Extended Jaccard Indexive Buffalo Optimized Clustering on Geo-social Networks with Big Data

M. Anoop
Research Scholar, Department of Computer Applications, Vels Institute of Science, Technology & Advanced Studies, Chennai, Tamil Nadu, India. E-mail: profanoopcs@rediffmail.com

P. Sripriya
Professor, Department of Computer Applications, Vels Institute of Science, Technology & Advanced Studies, Chennai, Tamil Nadu, India. E-mail: sripriya.phd@gmail.com

Received March 17, 2021; Accepted July 14, 2021
ISSN: 1735-188X
DOI: 10.14704/WEB/V18I2/WEB18314

Abstract

Clustering is a general task of data mining where partitioning a large dataset into dissimilar groups is done. The enormous growth of Geo-Social Networks (GeoSNs) includes users, who create millions of heterogeneous data with a variety of information. Analyzing such volume of data is a challenging task. The clustering of large volume of data is used to identify the frequently visited location information of the users in Geo-Social Networks. In order to improve the clustering of a large volume of data, a novel technique called Extended Jaccard Indexive Buffalo Optimized Data Clustering (EJIBODC) is introduced for grouping the data with high accuracy and less time consumption. The main aim of EJIBODC technique is to partition the big dataset into different groups. In this technique, many clusters with centroids are initialized to group the data. After that, Extended Jaccard Indexive Buffalo Optimization technique is applied to find the fittest cluster for grouping the data. The Extended Jaccard Index is applied in the Buffalo Optimization to measure the fitness between the data and the centroid. Based on the similarity value, using a gradient ascent function, the data finds the fittest cluster centroid for grouping. After that, the fitness value of cluster is updated and all the data gets grouped into a suitable cluster with high accuracy and minimum error rate. An experimental procedure is involved with big geo-social dataset and testing of different clustering algorithms. The series discussion is carried out on factors such as clustering accuracy, error rate, clustering time and space complexity with respect to a number of data. Experimental outcomes demonstrate that the proposed EJIBODC technique achieves improved performance in terms of higher clustering accuracy, less error rate, time consumption and space complexity when compared to previous related clustering techniques.

Keywords

Geo-Social Networks, Big Data, Extended Jaccard Similarity, Buffalo Optimization.

Introduction

The rapid development of Geo-Social Networks (GeoSNs) facilitates people to create their contents publically related to any social events with the geographic location information. Geo-social network data grants complete information on recent trends in humans, their activities, their livelihood fashion, events and occurrences, disasters, current health infection and much more information about the locations. There are a lot of other works that have been prepared using geo-social networks using location information. However, the entire work is done at a limited level and ignoring the dimensions of data generated. In this paper, a novel clustering technique is introduced to analyze the big size of data generated in the Geo-social network.

An Artificial Bee Colony (ABC) approach was introduced in (S. Sudhakar Ilango, 2019) to optimize the best cluster for reducing the execution time and minimizing the error rate. However, the accuracy of finding the optimal cluster was not improved. In (Dingming Wu, May 2018), a Density-based Clustering Places in Geo-Social Networks (DCPGS) was designed to determine the optimal cluster. But the designed DCPGS failed to find the quality of clusters.

An Extended Adaptive Density Peaks (EADP) clustering technique was introduced in (Mingli Xu, 2019) for identifying the overlapping community. The designed clustering technique failed to apply large scale social networks. A Non-Triangle Inequality (TI) clustering technique was introduced in (Sanjit Kumar Saha, 2020) but it failed to minimize the complexity of the algorithm. An Agglomerative Spectral Clustering technique was presented in (Ulzii-Utas Narantsatsralt, November 2017) for enhancing community detection in the social network. However, the performance of the error rate was not minimized.

The fuzzy logic concept was applied in density-based clustering in (Goldina Ghosh, 2016) for analyzing big social data. But the designed clustering method failed to apply the meta-heuristic efforts for finding the optimum cluster. A Clustering Algorithm based Structure Similarity (CASS) was introduced in (Jungrim Kim, 2018) for evaluating huge volume of social data by means of applying an optimization technique. Though the designed algorithm reduces the memory usage and execution time, the accuracy of clustering was not improved.

A Social Network Analysis Platform (SNAP) approach was designed in (Victor Chang, 2018) to perform Big Data Analytics. But the clustering technique was not applied to process the large volume of data for minimizing the memory usage. An incremental K-means clustering algorithm based on density was designed in (Weijia Lu, 2019) for achieving better clustering accuracy and minimal error. The designed algorithm failed to test the large data sets.

Clustering a temporal network was presented in (Joseph Crawford, 2018) depending on topological similarity. The designed clustering technique increases the accuracy and minimizes the running time but the memory consumption was not reduced.

Major Contribution

Various methods have been reviewed for clustering big data, but it did not provide an effective mechanism in terms of accuracy and optimal selection of clusters to handle the big data. The contributions of the proposed EJIBODC technique are shortened as follows:

- A novel EJIBODC technique is introduced to obtain optimal geo-social data clusters by partitioning the total dataset into different groups.
- The EJIBODC technique uses the buffalo optimization technique to find optimal cluster centroid in the search space for accurately grouping the geo-social data. On the contrary to existing distance-based clustering technique, the EJIBODC technique uses the extended Jaccard similarity index for fitness computation.
- Based on the fitness evaluation, the proposed optimization technique carries out global search methods by attempting to balance the exploration and exploitation process.
- Through extensive experimental results, the performance of the proposed EJIBODC technique is evaluated with conventional clustering methods. The results validate that the proposed EJIBODC technique outperforms for selecting optimal clusters to group the big social data with high accuracy and minimum error as well as time consumption.

The paper is divided into six different sections. Section two demonstrates the related work in the big data clustering environment and also reviews the issues of various techniques. In section three, a novel EJIBODC is described for the clustering of big geo-spatial data. Section four illustrates the experimental scenario and parameter description. Followed by this, the comparative performance analyses of proposed and other two existing algorithms are presented in section five. Section six describes the conclusion of the proposed EJIBODC.

Related Works

A local expansion technique by means of density-based clustering was introduced in (Xiaofeng Wang, 2017) to detect the intrinsic network communities, but the optimization was not applied to increase the clustering accuracy and minimize the error rate.

A Gaussian Pigeon-oriented Graph Clustering Algorithm was designed in (Yang Sun, 2019) for achieving the better clustering accuracy and lesser mean square error. However, a large volume of information and external information was not considered to perform the clustering in social network. Gaussian Process Regression (GPR) model was employed in (Gunasekaran Manogaran, May 2017) for processing big data based on cluster computing method. But, the clustering accuracy was not improved using the GPR model.

CorClustST–Correlation-based Clustering using big Spatio-Temporal information was introduced in (Marc Husch, April 2018). However, the clustering solution was not optimized for a particular quality condition since it has high complexity. A hybrid clustering algorithm was designed in (Sunil Kumar, 2019) to handle the large volume of data. But the designed hybrid clustering algorithm has the limitation of taking more execution time.

A Dynamic frequency-based parallel k-bat algorithm was designed in (Ashish Kumar Tripathi, 2018) to cluster the large scale data. However, the performance of clustering accuracy was not improved. A novel clustering technique named Eb & D was developed in (Xingqin Qi, 2017) for increasing the clustering accuracy of large scale networks with minimum error rate. But the designed method failed to predict the number of clusters. A new Intelligent Weighting k-means Clustering (IWKC) algorithm was developed in (Qian Tao, Chunqin Gu, 2020) by means of swarm intelligence. However, the algorithm failed to handle the clustering of high-dimensional multi-view data in the heterogeneous data sets. A fast clustering approach was presented in (Un Wu, June 2018) based on the projection similarity. The designed clustering method provides a higher clustering accuracy for small datasets.

A fuzzy consensus clustering (FCC) method was introduced in (Junjie Wu, December 2017) for communication detection by grouping the big data. The designed method failed to achieve accurate clustering results with minimum error.

Proposal Methodology

With the initiation of Big Data era, multi-source geo-tagged data offers a novel perception and data source for urban spatial analysis. In general, Geo-social community detection aims

at finding the users that concern both social and spatial constraints. Clustering is commonly used as a method for community detection in the Geo-social network. Based on this motivation, a novel EJIBODC technique is introduced. The proposed EJIBODC technique collects the input big geo-social data from the dataset. On the contrary to existing clustering technique, the proposed clustering uses the extended Jaccard Indexive Buffalo Optimization technique for increasing the clustering accuracy with minimum time.

The optimization algorithm is stimulated by the activities of African buffalos in the large African forests. The designed optimization algorithm is a population-based metaheuristic algorithm to track the location of the best buffalo in each iteration. In contrast to the basic African buffalos, the proposed optimization technique uses the extended Jaccard index for calculating the fitness to find the global optimum solution for the population. Extended Jaccard Indexive Optimization is an attempt to develop a robust and efficient algorithm that ensures the fast convergence rate by utilizing a few learning parameters. The block diagram of the EJIBODC technique based data clustering is illustrated in figure 1.

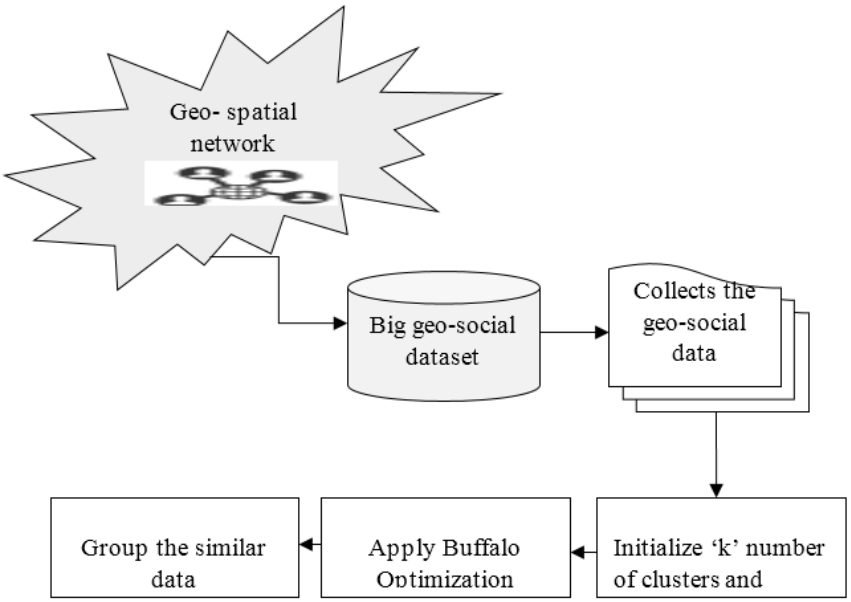

Figure 1 Architecture of proposed EJIBODC Technique

Figure 1 illustrates the architecture of the proposal EJIBODC technique to group the similar data. Large volumes of big data are generated from the geo-social network. Then, these data are stored in the dataset. In order to perform big data analysis, the big geo-social data are collected from the dataset.

$$d_i = \{d_1, d_2, d_3, \ldots . d_n\} \in D \quad (1)$$

Where, d_i denotes big geo-social data of $d_1, d_2, d_3, \ldots . d_n$ collected from the dataset 'D'. After collecting the big data, the clustering process is carried out to partition the total dataset into dissimilar groups for detecting the frequently visited location of the users in the social network.

The proposed technique starts to perform the clustering process where 'k' number of clusters $S_1, S_2, S_3, \ldots . S_k$ and the centroid $v_1, v_2, v_3, \ldots . v_k$ are initialized. In order to obtain the accurate clustering, the Extended Jaccard Indexive Optimization technique is applied.

By applying the proposed optimization technique, population of Buffalo's are initialized with a random location in the search space. Here the Buffalo's are related to the cluster centroid $v_1, v_2, v_3, \ldots . v_k$. After the initialization, the fitness of each Buffalo is calculated to find the fittest cluster for assigning the data. Based on the Extended Jaccard Similarity index, the fitness is computed between the data and the cluster centroid. The similarity is mathematically expressed as follows,

$$\varphi = \frac{BD_i \cap v_k}{\sum BD_i + \sum v_k - BD_i \cap v_k} \quad (2)$$

Where, φ indicates Extended Jaccard Similarity coefficient, BD_i denotes geo-social data, v_k indicates cluster centroid, the intersection symbol '$\bigcap$' designates mutual independence between the data and the cluster centroid which are statistically dependent, $\sum BD_i$ is the sum of BD_i score and $\sum v_k$ is the sum of $v_k$score. The Extended Jaccard Similarity coefficient provides the output values from 0 to 1 $[0 \leq \varphi \leq 1]$. The proposed technique finds the fittest cluster centroid to group the data based on the gradient ascent function to find the maximum of the given solution.

$$F = \arg\max \varphi \quad (3)$$

From (3), F denotes fitness function and $\arg max$ denotes argument of a maximum function to find a higher similarity. The higher similarity data is assigned to that particular cluster centroid. Based on the fitness value, the two processes namely exploration and exploitation are performed.

$$M_k\ (t+1) = M_k + d_{p1}\ (bf - W_k) + d_{p2}\ (bp.k - W_k) \quad (4)$$

Where, $M_k\ (t+1)$ denotes an update buffalos' exploitation of the 'k'th buffalo, W_k denotes an exploration of the buffalos', d_{p1} and d_{p2} are the learning parameters that set the values from 0.1 to 0.6, bf indicates herd's best fitness of buffalo's and bp refers to the individual

buffalo's best location. Followed by this, using the following equation, update the location of buffalos.

$$W_k(t+1) = \frac{(W_k + M_k)}{s} \quad (5)$$

Where, $W_k(t+1)$ denotes an update of the location of buffalos and s denotes parameter value that is set as $\pm$ 0.5. If the stopping criteria are not met, then go back for updating the buffalos, else stop the process. Based on the above mentioned process, the optimum cluster centroid is identified to group the data into the cluster. In this way, all the data are grouped into clusters with higher accuracy. The flow process of the Extended Jaccard Indexive Buffalo Optimized Data Clustering is given below,

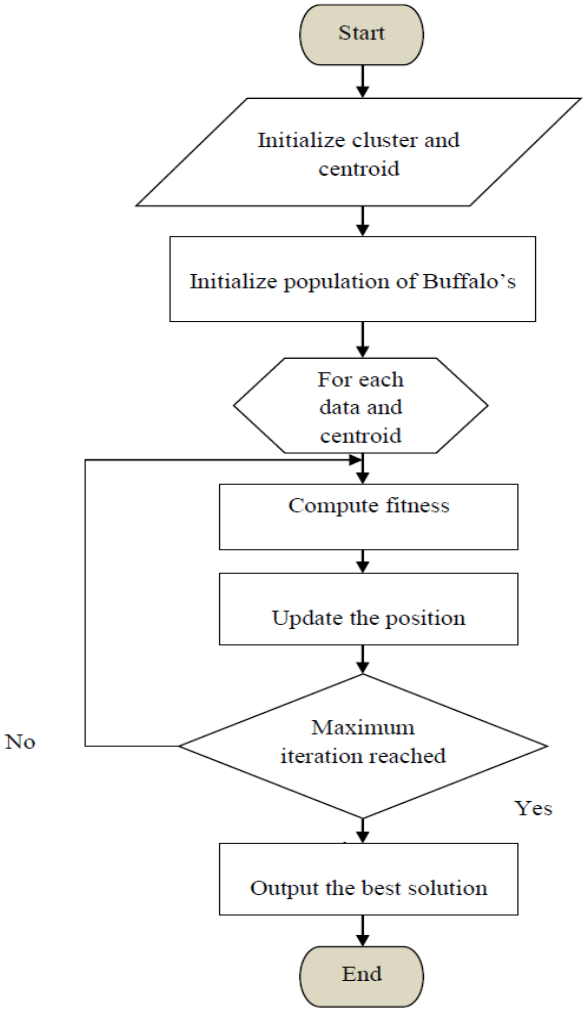

Figure 3 Flow diagram of Extended Jaccard Indexive Buffalo Optimized Data Clustering

Figure 3 illustrates the flow diagram of Extended Jaccard Indexive Buffalo Optimized Clustering of big data. The algorithmic process of clustering is described as follows,

Algorithm 1 Extended Jaccard Indexive Buffalo Optimized Clustering

Input: Number of geo-spatial data $d_1, d_2, d_3, \ldots . d_n$
Output: Increase clustering accuracy

Begin
Initialize 'k' number of clusters and centroid
Randomly initialize a population of the centroid $v_1, v_2, v_3, \ldots . v_k$
Initialize learning parameters d_{p1} and d_{p2}
For each data d_i
For each centroid 'v_k'
Measure similarity φ
Compute fitness $F = \arg\max \varphi$
While (t < max_ iteration)
if $\left(F(v_i) < F\left(v_j\right)\right)$ **then**
Update buffalos' exploitation $M_k\ (t+1)$
Update the location of buffalos $W_k(t+1)$
End if
end for
end for
t= t+1
end while
Obtain the best solution
End

The step by step process of Extended Jaccard Indexive Buffalo Optimized Clustering technique is described to improve the clustering accuracy with minimum time. Initially, the number of clusters and centroid are initialized. Then, randomly initialize the population of the centroid (i.e. buffalos) in search space. For each data and cluster centroid, the similarity is measured for evaluating the fitness. If the fitness of one buffalo is greater than the other, the position of buffalos' gets updated and it finds the optimal. This process is repeated until the maximum iteration is reached. These processes increase the clustering accuracy and minimize the error rate.

Experimental Scenario and Parameter Description

In this section, proposed EJIBODC technique and existing methods namely the ABC approach (S. Sudhakar Ilango, 2019) and DCPGS (Dingming Wu, May 2018) are illustrated and are implemented using Java Language. The Weeplaces Dataset (https://www.yongliu.org/datasets/) is used to conduct the experimental evaluation and datas are collected from the popular location-based social network services e.g., Facebook, Foursquare and Gowalla. This dataset comprises of 7,658,368 check-ins generated by 15,799 users over 971,309 locations. The check-in includes the information about the user, check-in-time, latitude, and longitude and location id. This information is used to perform

the clustering process. In order to conduct the fair simulation, the number of big data is taken from 1000 to 10000. Totally ten runs are performed for each technique. The performance of the proposed EJIBODC technique is evaluated using the following metrics.

- Clustering accuracy
- Error rate
- Clustering time
- Space complexity

Clustering Accuracy

The clustering accuracy is a significant metric to find the performance of the algorithm. It is referred as the ratio of the data that are correctly grouped to the total number of big data. The mathematical formula for evaluating the clustering accuracy is expressed as follows,

$$CA = \left[\frac{\varepsilon_{AC}}{\varepsilon}\right] * 100 \quad (6)$$

Where CA signifies clustering accuracy, 'ε_{AC}' denotes the number of data that are correctly grouped and 'ε' denotes the total number of data taken for conducting the experiment. The accuracy of big geo-social data is measured in terms of percentage (%).

Error Rate

The second important parameter is the Error Rate. It is determined as the ratio of number of data wrongly grouped into the cluster to the total number of data taken as input. The formula for measuring the error rate is expressed as follows,

$$ER = \left[\frac{\varepsilon_{WC}}{\varepsilon}\right] * 100 \quad (7)$$

Where ER specifies the error rate and ε_{WC} denotes the number of data wrongly grouped into the cluster from the total number of clusters which is denoted by ε. The error rate of big data clustering is measured in percentage (%).

Clustering Time

The third important parameter is the clustering time. It is used to find the time taken by the algorithm to group the data into dissimilar clusters. The clustering time is calculated using the given formula,

$$CT = \varepsilon * t[CS] \quad (8)$$

From (8), the clustering time 'CT' is measured using the number of data 'ε' that is considered for experimentation purpose and the time for grouping single data $t[CS]$ is measured in terms of milliseconds (ms).

Space Complexity

The final performance metric is the space complexity which measured as an amount of memory space consumed by the algorithm to store the clustered data. Therefore, the space complexity is mathematically written as follows.

$$com_{space} = \varepsilon * Mem[SD] \quad (9)$$

From (5), '$Mem[SD]$' symbolizes memory utilized to store a single geo-social data and 'ε' refers to the total number of geo-social data. The space complexity of the different clustering algorithms is measured in terms of Megabytes (MB).

Comparative Analysis

In this section, performances of different metrics are analyzed using the EJIBODC technique with two existing clustering algorithms namely, ABC approach (S. Sudhakar Ilango, 2019) and DCPGS (Dingming Wu, May 2018). Initially, the clustering accuracy is measured by means of correctly grouped geospatial data with the total number of data considered as input.

Table I Clustering Accuracy

Number of geo-social data	Clustering Accuracy (%)		
	ABC approach	DCPGS	EJIBODC
1000	83	79	88
2000	86	83	90
3000	85	82	91
4000	86	81	93
5000	85	81	92
6000	84	80	90
7000	83	79	89
8000	85	77	90
9000	84	76	89
10000	83	75	87

Table I reports the experimental outcome of clustering accuracy relating to the number of geo-social data. For better experimentation, ten runs are measured with the input data taken

in the count from 1000 to 10000. Table I concludes that the EJIBODC technique outperforms the existing methods (S. Sudhakar Ilango, 2019) and (Dingming Wu, May 2018) for all the runs. The performance results are indicated as shown in figure 4.

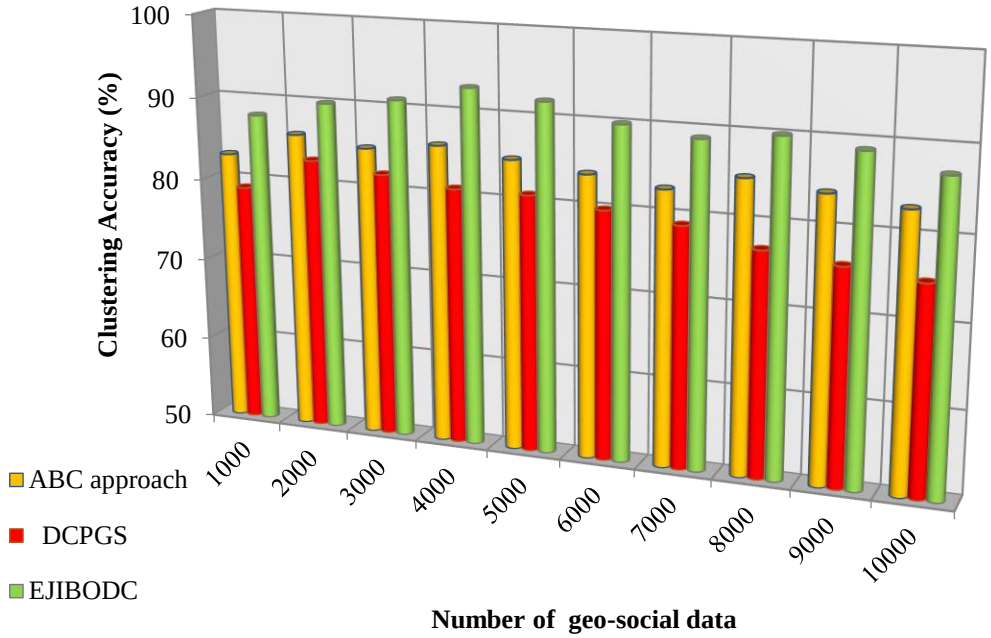

Figure 4 Performance results of clustering accuracy

Figure 4 depicts the comparison of the clustering accuracy using the ten best results obtained. The graphical results show that the 'x' axis represents 10000 different numbers of data considered for experimentation at a different time interval and the 'y' axis represents the clustering accuracy. To make the results more reliable, each clustering method runs for ten times with different input data. By comparing the EJIBODC technique with existing algorithms, it is inferred that the clustering accuracy is better. The reason behind the improvement is the application of the Extended Jaccard Indexive Buffalo Optimized Clustering technique. By applying this proposed clustering algorithm, the fitness Extended Jaccard similarity index is applied to group similar data into a suitable cluster. By applying the similarity index, the fittest cluster centroid is correctly identified for grouping the data resulting it in improving the accuracy. The observed results found that the EJIBODC technique when compared to the clustering methods have high clustering accuracy. The average clustering accuracy of the EJIBODC technique is increased by 7% when compared to the ABC approach (S. Sudhakar Ilango, 2019) and 13% when compared to DCPGS (Dingming Wu, May 2018).

Table II Error Rate

Number of geo-social data	Error Rate (%)		
	ABC approach	DCPGS	EJIBODC
1000	17	26	12
2000	14	22	10
3000	15	21	9
4000	14	24	7
5000	15	25	8
6000	16	25	10
7000	17	27	11
8000	15	28	10
9000	16	29	11
10000	17	30	13

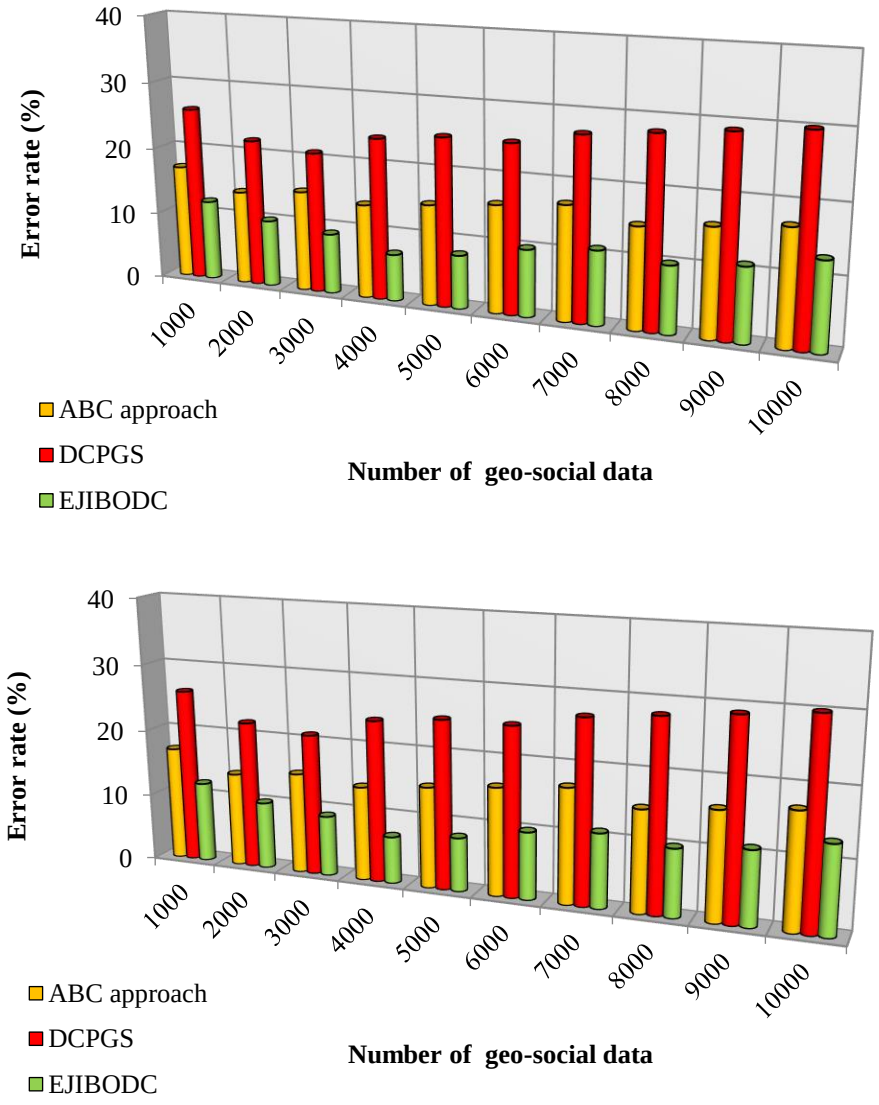

Figure 5 Performance results of error rate

Table II and figure 5 illustrates the experimental outcomes of error rate using three different clustering algorithms namely the proposed EJIBODC and the existing ABC approach (S. Sudhakar Ilango, 2019) and DCPGS (Dingming Wu, May 2018). The graphical illustration shows that the EJIBODC technique achieves better outcomes in terms of error rate. This is because; EJIBODC technique uses the similarity measure in the fitness calculation. Based on similarity value, an optimal fittest cluster is discovered by applying the gradient ascent function. Then the data is grouped into a suitable cluster resulting it in improving the accuracy and minimizing the error rate. By applying the input of 1000 data, the EJIBODC technique wrongly grouped 120 data and the error rate is 12%. By applying the ABC approach (S. Sudhakar Ilango, 2019) and DCPGS (Dingming Wu, May 2018), the number of data incorrectly grouped is 170 and 260 and their clustering accuracy is 17% and 26% respectively with a similar count of input. Thus, the overall error rate of the EJIBODC technique is reduced by 36% when compared to ABC technique (S. Sudhakar Ilango, 2019) and 60% when compared to DCPGS (Dingming Wu, May 2018).

Table III Clustering Time

Number of geo-social data	Clustering Time (ms)		
	ABC approach	DCPGS	EJIBODC
1000	28	37	23
2000	30	40	26
3000	36	48	30
4000	40	52	32
5000	43	48	35
6000	48	54	38
7000	53	57	42
8000	56	61	46
9000	58	65	49
10000	63	70	52

Table III details the experimental results of the time taken for grouping the number of geo-social data into a number of clusters. As revealed in table III, while increasing the number of geo-social data, the time taken to group the data also increases. This demonstrates that the clustering time is directly related to the data. However, the clustering time is found to be decreased using the EJIBODC technique than the existing clustering methods. This is the reason for calculating the fitness between the data and cluster centroid using buffalo optimization. Initially, the number of clusters and the centroid are initialized. In order to group similar data into the cluster, the extended Jaccard similarity is measured. This process helps to minimize the time to group similar data into the cluster. For example, 1000 data are considered to evaluate the clustering time. The proposed EJIBODC technique consumes $23ms$ for grouping the geo-social data while the time consumption of the ABC

approach (S. Sudhakar Ilango, 2019) and DCPGS (Dingming Wu, May 2018) are $28ms$ and $37ms$ respectively. Similarly, different runs are carried by increasing the input data upto 10000. Therefore, the overall time consumption of the EJIBODC technique is found to be reduced by 18% and 31% when compared to existing approaches (S. Sudhakar Ilango, 2019) and (Dingming Wu, May 2018) respectively.

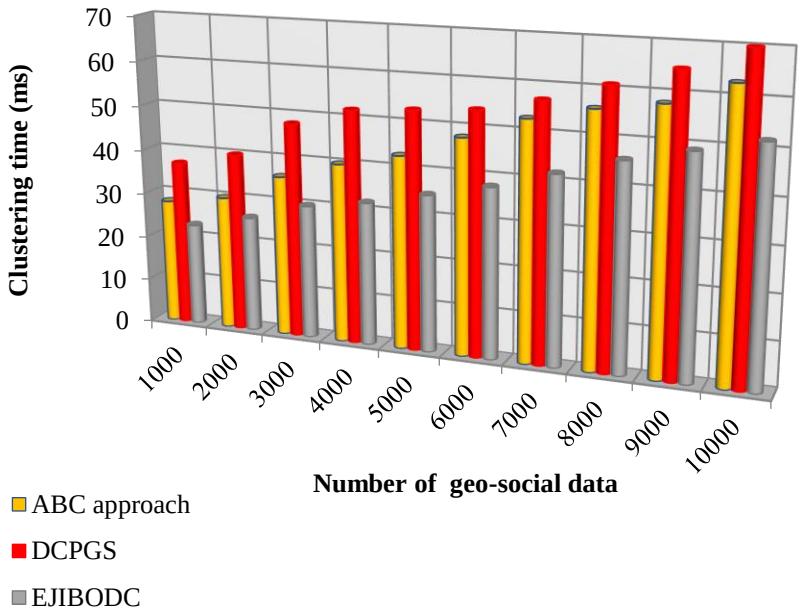

Figure 6 Performance results of clustering time

The performance results of clustering time versus a number of geo-social data are shown in figure 6. Compared to existing algorithms, the EJIBODC technique gets much better performance on clustering time.

Table IV Space Complexity

Number of geo-social data	Space Complexity (MB)		
	ABC approach	DCPGS	EJIBODC
1000	25	28	21
2000	28	34	24
3000	33	36	30
4000	36	39	32
5000	40	44	35
6000	42	46	39
7000	48	52	42
8000	50	53	46
9000	53	58	49
10000	56	61	52

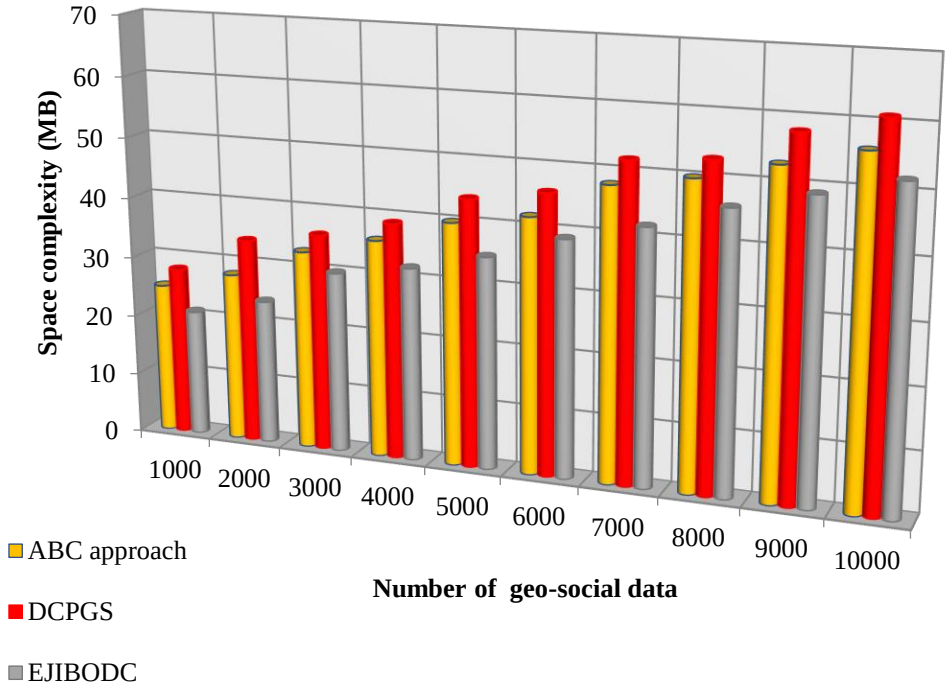

Figure 7 Performance results of space complexity

Table IV and figure 7 displays the effect of space complexity versus number of geo-social data in the range of 1000 to 10000. As portrayed in figure 7, the space complexity gets increased while increasing the number of data as given as input. But the impact of memory consumption is minimal using the EJIBODC technique. The reason for this proposed clustering technique minimizes the memory space while storing the geo-social data. Let us consider 1000 data, the space complexity of the EJIBODC technique is $21MB$ and the results of storage space of the other two methods namely, ABC approach (S. Sudhakar Ilango, 2019) and DCPGS (Dingming Wu, May 2018) are $25MB$ and $28MB$ respectively. The overall space complexity of the EJIBODC technique is compared with the existing results. The average comparison results prove that the space complexity is considerably reduced using the EJIBODC technique by 11% when compared to ABC approach (S. Sudhakar Ilango, 2019) and 19% when compared to DCPGS (Dingming Wu, May 2018).

Conclusion

As the utilization of Geo-Social networks has been increasing rapidly, it becomes quite difficult for clustering large dataset using a conventional clustering algorithm. In this paper, the EJIBODC technique is designed for the problem of solving the big data clustering in Geo-Social Networks. The designed EJIBODC technique provides the fastest completion

time of clustering and making it more efficient for all types of data. The extended Jaccard index is a similarity coefficient employed in buffalo optimization to find the fittest cluster for grouping the data. Therefore, the input geo social data are correctly grouped to reduce time consumption. To demonstrate the effectiveness, the proposed EJIBODC technique has experimented with the geo-social dataset. According to the evaluation results, it is observed that in the clustering stage, the proposed EJIBODC technique efficiently grouped the big data and hence it has a better clustering accuracy. Similarly, the EJIBODC technique also reduces the clustering time, error rate and space complexity when compared to conventional clustering algorithms.

References

Ilango, S.S., Vimal, S., Kaliappan, M., & Subbulakshmi, P. (2019). Optimization using artificial bee colony based clustering approach for big data. *Cluster Computing*, *22*(5), 12169-12177.

Wu, D., Shi, J., & Mamoulis, N. (2017). Density-based place clustering using geo-social network data. *IEEE Transactions on Knowledge and Data Engineering*, *30*(5), 838-851.

Xu, M., Li, Y., Li, R., Zou, F., & Gu, X. (2019). EADP: An extended adaptive density peaks clustering for overlapping community detection in social networks. *Neurocomputing*, *337*, 287-302.

Saha, S. K., & Schmitt, I. (2020). Non-TI Clustering in the Context of Social Networks. *Procedia Computer Science*, *170*, 1186-1191.

Narantsatsralt, U.U., & Kang, S. (2017). Social network community detection using agglomerative spectral clustering. *Complexity*, 1-10.

Ghosh, G., Banerjee, S., & Yen, N.Y. (2016). State transition in communication under social network: An analysis using fuzzy logic and Density Based Clustering towards big data paradigm. *Future generation computer systems*, *65*, 207-220.

Kim, J., Shin, M., Kim, J., Park, C., Lee, S., Woo, J., & Park, S. (2018). CASS: A distributed network clustering algorithm based on structure similarity for large-scale network. *PloS one*, *13*(10).

Chang, V. (2018). A proposed social network analysis platform for big data analytics. *Technological Forecasting and Social Change*, *130*, 57-68.

Lu, W. (2020). Improved K-means clustering algorithm for big data mining under Hadoop parallel framework. *Journal of Grid Computing*, *18*(2), 239-250.

Crawford, J., & Milenković, T. (2018). ClueNet: Clustering a temporal network based on topological similarity rather than denseness. *PloS one*, *13*(5).

Wang, X., Liu, G., Li, J., & Nees, J. P. (2017). Locating structural centers: A density-based clustering method for community detection. *PloS one*, *12*(1).

Sun, Y., Yin, S., Li, H., Teng, L., & Karim, S. (2019). GPOGC: Gaussian pigeon-oriented graph clustering algorithm for social networks cluster. *IEEE Access*, *7*, 99254-99262.

Manogaran, G., & Lopez, D. (2017). A Gaussian process based big data processing framework in cluster computing environment. *IEEE Transactions on Information Theory,* 63(5), 2954 – 2974.

Hüsch, M., Schyska, B.U., & von Bremen, L. (2020). CorClustST—Correlation-based clustering of big spatio-temporal datasets. *Future Generation Computer Systems, 110,* 610-619.

Kumar, S., & Singh, M. (2019). A novel clustering technique for efficient clustering of big data in Hadoop Ecosystem. *Big Data Mining and Analytics, 2*(4), 240-247.

Tripathi, A.K., Sharma, K., & Bala, M. (2018). Dynamic frequency based parallel k-bat algorithm for massive data clustering (DFBPKBA). *International Journal of System Assurance Engineering and Management, 9*(4), 866-874.

Qi, X., Song, H., Wu, J., Fuller, E., Luo, R., & Zhang, C.Q. (2017). Eb&D: A new clustering approach for signed social networks based on both edge-betweenness centrality and density of subgraphs. *Physica A: Statistical Mechanics and its Applications, 482,* 147-157.

Tao, Q., Gu, C., Wang, Z., & Jiang, D. (2020). An intelligent clustering algorithm for high-dimensional multiview data in big data applications. *Neurocomputing, 393,* 234-244.

Wu, Y., He, Z., Lin, H., Zheng, Y., Zhang, J., & Xu, D. (2019). A fast projection-based algorithm for clustering big data. *Interdisciplinary Sciences: Computational Life Sciences, 11*(3), 360-366.

Wu, J., Wu, Z., Cao, J., Liu, H., Chen, G., & Zhang, Y. (2017). Fuzzy consensus clustering with applications on big data. *IEEE Transactions on Fuzzy Systems, 25*(6), 1430-1445.

Rostami, R.R., Karbasi, S. (2020). Detecting fake accounts on twitter social network using multi-objective hybrid feature selection approach. *Webology, 17*(1), 1-18.