inspired methodology which is stirred by dragon fly`s stagnant and energetic behaviour on the basis of examination and utilization. DFO offers three crucial standard Severance (SR), Configuration (CF) and Consistency (CS) and two former significant convictions of brimming Foodstuff sources Appeal (FA) and Opponent Escaping (OE) represented in (1) to (5).

$$SR_p = -\sum_{q=1}^{N_n} (X - X_q) \quad (1)$$

$$CF_p = \frac{\sum_{q=1}^{N_n} V_q}{N_n} \quad (2)$$

$$CS_p = \frac{\sum_{q=1}^{N_n} X_q}{N_n} - X \quad (3)$$

$$FA_p = X^+ - X \quad (4)$$

$$OE_p = X^- + X \quad (5)$$

Here, X =dragonfly individual location, X^+ =foodstuff location, X^- =opponent location, N_n =neighbours number, V_q & X_q = q^{th} individual`s velocity and location.

The speed vector is evaluated by utilizing (6), then dragonfly`s location is updated through (7).

$$\nabla X_{t+1} = (sr.SR_p + cf.CF_p + cs.CS_p + fa.FA_p + oe.OE_p) + wt.\nabla X_t \quad (6)$$

$$X_{t+1} = X_t + \nabla X_{t+1} \quad (7)$$

Here, sr, cf, cs, fa, oe and wt are steady coefficient.

DFOC approach

START

Assign N data entities as cluster centroids randomly.

For each clusters

Initialize standards of dragonfly population (X_p) and speed vector (X_p) with p=1,2,3,.....N_n

While finish circumstance is not pleased

Calculate entire dragonfly`s intention standards

Update foodstuff and opponent source

Update sr, cf, cs, fa, oe and wt

Calculate SR, CF, CS, FA, and OE by (1) to (5)

Update neighbour`s area

If (minimum 1 neighbour locates in dragonfly area)

Update speed vector by (6)

Update location vector by (7)

Else

Update location vector by (7)

End If

Confirm and accurate next location of dragonfly based on capricious restrictions

End While

End For

STOP

In DFOC, the DFO is applied on multidimensional clinical datasets to obtain optimal clusters with cluster heads (centroids) with minimizing the intra-cluster distances among data elements. In DFO, every cluster is assigned as

dragonfly and each data entities are assigned as explore agents. All dragon fly`s positions are updated according to fitness standards with reducing the intra-cluster distances among data entities to find out the optimal clusters with centroids.

B. Multidimensional Clinical Datasets

The DFOC is applied on several multidimensional clinical datasets describing in table 1.

TABLE I. MULTIDIMENSIONAL CLINICAL DATASETS

Sr. No.	**Clinical Dataset**			
	Dataset	***No. of instances***	***No. of dimensions***	***No. of Clusters***
1	Cancer	683	9	2
2	Cryotherapy	90	7	2
3	Liver Patient	583	10	2
4	Heart Patients	297	14	4

3. Result and Analysis

The DFOC is implemented on all four clinical data sets (table 1) on MATLAB 2019a tool. The results are obtained in terms of intra-cluster distance, purity index, F-measure, and standard deviation over 1000 repetitions.

A. Intra-cluster distance

It is explained as the mean distance among data entities in identical cluster. It must have least value for optimized clustering.

B. Purity Index

It is illustrated the frequent clustering of data entities by using (8). It must have maximum value for optimized clustering.

$$P_I = \sum_{s=1}^{K} \frac{\left(\left|CR_r\right| \frac{\max(\left|CR_{rs}\right|)}{\left|CR_s\right|} \right)}{\left|D_S\right|} \tag{8}$$

Here, K = clusters number,

$\left|CR_r\right|$ and $\left|CR_s\right|$ = rth class and sth cluster length

$\left|D_S\right|$ = dataset length

$\left|CR_{rs}\right|$ = data entities of rth class locate to sth cluster.

C. -Measure

It is obtained from precision (prec) and recall (rcl) for data reclamation by (9) to (12). It must have maximum value for optimized clustering.

$$prec(r,s) = \frac{\left|CR_{rs}\right|}{\left|CR_s\right|} \tag{9}$$

$$rcl(r,s) = \frac{\left|CR_{rs}\right|}{\left|CR_r\right|} \tag{10}$$

$$Fun(r,s) = \frac{2.prec(r,s).rcl(r,s)}{prec(r,s) + rcl(r,s)} \tag{11}$$

$$FM = \sum_{r=1}^{K} \frac{|CR_r|}{|D_S|} \max\{Fun(r,s)\} \quad (12)$$

D. Standard Deviation

It is explained the data clustering strength about the mean standards using (13). It must have least value for optimal clustering.

$$S_D = \sqrt{\frac{\sum(de - \overline{de})}{|D_S|}} \quad (13)$$

Here, de = data entity in dataset,

$|\overline{de}|$ = mean of data entities in a dataset.

TABLE II. RESULTS FOR CANCER DATASET

Approaches	Performance Parameters			
	Intra-cluster distance	***Purity index***	***Standard deviation***	***F-Measure***
K-Means	94.2641	0.86	0.5248	0.84
GA	0.3265	0.87	0.2153	0.85
ACO	0.08587	0.90	0.1042	0.87
DFOC	0.002514	0.95	0.024	0.92

TABLE III. RESULTS FOR CRYOTHERAPY DATASET

Approaches	**Performance Parameters**			
	Intra-cluster distance	**Purity index**	**Standard deviation**	**F-Measure**
K-Means	**19.3625**	**0.82**	**0.3521**	**0.76**
GA	**0.3142**	**0.90**	**0.1241**	**0.85**
ACO	**0.0541**	**0.91**	**0.0624**	**0.86**
DFOC	**0.00325**	**0.95**	**0.00786**	**0.90**

TABLE IV. RESULTS FOR LIVER PATIENTS DATASET

Approaches	**Performance Parameters**			
	Intra-cluster distance	**Purity index**	**Standard deviation**	**F-Measure**
K-Means	**42.3214**	**0.87**	**0.4215**	**0.85**
GA	**0.4201**	**0.88**	**0.2641**	**0.86**
ACO	**0.0845**	**0.90**	**0.0758**	**0.87**
DFOC	**0.00464**	**0.91**	**0.00882**	**0.88**

TABLE V. RESULTS FOR HEART PATIENTS DATASET

Approaches	**Performance Parameters**			
	Intra-cluster distance	**Purity index**	**Standard deviation**	**F-Measure**
K-Means	**12.3654**	**0.78**	**0.20365**	**0.74**
GA	**0.50241**	**0.81**	**0.07548**	**0.76**

Approaches	Performance Parameters			
	Intra-cluster distance	Purity index	Standard deviation	F-Measure
ACO	0.0365	0.84	0.02364	0.80
DFOC	0.00124	0.91	0.0074	0.95

TABLE VI. RESULTS FOR AVERAGE RANK FOR ALL DATASETS BASED ON INTRA-CLUSTER DISTANCE

Approaches	Datasets				
	Cancer	Cryotherapy	Liver Patients	Heart Patients	Average Rank
K-Means	94.2641 (4)	19.3625 (4)	42.3214 (4)	12.3654(4)	4
GA	0.3265 (3)	0.3142 (3)	0.4201 (3)	0.50241 (3)	3
ACO	0.08587 (2)	0.0541 (2)	0.0845 (2)	0.0365 (2)	2
DFOC	0.002514 (1)	0.00325 (1)	0.00464 (1)	0.00124 (1)	1

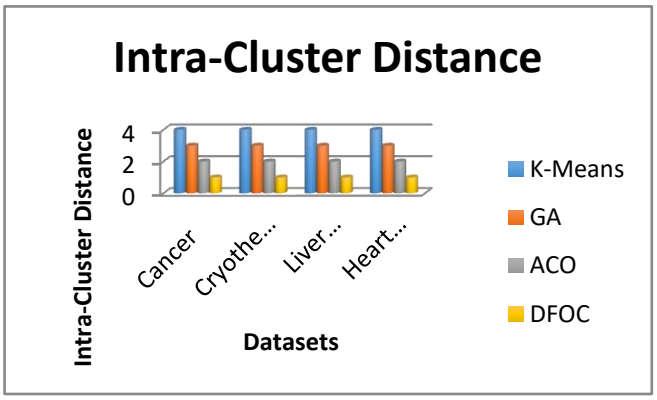

Fig. 1. Average Rank for all datasets based on Intracluster Distance

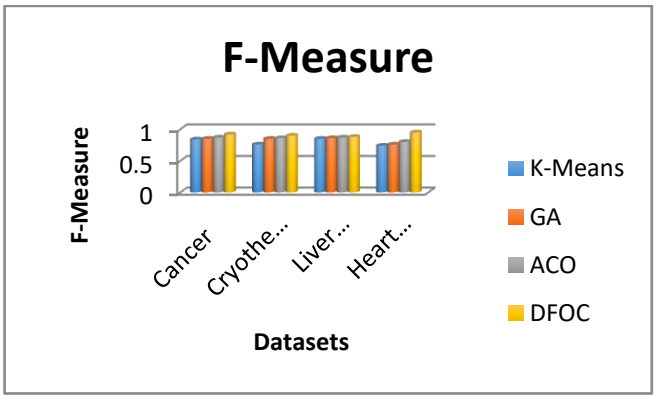

Fig. 2. F-Measure for all datasets

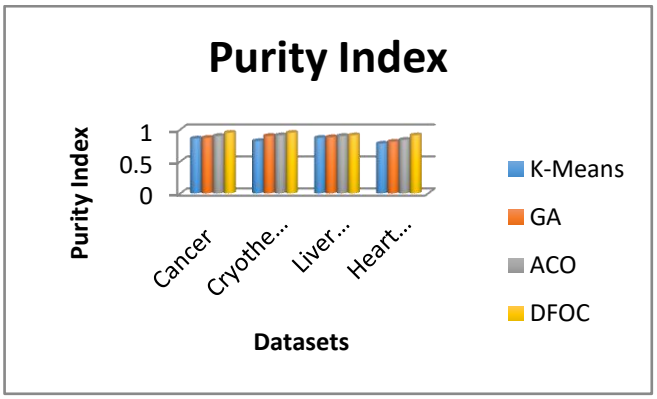

Fig. 3. Purity Index for all dataets

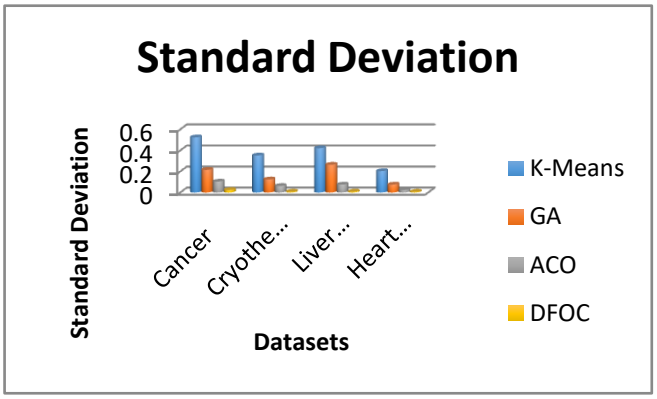

Fig. 4. Standard Devition for all datasets

The results in table II to table VI and figure 1 to figure 4 illustrates the better quality results of DFOC on all four multidimensional clinical datasets against K-Means, GA and ACO in terms of intra-cluster distance, F-measure, purity index and standard deviation. Due to better examination and utilization, DFOC improves the search space in global area for generating optimal cluster, hence DFOC generates enhanced outputs as compare to prior approaches.

4. Conclusion

In this work, a Dragon Fly Optimization based Clustering (DFOC) approach is implemented to improve the performance of data clustering by obtaining optimized clusters from multidimensional clinical data for OLAP. The outcomes are examined on MATLAB 2019a tool and illustrated the superior efficiency of DFOC as compared to prior approaches ACO, GA and K-Means in terms of intra-cluster distance, purity index, F-measure, and standard deviation.

References

A. Vaisman, and E. Zimanyi, "Mobility Data Warehouses. International Journal of Geo-Information," MDPI, Vol. 8 (170), pp. 1-22, 2019.

G. Agapito, C. Zucco, and M. Cannataro, "COVID-WAREHOUSE: A Data Warehouse of Italian COVID-19, Pollution, and Climate Data," International Journal of Environment Research and Public Health, MDPI, Vol. 17 (5596), pp. 1-22, 2020.

N. Jukic, B. Jukic, and M. Malliaris, "Online Analytical Processing (OLAP) for Decision Support," pp. 1-25, 2008.

A. Papacharalampopoulos, C. Giannoulis, P. Stavropoulos, and D. Mourtzis, "A Digital Twin for Automated Root-Cause Search of Production Alarms Based on KPIs Aggregated from IoT," Applied Science, MDPI,Vol. 10 (2377), pp. 1-16, 2020.

N. Stefanovic, "Proactive Supply Chain Performance Management with Predictive Analytics," The Scientific World Journal, Hindawi, pp. 1-18, 2014.

B. Noh, J. Son, H. Park, and S. Chang, "In-Depth Analysis of Energy Efficiency Related Factors in Commercial Buildings Using Data Cube and Association Rule Mining," Sustainability, MDPI, Vol. 9 (2119), pp. 1-20, 2017.

D. R. D. Almeida, C. D. S. Baptista, F. G. D. Andrade, and A. Soares, "A Survey on Big Data for Trajectory Analytics. International Journal of Geo-Information," MDPI, Vol. 9 (88), pp. 1-24, 2020.

C. Ciferri, R. Ciferri, L. Gomez, M. Schneider, A. Vaisman, and E. Zimanyi, "Cube Algebra: A Generic User-Centric Model and Query Language for OLAP Cubes," International Journal of Data Warehousing and Mining, pp. 1-23, 2012.

N. E. Emmanuel, J. A. Obiageli, and V. Osinachi, "Design and Implementation of Multidimensional Students Result Analytical Processing of tertiary Institutions," International Journal of Engineering and Computer Science, Vol. 8 (8), pp. 24814-24828, 2019.

L. Shen, S. Liu, S. Chen and X. Wang, "The Application Research of OLAP in Police Intelligence Decision System," International Workshop on Information and Electronics Engineering (IWIEE), Elsevier, Vol. 29, pp. 1-6, 2012.

Venkatraman, S., 2017. A Proposed Business Intelligent Framework for Recommender Systems. Informatics, MDPI, 4 (40), pp. 1-12.

I. L. Cruz, R. Berlanga and M. J. Aramburu, "Modelling Analytical Streams for Social Business Intelligence," Informatics, MDPI, Vol. 5 (53), pp. 1-17, 2018.

W. Fuertes, F. Reyes, P. Valladares, F. Tapia, T. Toulkeridis, and E. Perez, "An Integral Model to Provide Reactive and Proactive Services in an Academic CSIRT Based on Business Intelligence," Systems, MDPI, Vol. 5 (52), pp. 1-20, 2017.

W. Q. Qwaider, "Apply On-Line Analytical Processing (OLAP) With Data Mining For Clinical Decision Support," International Journal of Managing Information Technology (IJMIT), Vol. 4 (1), pp. 1-13, 2012.

J. N. S. Rubi and P. R. L. Gondim, "IoMT Platform for Pervasive Healthcare Data Aggregation, Processing, and Sharing Based on OneM2M and OpenEHR," Sensors, MDPI, Vol. 19 (4283), pp. 1-25, 2019.

J. L. S. Cervantes, M. Radzimski, C. A. R. Enriquez, G. A. Hernandez, L. R. Mazahua, C. S. Ramirez, and A. R. Gonzalez, "SREQP: A Solar Radiation Extraction and Query Platform for the Production and Consumption of Linked Data from Weather Stations Sensors," Journal of Sensors, Hindawi, pp. 1-19, 2016