Abstract - Keywords: Big social data, Stochastic Extreme Gradient Boost clustering, adaptive Fuzzy K-Means Clustering, Minkowski distance

Article · June 2021

DOI: 10.1109/ICACITE51222.2021.9404574

2 authors:

M. Anoop
Alpha Arts and Science College
9 PUBLICATIONS **1** CITATION
SEE PROFILE

Sripriya P.
Vels University
33 PUBLICATIONS **44** CITATIONS
SEE PROFILE

2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)
Department of Electrical & Electronics Engineering, Galgotias College of Engineering and Technology, Gr. Noida, India

Ensembled Adaptive Fuzzy K-Means With Stochastic Extreme Gradient Boost Big Data Clustering on Geo-Social Networks

M. Anoop, P. Sripriya
Research Scholar, Department of Computer Applications, Vels Institute of Science, Technology & Advanced Studies, Chennai, Tamil Nadu, India.profanoopcs@rediffmail.com
Professor, Department of Computer Applications, Vels Institute of Science, Technology & Advanced Studies, Chennai, Tamil Nadu, India. sripriya.phd@gmail.com

***Abstract* – The szwfast growth of Geo-Social Networks (GeoSNs) offers a novel and unlikely form of data. Handlers of GeoSNs arrest their geographic locations and segment them with other users to form a community. Public detection is an efficient tool for analyzing the social relationship between users. The existing algorithm typically focuses on clustering the data but the number of expected clusters with higher accuracy is the major challenging one. In order to improve the clustering accuracy, Ensembled Adaptive Fuzzy K-Means with Stochastic Extreme Gradient Boost data clustering (EAFK-SEGBBDC) technique is introduced. The main aim of the EAFK-SEGBBDC technique is to analyze the geosocial network data and to form the cluster with higher accuracy and minimal error rate. In the EAFK-SEGBBDC technique, Stochastic Extreme Gradient Boost Cluster is an ensemble technique to construct a strong cluster by combining the number of weak learners as adaptive Fuzzy K-Means Cluster. The input geosocial network data are collected from the dataset and it is given to the adaptive Fuzzy K-Means Cluster for grouping the similar data. Adaptive Fuzzy C-Means Clustering model partitions the number of input data into different groups. Minkowski distance is calculated between the cluster centroid and the geo-social network data for grouping similar data to form the cluster. Finally, entirely the weak learners are combined to obtain the final strong clustering results with higher clustering accuracy. The observed results indicate that the proposed EAFK-SEGBBDC technique provides better performance in terms of achieving higher accuracy and lesser time than the conventional methods.**

Keywords: Big social data, Stochastic Extreme Gradient Boost clustering, adaptive Fuzzy K-Means Clustering, Minkowski distance

I. INTRODUCTION

A new intelligent weighting k-means clustering (IWKM) algorithm was introduced in [1]. The designed algorithm was not effectual in the clustering of high-dimensional data with minimum complexity and higher accuracy. A Density-based Clustering Places in Geo-Social Networks (DCPGS) was developed in [2] for partitioning the network based on Spatio-temporal information and the social interaction among the users. However, the algorithm failed to achieve a higher quality of clustering results.

[3] For community detection,a density algorithm. However, the higher accuracy was not achieved with lesser time consumption. The graph clustering technique was introduced in [4] depends on collaborative similarity for community detection. However, it takes more time to validate the large volume of data. In [5], a novel hybrid fuzzy c-means clustering algorithm was integrated with dual soft computing techniques for community detection. However, the error rate was not minimized.

An incremental K-means clustering algorithm was introduced in [6] depends on density to improve the clustering accuracy using the Hadoop platform. However, the time complexity of the data clustering was not minimized. A Fast and efficient clustering algorithm was introduced in [7] for clustering the big volume of data efficiently to detect the communities. However the clustering algorithm rises the accuracy and scalability, space and time complexity analysis was not performed.

[8] A theory based on the large Spatio-temporal data. The clustering approach was efficient for big data processing but the accurate results were not obtained with minimum complexity. A novel clustering algorithm known as hybrid clustering was introduced in [9] to design scalable clustering results with higher precision. However, complexity analysis was not performed. A Random Sample Partition (RSP) method was introduced in [10] based on the distributed data to partition a big data set. The designed method considerably decreases the computational burden of big data but the higher accuracy was not obtained.

II. PROPOSAL METHODOLOGY

A novel proposed model is developed with the aim to efficiently perform the clustering process with big geosocial data. Big data analytics is a process of estimating the vast size of geosocial data to extract useful patterns. With the extensive development in a geosocial network, big data assessment is effectively used for community detection. Due to the large volume of data generation, the learning of whole attributes is typically not possible and not accurate since the big dataset consists of more attributes and a number of instances. Therefore, the dimensionality of the big geosocial data needs to minimize for accurate processing with lesser time consumption. The dimensionality of the dataset is reduced by performing the clustering of the big geosocial data. Hence the process of the clustering technique reduces the time consumption as well as space complexity. Based on the above motivation, the proposed EAFK-SEGBBDC technique is designed with the application of the ensemble clustering technique.

978-1-7281-7741-0/21/$31.00 ©2021 IEEE

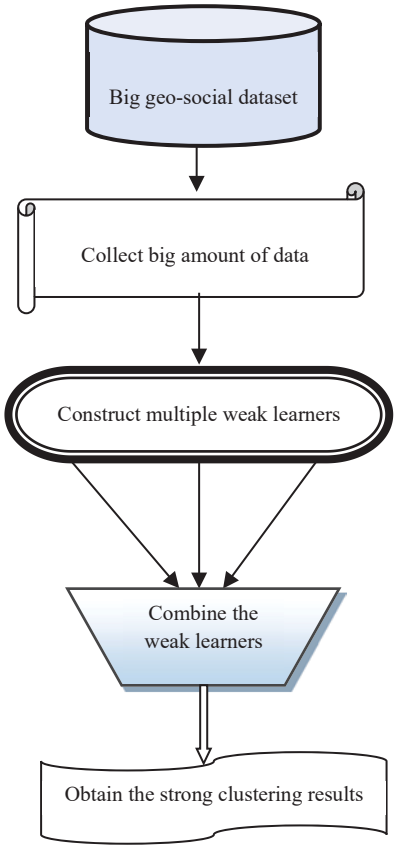

Figure 1 flow process of the EAFK-SEGBBDC technique

Figure 1 demonstrates an architecture diagram of the future EAFK-SEGBBDC technique to obtain the final clustering results with higher accuracy and lesser time consumption. A big volume of geosocial data is collected from the big dataset.

Stochastic Extreme Gradient Boost ensemble clustering technique uses the adaptive fuzzy k means clustering technique is a weak learner. The adaptive fuzzy k means clustering algorithm focuses on reducing the value of an objective function. Here, the objective function calculates the quality of the partitioning of the dataset into 'k' clusters. Clustering is the process of discovering the grouping of data samples based on their distance similarity between the cluster centers and their data samples, along with the membership degree.

By applying the adaptive fuzzy k means clustering, 'k' number of clusters $\beta_1, \beta_2, \beta_3, \ldots . \beta_k$ and the centroids $\varphi_1, \varphi_2, \varphi_3, \ldots . \varphi_k$ are initialized. The adaptive fuzzy k means clustering technique helps to reduce the objective function i.e. distance between the cluster centroid and the data. The objective function is to obtain as given below,

$$F = \arg\min \ \mu_{ij} \ D_{ij} \ (1)$$

Where, F denotes an objective function, $\arg min$ denotes an argument of the minimum function, μ_{ij} denotes a membership degree to which the data d belongs to a cluster center φ_j, D_{ij} indicates the distance between the i^{th} data and the j^{th} cluster centroid. The Fuzzy membership is measured based on the distance measure.

$$\mu_{ij} = \frac{1}{\sum_{i=1}^{n} \sum_{j=1}^{c} \left(\frac{D_{ij}}{D_{ik}}\right)^{\left(\frac{2}{n-1}\right)}} \quad (2)$$

In (2),μ_{ij} represents the fuzzy membership function helps to identify the member of that particular group, in a range of values [0 1], D_{ij} means a distance between the i^{th} data and j^{th} cluster center, D_{ik} represents a distance between the i^{th} data and k^{th} cluster center, n denotes a fuzzifier. The Minkowski distance between the data and the cluster center is measured as given below,

$$D_{ij} = \left(\left|d_i - \varphi_j\right|^q\right)^{1/q} \quad (3)$$

Where, D_{ij} denotes a distance between the i^{th} data 'd_i' and j^{th} cluster center φ_j , q denotes Minkowski distance of order $q = 1$.

$$D_{ik} = (|d_i - \varphi_k|^q)^{1/q} \quad (4)$$

Where, D_{ik} indicates the distance between the i^{th} data and k^{th} cluster center. The data which is closer to the cluster centroid is grouped. Likewise, all the data are grouped into a particular cluster until the convergence is met (i.e. all the data moved into the cluster). The flow process of the adaptive fuzzy k means clustering technique is shown in figure 2.

From the above definitions, the essentials can fit into more than one cluster with dissimilar degrees of membership. The entire "member- ship" of an element is regularized to 1 and a single cluster cannot contain all data points.

The weak clustering results have some training error and its lack of providing accurate results. Therefore, the boosting ensemble technique combines all the weak learners and to make a strong. The ensemble of the weak learners are expressed as given below,

$$E = \sum_{i=1}^{b} w_i \quad (5)$$

Where 'E' symbolizes the strong clustering result that provides strong clustering results through the linear combination of the weak learners $\sum_{i=1}^{b} w_i$. Then the weights are initialized to each weak learner is expressed as follows,

$$E = \sum_{i=1}^{b} w_i * \phi \quad (6)$$

Where 'ϕ' represents weights is the integer number to validate the clustering performance of weak classifiers. But the weak classifier has some training loss which is expressed as given below,

$$E = \sum_{i=1}^{b} w_i + loss_t \quad (7)$$

From (7), $loss_t$ designates a training loss of weak learner 'w_i'. From (), the squared error loss is estimated as given below,

$$loss_t = (E_a - E_o)^2 \quad (8)$$

Where, 'E_a' denotes an actual result, E_o indicate the observed results. Followed by, the weights get restructured according to the weight value. The proposed ensemble technique uses stochastic gradient descent step-size

task to find the weak learner results which have minimum training loss than the other.

$$f\ (x) = arg\ \min\ [\ loss_t\ \ (w_i)]\quad (9)$$

From (9), $f\ (x)$ symbolizes a gradient descent step-size function, $arg\ min$ indicates the argument of the minimum function, $loss_t(w_i)$ specifies a training loss of weak learners. Accordingly, the weak learner which has minimum training loss is taken as final clustering results. This helps to enhance the clustering accuracy and reduces the error rate.

\\ Algorithm 1: Ensembled Fuzzy K-Means with Stochastic Extreme Gradient Boost data clustering

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

Output: Improve Clustering accuracy

Begin

1. Collect number of geo-spatial data $d_1, d_2, d_3, \ldots . d_n$
2. **for each** data 'd_i'
3. Construct a set of weak learners $w_1, w_2, w_3 \ldots ., w_b$
4. Initialize 'k' number of clusters
5. **for each** cluster 'k'
6. Initialize 'clusters centroid 'φ_j
7. **end for**
8. **for each** data sample 'd_i'
9. **for each** cluster centroid 'φ_j'
10. Find objective function $F = \arg\min\ \mu_{ij}\ D_{ij}$
11. Measure membership grade 'μ_{ij}'
12. Group the d_i to cluster based on the minimum distance
13. **End for**
14. **End for**
15. **End for**
16. Combine the set of weak learners '$E\ = \sum_{i=1}^{b} w_i$ '
17. **For** each w_i
18. Assign the weight 'ϕ'
19. Calculate training loss '$loss_t\ = (E_a - E_o)^2$'
20. Update the weight '$\nabla\phi$'
21. Find weak learner with minimum error $f\ (x) = arg\ \min\ [\ loss_t\ \ (w_i)]$
22. Obtain strong clustering results
23. **End for**

End

III. IMPLEMENTATION SCENARIO

The performance of the planned EAFK-SEGBBDC has discussed with Weeplaces Dataset. The dataset consists of different traits such as User id, place Id, DateTime, latitude, longitude, and city. In order to perform the clustering process, let us consider the latitude value to find the user with a similar location.

The adaptive fuzzy k means clustering technique helps to reduce the impartial function i.e. space between the cluster centroid and the data.

$$w_i = F = \arg\min\ \mu_{ij}\ D_{ij}$$

Let us consider $D_{ij} = 0.010\ \rightarrow F = 0.8 * 0.010$ =0.0080

Let us consider $D_{ik} = 0.008\ \rightarrow F = 0.8 * 0.008$ =0.0064

From the above calculation, the data d belongs to cluster 1 i.e. D_{ij} since the remoteness between the data and the cluster center is minimal i.e. 0.010

Where, F denotes an objective function, $\arg min$ denotes an disagreement of the minimum function, μ_{ij} denotes a association degree to which the data d belongs to a cluster center φ_j, D_{ij} indicates the remoteness between the i^{th} data and the j^{th} cluster centroid. The Fuzzy membership is measured based on the remoteness measure. Let us consider the fuzzifier '$n = 3$'

$$\mu_{ij} = \frac{1}{\sum_{i=1}^{n}\sum_{j=1}^{c}\left(\frac{D_{ij}}{D_{ik}}\right)^{\left(\frac{2}{n-1}\right)}}\quad (2)$$

$$\mu_{ij} = \frac{1}{\sum_{i=1}^{n}\sum_{j=1}^{c}\left(\frac{0.010}{0.008}\right)^{\left(\frac{2}{3-1}\right)}} = \frac{1}{(1.125)^{(1)}} = 0.8$$

Let us consider cluster 1 and their centroid is φ_j =40.7363 (latitude value),$d_i = 40.726144, q = 1$. The Minkowski distance among the data and the cluster center is measured as given below,

$$D_{ij} = \left(\left|d_i - \varphi_j\right|^q\right)^{1/q}\quad (3)$$

$$D_{ij} = (|40.726144 - 40.7363|^1)^{1/1} = 0.010$$

Let us consider cluster 2 and their centroid is $\varphi_k =$ 41.556974, $d_i = 40.726144, q = 1$. The Minkowski distance among the data and the cluster center is measured as given below,

$$D_{ik} = (|d_i - \varphi_k|^q)^{1/q}$$

$$D_{ik} = (|40.726144 - 40.734144|^1)^{1/1} = 0.008\quad (4)$$

The weak clustering results have some training error and its lack of providing accurate results. Therefore, the boosting ensemble technique combines all the weak learners, and weights are initialized to make a strong. Let us consider $w_i = 0.0080, \phi = 1$

$$E\ = \sum_{i=1}^{b} w_i * \phi = 0.008 * 1 = 0.008\quad (6)$$

$$E\ = \sum_{i=1}^{b} w_i + loss_t\quad (7)$$

$$E\ = 0.008\ + 0.0084\ \ \ = 0.0164$$

From (7), $loss_t$ designates a training loss of weak learner 'w_i'. From (), the squared error loss is estimated as given below,

$$loss_t = (E_a - E_o)^2 \quad (8)$$

$$loss_t = (0.1 - 0.008)^2 = 0.0084 \quad (8)$$

Where, 'E_a' denotes actual results, E_o indicate the observed results. The weights get updated according to the weight value. $\phi = 0.95$

$$E = \sum_{i=1}^{b} w_i * \phi = 0.008 * 0.95 = 0.0076$$

Then the training loss is calculated with the updated value.

$$loss_t = (0.1 - 0.0076)^2 = 0.0085$$

. The obtained results have a higher error than the previous error i.$e.$ $0.0085 > 0.0084$. The proposed ensemble method uses stochastic gradient descent step-size function to discover the weak learner results which have minimum training loss than the other discoveries the weak learner fallouts with minimum error i.e. 0.008. As a result, accurate clustering results are obtained. In other words, the input data is grouped into cluster 1.

IV. Results and Discussions

Clustering accuracy: It is measured as the ratio of the data that are correctly gathered into the particular cluster to the whole number of big data taken for experimental evaluation. The clustering accuracy is calculated using the mathematical formula,

$$CA = \left[\frac{\varepsilon_{AC}}{\varepsilon}\right] * 100 \quad (10)$$

From (10),CA denotes a clustering accuracy, 'ε_{AC}' indicates the amount of data that are accurately clustered, 'ε' represents a entire number of data occupied for performing the experimental evaluation. The correctness of clustering is measured in percentage (%).

Error Rate: It is measured as the relation of an amount of data incorrectly grouped into the cluster to the total amount of data. The error rate is calculated as given below,

$$ER = \left[\frac{\varepsilon_{WC}}{\varepsilon}\right] * 100 \quad (11)$$

From (11), ER indicates the error rate, ε_{WC} denotes the quantity of data incorrectly grouped into the cluster, 'ε' the total quantity of data. The error rate is measured in percentage (%).

Clustering Time: It is defined as an total time consumed by the algorithm to group related data into different clusters. The clustering time is calculated as given below,

$$CT = \varepsilon * t[CS] \quad (12)$$

Where CT denotes a clustering time, 'ε' denotes the amount of data taken for experimentation. The clustering spell is measured in milliseconds (ms).

Space complexity: It is measured as the extent of storage space consumed by the algorithm to stock the data. Therefore, the space complexity is expressed as given below,

$$com_{space} = \varepsilon * Mem[SD] \quad (13)$$

Where, com_{space} denotes a space complexity, ε denotes the number of data, '$Mem[SD]$' symbolizes a memory consumed to store a single geo-social data and 'ε' refers to a total number of geosocial data. The space complexity is measured in terms of Megabytes (MB).

Number of geo-social data	Clustering Accuracy (%)		
	EAFK-SEGBBDC	IWKM	DCPGS
1000	90	85	79
2000	92	88	83
3000	92	86	82
4000	94	87	81
5000	93	87	81
6000	92	85	80
7000	91	84	79
8000	92	86	77
9000	91	85	76
10000	92	84	75

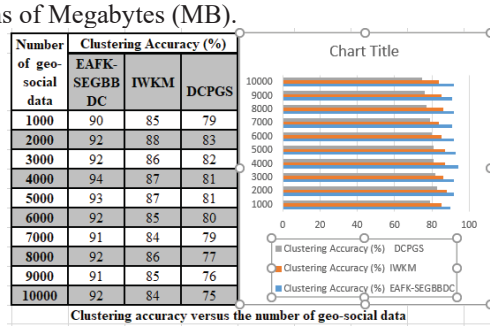

Fig 1: Clustering accuracy vs number of geo-social data

Number of geo-social data	Clustering time (ms)		
	EAFK-SEGBBDC	IWKM	DCPGS
1000	22	28	37
2000	24	32	40
3000	28	39	48
4000	30	46	52
5000	33	50	53
6000	37	51	54
7000	40	54	57
8000	44	58	61
9000	47	61	65
10000	50	65	70

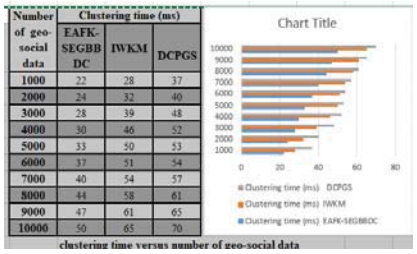

Fig 2: Clustering time vs number of geo-social data

Number of geo-social data	Error rate (%)		
	EAFK-SEGBBDC	IWKM	DCPGS
1000	10	15	26
2000	8	13	22
3000	8	14	21
4000	6	13	24
5000	7	13	25
6000	8	15	25
7000	9	16	27
8000	8	14	28
9000	9	15	29
10000	8	16	30

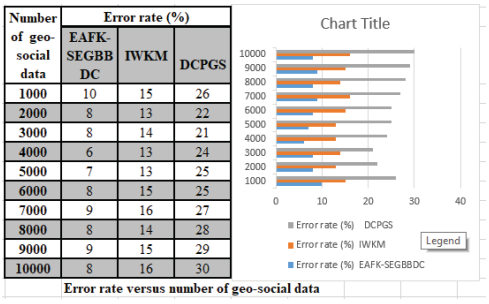

Fig 3: Error-rate vs number of geo-social data

Number of geo-social data	Space complexity (MB)		
	EAFK-SEGBBDC	IWKM	DCPGS
1000	20	24	28
2000	22	28	34
3000	27	33	36
4000	30	36	39
5000	32	40	44
6000	37	42	46
7000	40	46	52
8000	43	50	53
9000	46	53	58
10000	50	56	61

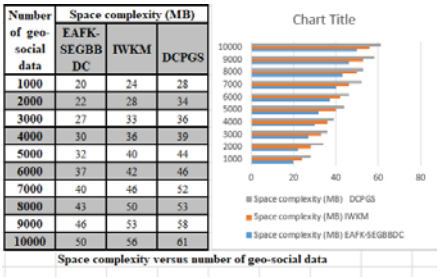

Fig 4: space complexity vs number of geo-social data

I. CONCLUSION

With the rise of the Big Data era, a novel distributed network clustering algorithm is required for a more efficient and valuable data examination. In order to deal with these problems in the geosocial network, a novel technique called EAFK-SEGBBDC is introduced. First, a big size of the data is gathered from the dataset. Then the Stochastic Extreme Gradient Boost clustering technique is used for clustering the big geo-social data into dissimilar clusters. The ensemble technique initially constructs the adaptive fuzzy k means clustering technique as a weak learner to group similar data into dissimilar clusters. The ensemble technique integrates weak learners to make a strong one. The Stochastic gradient step size is applied to find the strong clustering results with minimal error rate. Finally, it concluded that the proposed EAFK-SEGBBDC technique effectively handles the big data clustering in a reliable and better way. An experimental assessment is conducted for proposed and existing methods to show the performance improvement with different metrics. The performance results state that the proposed EAFK-SEGBBDC technique achieved better performance in achieving higher huddling accuracy, and minimum error rate, clustering time along with space complexity than the other related approaches.

II. REFERENCES

[1] Qian Tao, Chunqin Gu, Zhenyu Wang, Daoning Jiang, "An intelligent clustering algorithm for high-dimensional multiview data in big data application", Neuro computing, Elsevier, Volume 393, 14 June 2020, Pages 234-24

[2] Dingming Wu, Jieming Shi, and Nikos Mamoulis, "Density-based Place Clustering using Geo-Social Network Data", IEEE Transactions on Knowledge and Data Engineering, Volume 30, Issue 5, Pages 838 – 851, May 2018

[3] R.George, K.Shujaee, M.Kerwat, Z.Felfli, D.Gelenbe, K.Ukuwu, "A Comparative Evaluation of Community Detection Algorithms in Social Networks", Procedia Computer Science, Elsevier, Volume 171, 2020, Pages 1157-1165

[4] Smita Agrawal, Atul Patel, "SAG Cluster: An unsupervised graph clustering based on collaborative similarity for community detection in complex networks", Physica A: Statistical Mechanics and its Applications, Elsevier, 2020, Pages 1-21

[5] Yu Lei, Ying Zhou, Jiao Shi, "Overlapping communities detection of social network based on hybrid C-means clustering algorithm", Sustainable Cities and Society, Elsevier, Volume 47, 2019, Pages 1-8

[6] Weijia Lu, "Improved K-Means Clustering Algorithm for Big Data Mining under Hadoop Parallel Framework", Journal of Grid Computing, Springer, Volume 18, 2020, Pages 239-250

[7] Michele Ianni, Elio Masciari, Giuseppe M.Mazzeo, Mario Mezzanzanica, Carlo Zaniolo, "Fast and effective Big Data exploration by clustering", Future Generation Computer Systems, Elsevier, Volume 102, January 2020, Pages 84-94

[8] Marc Hüsch, Bruno U.Schyska, Luedervon Bremen, "CorClustST—Correlation-based clustering of big spatio-temporal datasets", Future Generation Computer Systems, Elsevier, Volume 110, September 2020, Pages 610-619

[9] Sunil Kumar and Maninder Singh, "A novel clustering technique for efficient clustering of big data in Hadoop Ecosystem", Big Data Mining and Analytics, Volume 2, Issue 4, 2019, Pages 240 – 247

[10] Salman Salloum, Joshua Zhexue Huang; Yulin He, "Random Sample Partition: A Distributed Data Model for Big Data Analysis", IEEE Transactions on Industrial Informatics, Volume 15, Issue 11, 2019, Pages 5846 – 5854