IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-8, Issue-2, Feb 2020 E-ISSN: 2347-2693

An Enhanced Fuzzy Based Linkage Clustering Algorithm (EFCA) in High Dimensional Data

R. Kiruthika[1*], V. Vijayakumar[2]

[1,2]Dept. of Computer Science, Sri Ramakrishna College of Arts and Science, Coimbatore, India

DOI: https://doi.org/10.26438/ijcse/v8i2.1217 | Available online at: www.ijcseonline.org

Accepted: 06/Feb/2020, Published: 28/Feb/2020

***Abstract*:** In data mining, clustering algorithm is a powerful meta-learning tool to precisely examine the huge volume of data created by recent applications. In particular, their major objective is to group data into clusters such that data points are grouped in the similar cluster when they are "similar" according to specific metrics. Several clustering algorithms have been developed to deal with very large number of features or with a very high number of dimensions, but they are often not practical when the data is large in both aspects. To address these issues, this paper work, developed an Enhanced Fuzzy based Linkage Clustering Algorithm (EFCA), which combines FCM and cluster assignment strategy to solve the optimization problem during high dimensional data processing. The proposed EFCA approach it can work with large volumes of high dimensional dataset for discovering the outliers. The experimental results shown that the proposed EFCA performance to improve 21.9% especial in terms of Partition Accuracy (PA), Dunn Index (DI) improves 28 %, and Computational time improves 16.4% compared with other existing clusiVAT and FensiVAT algorithms.

Keywords: Data mining, Big data cluster analysis, Fuzzy, Linkage.

I. INTRODUCTION

Data mining refers to the development of extracting or mining knowledge from sample amounts of data [1]. It is the process of searching available patterns by scanning the huge amount of data. Storing enormous quantity of data is utile to extract precious knowledge. To seek out constructive patterns within the data, there are different kinds of algorithms which can categorize the data either automatically or semi-automatically. These patterns are used to obtain the sets of rules. The patterns discovered must be meaningful such that they may lead to many advantages like decisions making, market analysis, financial growth, business intelligence etc. To get such meaningful patterns, significantly large amount of data is required.

In data mining, clustering is one of the framework in which data objects are grouped together without consulting a known class label. In clustering data groupings are not pre-defined; instead they are generated by finding the similarities between the data objects according to the characteristics found in the actual data. Based on this similarity the dataset is partitioned into several groups or clusters in such a way that objects within a cluster have high similarity in comparison with one another but are very dissimilar to objects in other clusters. In other words, a good clustering algorithm should maximize the intra-cluster similarity and minimize the inter-cluster similarity [9].

Big data is one of the new challenges in data mining because large volumes of high dimensional data and different varieties must be taken into account. The common methods and tools for data processing and analysis are unable to manage such amounts of data, even if powerful computer clusters are used. To analyze big data, many new data mining and machine learning algorithms as well as technologies have been developed. So, big data do not only yield new data types and storage mechanisms, but also new methods of analysis.

When dealing with big data, a data clustering problem is one of the most important issues. Often data sets, especially big data sets, consist of some groups (clusters) and it is necessary to find the groups. Clustering methods have been applied to many important problems [2], for example, to discover healthcare trends in patient records, to eliminate duplicate entries in address lists, to identify new classes of stars in astronomical data, to divide data into groups that are meaningful, useful, to cluster millions of documents or web pages. To address these applications and many others a variety of clustering algorithms has been developed. There exist some limitations in the existing clustering methods; most algorithms require scanning the data set for several times, thus they are unsuitable for big data clustering. There are a lot of applications in which extremely large or big data sets need to be explored, but which are much too large to be processed by traditional clustering methods.

To deal with large amounts of high-dimensional data, this paper introduces a rapid, Enhanced Fuzzy based Linkage

Clustering Algorithm (EFCA), which efficiently integrates (i) cluster assignment technique; and (ii) Linkage Clustering. The objective of this paper is an effective aggregation of cluster partitions, which are obtained using the EFCA on synthetic and real-world datasets.

The rest of the paper is organized as follows: Related work is detailed in Sect. 2. In Sect. 3, Proposed Methodology and performance metrics are described in Sect. 4. The conclusion is in Sect. 5.

II. LITERATURE REVIEW

The related work of several research filed or subject is must essential, before contributing in the research of that field. The preceding and continuing work connected to a clustering algorithm for large volumes of high dimensional data issues are focused in distributed manner. Clustering in high dimensional data algorithms rely on the competent computation of small sub-problems which leads to useful implementation. Therefore, this section presents a detailed literature review of the area taken (i.e., A Rapid Hybrid Clustering Algorithm for Large Volumes of High Dimensional Data).

***H. Gunadi* (2010)** [3] discussed a Nearest neighbor search, a problem that asks for a nearest point in the database given a query, is a problem in areas of computer science such as information retrieval, pattern recognition, image, and text processing. There is still ongoing research toward the improvement of the performance. Despite that, there is no comparison between methods proposed. We have no idea whether a method could work in high dimension or not, or how about the time and space complexity, or what about the results returned by specific method. The authors compared six of them here, they are Exhaustive, Vantage Point, Random Projection Matrix, Random Projection Tree (RP Tree), Random Ball Cover (RBC), and Locality-Sensitive Hashing (LSH). The authors consisted of comparisons between each method mentioned so that readers could see some characteristics of these methods to solve nearest neighbor search problem.

***T. C. Havens and J. C. Bezdek* (2012)** [4] discussed a VAT algorithm is a visual method for determining the possible number of clusters in, or the cluster tendency of a set of objects. The improved VAT (iVAT) algorithm uses a graph-theoretic distance transform to improve the effectiveness of the VAT algorithm for "tough" cases where VAT fails to accurately show the cluster tendency. The authors presented an efficient formulation of the iVAT algorithm which reduces the computational complexity of the iVAT algorithm from O(N 3) to O(N 2). Authors also proved a direct relationship between the VAT image and the iVAT image produced by our efficient formulation. They concluded with three examples displaying clustering tendencies in three of the Karypis data sets that illustrate the improvement offered by the iVAT transformation.

***Kumar, et.al.,* (2013)** [5] presented compared single linkage clustering based on MSTs built with the Filter-Kruskal method to the proposed clusiVAT algorithm, which is based on sampling the data, imaging the sample to estimate the number of clusters, followed by non-iterative extension of the labels to the rest of the big data with the nearest prototype rule. Numerical experiments with both synthetic and real data confirm the theory that clusiVAT produces true single linkage clusters in compact, separated data. They also showed that single linkage fails, while clusiVAT finds high quality partitions that match ground truth labels very well.

***Fahad, et.al.,* (2014)** [6] introduced a concepts and algorithms related to clustering, a concise survey of existing (clustering) algorithms as well as providing a comparison, both from a theoretical and an empirical perspective. From a theoretical perspective, authors developed a categorizing framework based on the main properties pointed out in previous studies. Empirically, they conducted extensive experiments where they compared the most representative algorithm from each of the categories using a large number of real (big) data sets. The effectiveness of the candidate clustering algorithms is measured through a number of internal and external validity metrics, stability, runtime, and scalability tests. In addition, they highlighted the set of clustering algorithms that are the best performing for big data.

***Popescu, et.al.,* (2015)** [7] discussed many contemporary biomedical applications such as physiological monitoring, imaging, and sequencing produce large amounts of data that require new data processing and visualization algorithms. Algorithms such as principal component analysis (PCA), singular value decomposition and random projections (RP) have been proposed for dimensionality reduction. The authors proposed a new random projection version of the fuzzy c-means (FCM) clustering algorithm denoted as RPFCM that has a different ensemble aggregation strategy than the one previously proposed, denoted as ensemble FCM (EFCM). RPFCM is more suitable than EFCM for big data sets (large number of points, *n*). To evaluated their method and compare it to EFCM on synthetic and real datasets.

***Kumar, et.al.,* (2016)** [8] presented a new clusiVAT algorithm and compare it with four other popular data clustering algorithms. Three of the four comparison methods are based on the well known, classical batch k-means model. Specifically, they used k-means, single pass k-means, online k-means, and clustering using representatives (CURE) for numerical comparisons. clusiVAT is based on sampling the data, imaging the reordered distance matrix to estimate the number of clusters in the data visually, clustering the samples using a relative of single linkage (SL), and then non-iteratively extending the labels to the rest of the data-set using

the nearest prototype rule. Previous work has established that clusiVAT produces true SL clusters in compact-separated data. They have performed experiments to showed that k-means and its modified algorithms suffer from initialization issues that cause many failures. On the other hand, clusiVAT needs no initialization, and almost always finds partitions that accurately match ground truth labels in labeled data.

***J. C. Bezdek* (2017)** [9] developed the necessary concepts such as similarity, distance, clusters, computer view point, human view point, and cluster validation in a very logical and lucid manner with plenty of easy-to-follow examples and creative pictures. The authors focused on four types of popular clustering algorithms, it provides adequate materials and pointers for interested readers to sail through a much wider family of clustering algorithms.

***Rathore, et.al.*, (2018)** [10] proposed a new random projection, fuzzy c-means based cluster ensemble framework for high-dimensional data. Their framework uses cumulative agreement to aggregate fuzzy partitions. Fuzzy partitions of random projections are ranked using external and internal cluster validity indices. The best partition in the ranked queue is the core (or base) partition. Remaining partitions then provide cumulative inputs to the core, thus, arriving at a consensus best overall partition built from the ensemble.

***Punit Rathore, et.al.*, (2019)** [11] proposed a novel fast clustering algorithm called FensiVAT. FensiVAT is a hybrid, ensemble-based clustering algorithm which uses fast data-space reduction and an intelligent sampling strategy. In addition to clustering, FensiVAT also provides visual evidence that is used to estimate the number of clusters (cluster tendency assessment) in the data. In the experiments, authors compared FensiVAT with nine state-of-the-art approaches which are popular for large sample size or high-dimensional data clustering.

This section provides a brief Literature review framework study rapid hybrid clustering in large volume of high dimensional data techniques details in this research. Meanwhile, this section discussed a various clustering techniques in various domain procedures.

III. PROPOSED METHODOLOGY

This paper presents a novel Enhanced Fuzzy based Linkage Clustering Algorithm (EFCA) in MATLAB simulation is applied to the synthetic and real-world dataset. This overall proposed flow diagram in figure 1 follows a high dimensional data clustering procedure form start to end state.

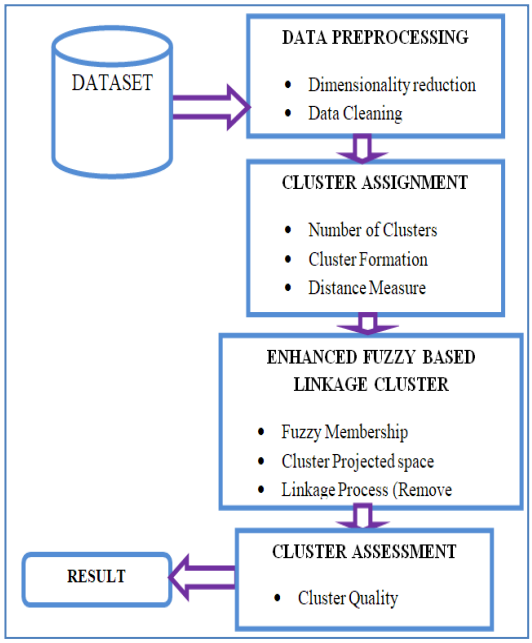


Figure 1: Proposed Flow diagram

A. *DATA PREPROCESSING*

Data preprocessing is one of the significant parts of machine learning. It plays a significant part in building a model. This process includes deleting duplicate/ redundant or irrelevant values from the high dimensional dataset. Duplicate observations most frequently arise during data collection and irrelevant observations are those that don't actually fit the specific problem that to trying to solve. Redundant observations alter the efficiency by a great extent as the data repeats and may add towards the correct side or towards the incorrect side, thereby producing unfaithful results. Irrelevant observations are any type of data that is of no use to be removed directly.

The predictable data preprocessing technique is respondings as it creates with data that is unspecified prepared for examination and there is no feedback and communicates for the method of data collection. The data variation among data sets is the major difficulty for the data preprocessing. In data preprocessing, clear out customs work to uncontaminated the data by satisfying in absent values, smoothing noisy data, recognizing or discarding outliers, and determining irregularities. The real procedure of data cleansing might occupy removing typographical errors or authenticating and accurating values against a known list of entities.

B. *CLUSTER ASSIGNMENT*

Cluster assignment is used to assign data to the clusters that were earlier generated by some clustering methods such as K-means or DBSCAN (Density-Based Spatial Clustering of Applications with Noise). The process follows for clusters generated by DBSCAN, all core objects are stored. For each piece of new data, the algorithm tries to find a core object in

some formed cluster whose distance is less than the value of the RADIUS parameter. If such a core object is found, the new data is then assigned to the corresponding cluster, otherwise it is assigned to cluster -1, indicating that it is noise. It is possible that a piece of data can belong to more than one cluster, which can be further divided into the following two cases:

- If the number of core objects whose distances to the new data is less than the MINPTS parameter value, meaning that the new data is a border object, the new data is assigned to the cluster where there is a core object having the smallest distance to the new data.
- If the number of core objects whose distances to the new data is not less than MINPTS, which means the new data is also a core object, it is then assigned to cluster -2, indicating that it belongs to more than one cluster. In this case, re-running the DBSCAN function is highly suggested.

The local density ρi of object xi is defined as

$$\rho_i = \sum_{j=1}^{T} e^{-d_{ij}^2/dc^2} \quad \text{eqn. (1)}$$

where d_{ij} is the distance between two objects x_i and x_j and dc is a cutoff distance.

Algorithm 1: CLUSTER ASSIGNMENT
Input: Dataset and p
Output: V and $c$max
$D = (dij)n\times n$;
dc = cutoffdis (p); /∗Calculate the cutoff distance dc. ∗/
$\rho = (\rho i)n$;
idx = arg($sort(\rho, descent)$);
/∗The number of object corresponding to sorted
ρ. ∗/
$k = 1$;
$cl = (-1)1\times n$;
/∗The cluster number that each object belongs to, with initial value −1. ∗/
for $i = 1$ to $n - 1$ do
if $clidxi \sim = -1$ && $\#\{Neighbor(xidxi) > 1\}$ then /∗$xidxi$ does not belong to any cluster and the number of its neighbors is greater than 1. ∗/
V$k = xidxi$;
for $j \in Neighbor(xidxi)$ do
$clj = k$;
$k = k + 1$;
$c$max = k;

C. ENHANCED FUZZY BASED LINKAGE CLUSTERING

The Enhanced Fuzzy based Linkage Clustering Algorithm (EFCA) is an extension version of FCM algorithm and convergence properties of the proposed method are recapitulated. Proposed method dynamically updates feature-weight vectors in its training phase rather than utilizing a fixed feature-weight linkage vector. Because the feature-weight linkage vector of the traditional fuzzy c-means algorithm remains fixed during the clustering procedure, the significance of certain features to the changing cluster information cannot be appropriately manifested.

Let the dataset set $D = \{ds_1, ds_2, ds_3 \ldots ds_n\}$ denote a set of data points to be portioned into c clusters, where x_i ($i = 1, 2, 3 \ldots n$) is the data points. The fuzzy objective function is to discover nonlinear relationships among the data methods use embedding linking's that connectivity features of data to new feature weight spaces. The proposed technique Fuzzy based linkage clustering algorithm is an iterative clustering technique that minimizes the objective function.

Given an dataset, $D = \{x_1 \ldots x_n\} \subset Rp$, the proposed algorithm partitions X into c fuzzy partitions by minimizing the following objective function as,

$$J(U, V, D) = \sum_{i=1}^{C} \sum_{j=1}^{N} \mu_{ij}^{m} [d_{ij}^{w}]^2 \quad \text{eqn. (2)}$$

Where C is the number of clusters and selected as a specified value, N the number of data points, u_{ik} the membership link of x_k in class i, satisfying the $\sum_{i=1}^{c} u_{ik} = 1$, m the quantity scheming clustering, and V the set of cluster centers or prototypes ($v_i \in$Rp). The proposed clustering algorithm works message passing among data points. Each data points receive the availability from others data points (from pattern) and send the responsibility message to others data points (to pattern). Sum of responsibilities and availabilities for data points identify the cluster patterns.

PSEUDO CODE OF ENHANCED FUZZY BASED LINKAGE CLUSTER
Input: Dataset, number of clusters
Output: Cluster Result
Process
Step 1: Initialize number of clusters C, fuzzy exponent m and fuzzy partition matrix
Step 2: Initialize feature-weight linkage vector using normalized Term Variance:
Step 3: while (not achieve termination condition)
Update the cluster centers
Calculate the distances
Update the fuzzy partition matrix:
Update the elements in the feature-weight linkage vector
End while
End Process

D. CLUSTER ASSESSMENT

The cluster assessment is performed by distance matrix that chooses a subset of the compound space which consists only compounds which have sufficient number of close neighbors.

This is obtained based on the descriptor chosen in the earlier step. The similarity measures often used in calculation of similarity between chemical compounds are Euclidean measures. The similarity measure chosen is the Euclidean distance, which is based on the triangle inequality. Euclidean measure is chosen because it shows that it was best used in shared-Neighbor clustering.

Euclidean distances are usually computed from raw data and the advantage of this method is that the distance between any two object is not affected if we add new objects (such as outliers) into the analysis. The similarity measures using Euclidean distance is measured based on inter-point distance d(x_1, x_2) and the equations for binary descriptor is as follows:

$$d(x_1,x_2)=\sqrt{x_1 - x_2{}^2} \qquad \text{eqn. (3)}$$

The distance of the similarity matrix, the result gained will be the input for the calculation of the cluster method chosen.

IV. EXPERIMENAL RESULTS AND DISCUSSIONS

The proposed developed an Enhanced Fuzzy Based Linkage Clustering (EFCA) algorithm and evaluate the performance of Partition Accuracy (PA), Dunn Index (DI) and Run Time (RT) measures in large volumes of high dimensional data clustering in MATLAB simulation with Intel I5-6500 series 3.20 GHz 4 core processor, 8GB main memory, and runs on the Windows operating system. To evaluate the performance of the proposed EFCA in US Census 1990, KDD CUP'99 and FOREST real world dataset use an internal cluster validity index, Dunn's Index (DI) [13], [14], Partition Accuracy to evaluate the quality of output partitions for all clustering algorithms for this dataset. DI is a metric of how well a set of clusters represent compact separated clusters. DI for a partition U, is defined as,

$$DI(k,U) = \frac{min_{1\le i,j\le k\ i\ne j} dist(C_i, C_j)}{max_{1\le l\le k}\, diam(C_l)} \quad eqn.(4)$$

where C is the i^{th} cluster, dist(C_i, C_j) is the distance between two clusters, and diam(C_l) is the cluster diameter (maximum distance within a cluster).

Table 1 shows the comparison of Average Partition Accuracy PA (%) values with existing FensiVAT, clusiVAT and proposed EFCA.

Table:1 Comparison of Average PA (%) with existing FensiVAT, clusiVAT and proposed EFCA

Methods	KDD	FOREST
clusiVAT [5]	96.1	48.9
FensiVAT [12]	96.1	48.9
EFCA	**96.8**	**21**

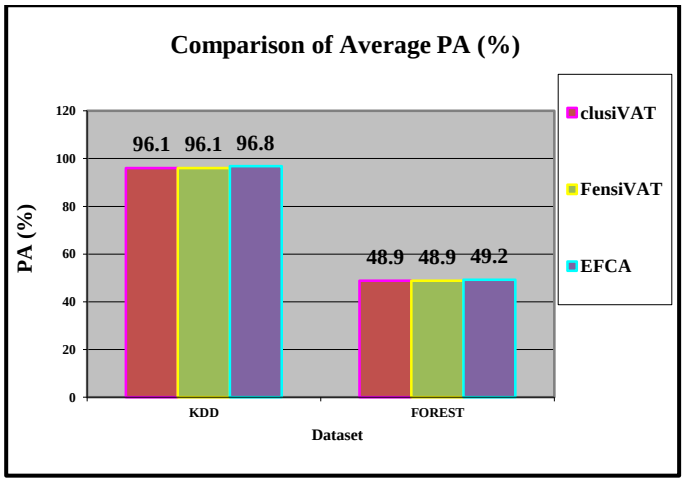
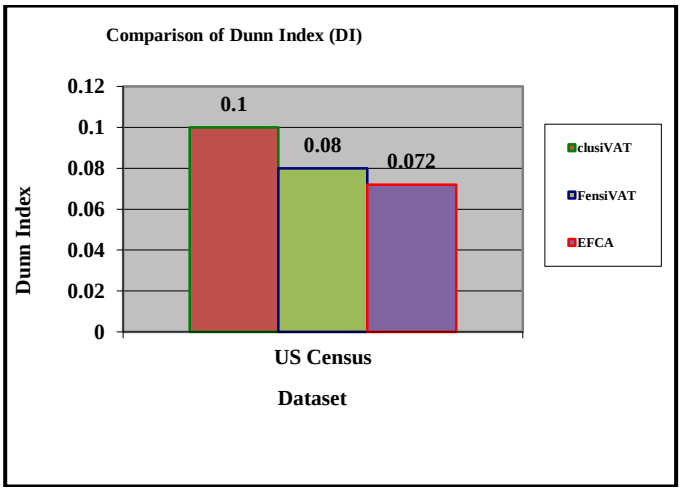


Table 2 shows the comparison of Dunn Index values with existing FensiVAT, clusiVAT and proposed EFCA.

Table 2: Comparison of Dunn Index (DI) with existing FensiVAT, clusiVAT and proposed EFCA

Methods	US Census
clusiVAT [5]	0.10
FensiVAT [12]	0.08
EFCA	**0.072**

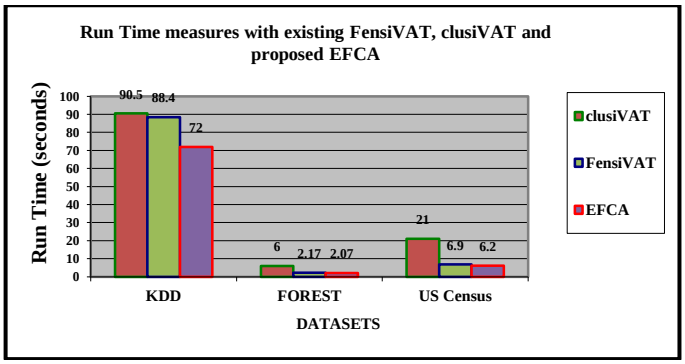


In table 3 represents the performance of Run Time of the testing datasets.

$$Run\ Time = \frac{Total\ time\ taken\ by\ all\ Running\ tasks}{Number\ of\ Running\ tasks} \quad \text{eqn. (5)}$$

Table 3: Run Time measures with existing FensiVAT, clusiVAT and proposed EFCA.

Methods	KDD	FOREST	US Census
clusiVAT [5]	90.5	6	21
FensiVAT [12]	88.4	2.17	6.9
EFCA	**72**	**2.07**	**6.2**



The experimental results shown that the proposed EFCA performance to improve 21.9% especial in terms of Partition Accuracy (PA), Dunn Index (DI) improves 28 %, and Computational time improves 16.4% compared with other existing clusiVAT and FensiVAT algorithms.

V. CONCLUSION

In this paper designed and implemented an enhanced method of clustering algorithm for large volumes of high dimensional data using Enhanced Fuzzy based Linkage Clustering Algorithm (EFCA), which combines FCM and cluster projected space strategy to solve the optimization problem during high dimensional data processing. The proposed Enhanced Fuzzy based Linkage Clustering Algorithm in high dimensional data clustering is in core variations of fuzzy clustering algorithm using different weight linkage measures applied to the vector of base-level clustering's baseline on both synthetic and real-world data. According to the proposed EFCA approach it can work with large volumes of high dimensional dataset for discovering the outliers.

REFERENCES

[1] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, **1996.**

[2] Hoppner, F.; Klawnn, F.; Kruse, R.; Runkler, T.;. Fuzzy Cluster Analysis: "Methods for classification, data analysis and image recognition", John Wiley & Sons, Inc., New York NY., **2000.**

[3] H. Gunadi, "Comparing nearest neighbor algorithms in highdimensional space," **2011.**

[4] T. C. Havens and J. C. Bezdek, "An efficient formulation of the improved visual assessment of cluster tendency (iVAT) algorithm," IEEE Trans. Knowl. Data Eng., **vol. 24**, **no. 5**, pp. 813–822, **May 2012.**

[5] D. Kumar, M. Palaniswami, S. Rajasegarar, C. Leckie, J. C. Bezdek, and T. C. Havens, "clusiVAT: A mixed visual/numerical clustering algorithm for big data," in Proc. IEEE Int. Conf. Big Data, **pp. 112–117, 2013**.

[6] A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A. Y. Zomaya, S. Foufou, and A. Bouras, "A survey of clustering algorithms for big data: Taxonomy and empirical analysis," IEEE Trans. Emerging Topics Comput., **vol. 2**, **no. 3**, pp. 267–279, **Sep. 2014.**

[7] M. Popescu, J. Keller, J. Bezdek, and A. Zare, "Random projections fuzzy c-means (RPFCM) for big data clustering," in Proc. IEEE Int. Conf. Fuzzy Syst**., pp. 1–6, 2015**.

[8] D. Kumar, J. C. Bezdek, M. Palaniswami, S. Rajasegarar, C. Leckie, and T. C. Havens, "A hybrid approach to clustering in big data," IEEE Trans. Cybern., **vol. 46**, **no. 10**, pp. 2372–2385, **Oct. 2016.**

[9] J. C. Bezdek, Primer on Cluster Analysis: Four Basic Methods that (Usually) Work, vol. 1. Sarasota, FL, USA: First Edition Design Publishing, **2017.**

[10] P. Rathore, J. C. Bezdek, S. M. Erfani, S. Rajasegarar, and M. Palaniswami, "Ensemble fuzzy clustering using cumulative aggregation on random projections," IEEE Trans. Fuzzy Syst., **vol. 26**, **no. 3**, pp. 1510–1524, Jun. **2018.**

[11] P. Rathore, A. S. Rao, S. Rajasegarar, E. Vanz, J. Gubbi, and M. Palaniswami, "Real-time urban microclimate analysis using internet of things," IEEE Internet Things J., **vol. 5**, **no. 2**, pp. 500–511, **Apr. 2018.**

[12] Punit Rathore, Dheeraj Kumar, James C. Bezdek, Sutharshan Rajasegarar, and Marimuthu Palaniswami, "A Rapid Hybrid Clustering Algorithm for Large Volumes of High Dimensional Data", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, **VOL. 31**, **NO. 4**, **APRIL 2019.**

[13] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," J. Cybern., **vol. 3**, **no. 3**, pp. 32–57, **1973.**

[14] P. Rathore, Z. Ghafoori, J. C. Bezdek, M. Palaniswami, and C. Leckie, "Approximating Dunn's cluster validity indices for partitions of big data," IEEE Trans. Cybern.

