Buddy System Based Alpha Numeric Weight Based Clustering Algorithm with User Threshold

M PRASAD * and Srikanth Thota

Posted Date: 24 August 2023

doi: 10.20944/preprints202308.1676.v1

Keywords: Clustering; hierarchical agglomerative clustering; Alpha numeric Weight based of Object Positional Value for a Term / Field / Attribute; Clustering Ranges; Buddy System

Article

Buddy System Based Alpha Numeric Weight Based Clustering Algorithm with User Threshold

Maradana Durga Venkata Prasad[1,*] and Dr.Srikanth T[2]

[1] Research Scholar, Department of Computer Science and Engineering, Gandhi Institute of Technology and Management (GITAM), Visakhapatnam, Andhra Pradesh, India; powersamudra@gmail.com
[2]Associate Professor, Department of Computer Science and Engineering, Gandhi Institute of Technology and Management (GITAM), Visakhapatnam, Andhra Pradesh, India; sthota@gitam.edu

Abstract: Data is present in the data sources like Files and Data bases, Retrieval of information from that data sources is one of the important issue nowadays. So for retrieval information from the data sources clustering is used. In the present market different types of clustering algorithms were available. But opting of the clustering is based on user requirements. This paper focuses on the study of hierarchical clustering approach on different conditions or measures or with customer choices like clustering process number of clusters generated at each level, number of levels, attributes range for performing the clustering on the given data set. In brief overview we discuss the hierarchical approach for clustering algorithm with the user opting choices.

Keywords: clustering; hierarchical agglomerative clustering; Alpha numeric weight based of Object Positional Value for a Term/Field/Attribute; Clustering Ranges; buddy system

I. Introduction

Clustering is a collecting of grouping set of objects where all similar come into one group and dissimilar objects will come into other group [1].Clustering is a one of the method which is used in the data mining process, feature extraction and data classification. Among all the clustering algorithms, hierarchical clustering approach is a hot topic in the current era. Hierarchical clustering approaches are of two kinds. They were Agglomerative clustering approach and Divisive clustering approach [2]. Divisive approach is a top down approach for clustering the given data set and forms a hierarchical clustering tree. In order to get good clusters from the given data set, we have to go for user preferences. Clustering Process Creates 2 Groups of clusters. They were

Table 1. Types of Objects in the clustering Process.

Objects Group	Details
Similar	All Objects of same type
Dissimilar	All Objects of different type

Clustering Distances

For the given data set, the distance between any two clusters is called as Clustering Distance. It is of two types. They were Inter-cluster and Intra cluster Distance.

Table 2. Types of cluster Distances.

Type of cluster Distance	Details
Intra cluster	It is the distance between the centroid of a cluster and a data item present within a cluster.
Inter cluster	It is the distance between the data items in distinct clusters.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 August 2023 doi:10.20944/preprints202308.1676.v1

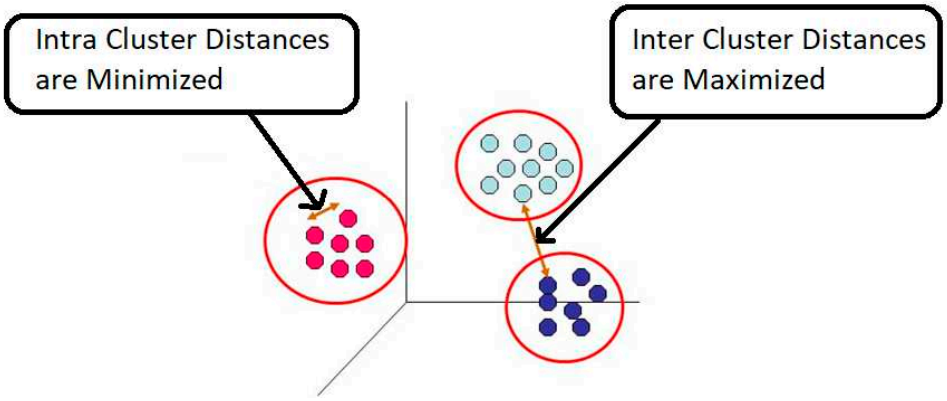

Figure 1. Types of cluster Distances.

Table 3. Clustering Outline.

Techniques of Data Mining	**Clustering, Classification, Association mining, Text mining, Sequential patterns, prediction, Decision trees, and Regression.**
Technique	Unsupervised
Classes	Used for Finite set of classes
Data Mining Task	Descriptive
Goal	Used for finding similarities in the data set.
Data Set	Used for finite set of data
Objects Similarity	Defined by similarity function

Table 4. Outline of Supervised and Unsupervised Learning [3].

Property	**Supervised Learning**	**Unsupervised Learning**
Definition	It uses datasets which are labeled on algorithms to train, classify data or predict outcomes correctly.	Machine learning is used in the Unsupervised Learning concepts to understand, cluster unlabeled datasets.
Number of Classes	Unknown	Unknown
Labeling	Input data Labeled	Input data unlabeled
Output	Known	Unknown
Uses	Training Data Set	Input Data set
Knowledge Required on	Training Set	No previous knowledge
Classify Used for	later observations	data understanding
performance Measure	Accurately	Indirect / Qualitative
Used for	Analysis	Prediction
Examples	Classification, regression e.t.c	Clustering, Association e.t.c

Types of Hierarchical Clustering

It is of two types. They were Divisive and Agglomerative.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 August 2023 doi:10.20944/preprints202308.1676.v1

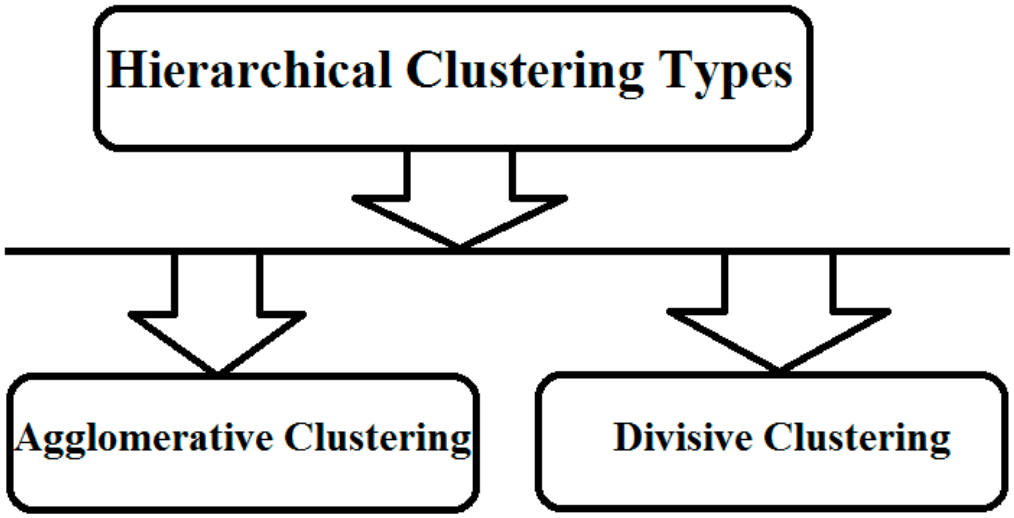

Figure 2. Types of Hierarchical Clustering.

Divisive

It starts with complete data set and divides it into successfully small clusters [4].

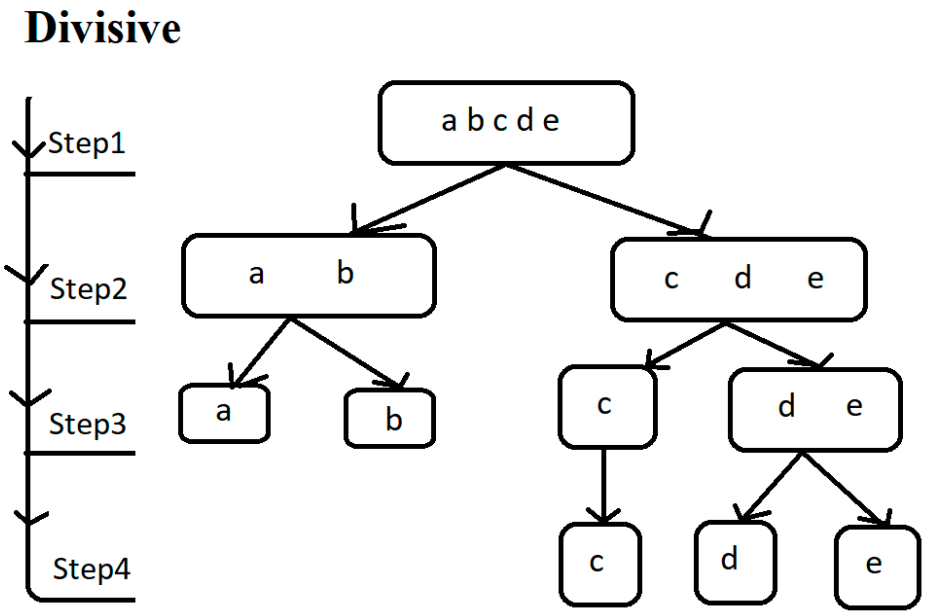

Figure 3. Divisive Clustering Example.

Agglomerative

It starts with single element as a unique cluster and combines them into successfully big clusters [5].

I.e. Data mining means extraction of data from data sources. So, whatever the data extracted will be used by the user. So, we have to take the preferences of the user before the clustering process. So, at the end of the clustering the user will get good Clustering resultsfrom a given dataset.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 August 2023 doi:10.20944/preprints202308.1676.v1

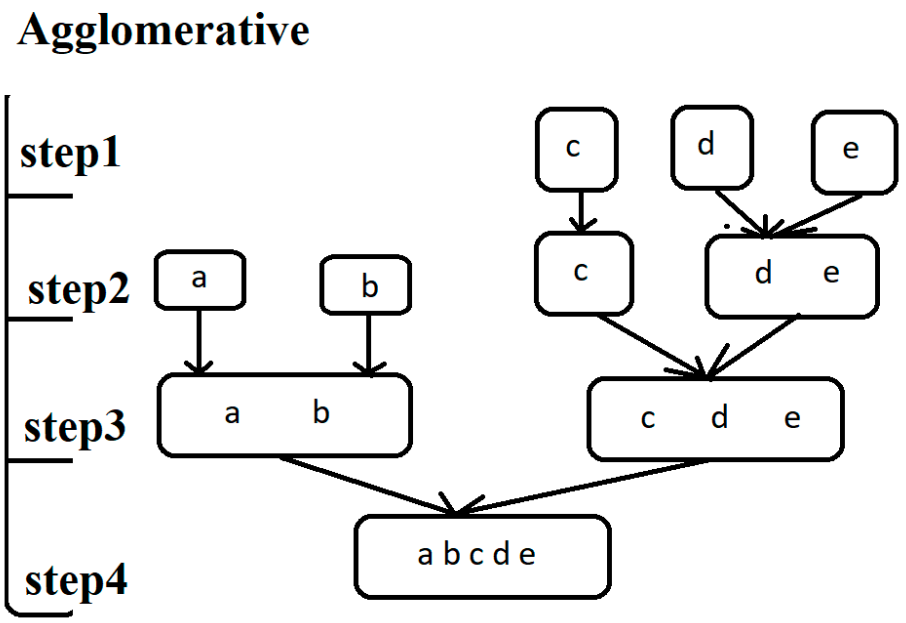

Figure 4. Agglomerative Clustering Example.

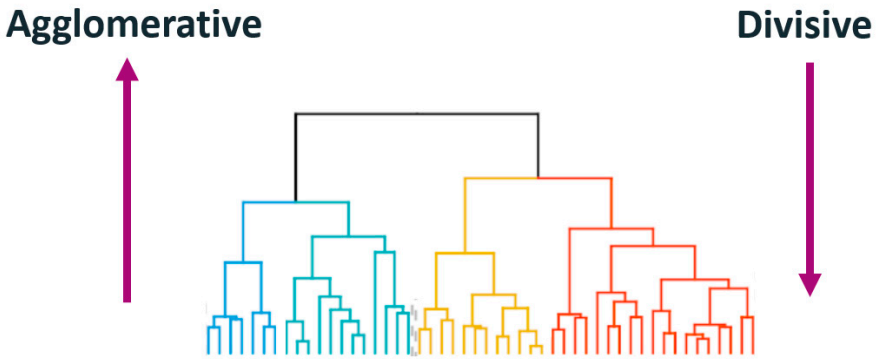

Figure 5. Agglomerative and Divisive clustering indicating dendrograms.

Applications of Clustering

In the current era the data in the servers is growing day by day so Clustering is used in many Areas such Market Analysis, Outlier Detection, Classification of Documents, Data Mining Function, Pattern Recognition, Image Processing, Anomaly Detection, Medical Image processing, Grouping of Search Results, for the Analysis of Social Networks e.t.c.

Table 5. Clustering Application Areas.

Clustering Application Areas	Clustering Purpose
Market Analysis	It is used to identify different groups of people in the Market and to identify their requirement to provide the products Required by the Customers [6].
Outlier Detection	It is used in outlier detection applications. Example of credit card fraud [7].
Classification Of Documents	It is used to classify the documents in WWW (world wide web) [8].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 August 2023 doi:10.20944/preprints202308.1676.v1

Data Mining Function	It is used in cluster analysis (to observe characteristics of each cluster.)[9].
Pattern Recognition	It is used in traffic Pattern Recognition to clear traffic problems [10].
Image Processing	It is used in Image Processing for segmentation of image [11].
Anomaly Detection	It is used in anomaly detection is to study normal modes in the data available and it is used to point out anomalous are there or not [12].
Medical Imaging	It is used in Medical Imaging for segmentation of the images and analyzes it [13].
Search Result Grouping	It is used in the grouping of search results from the WWW when the users so the Search [14].
Social Network Analysis	It is used to merge the entities of a social network into distinct Classes depends on their relationships and links between the classes [15].

e.t.c

Stages of Clustering

Total there are three stages in the clustering process [16]. They were

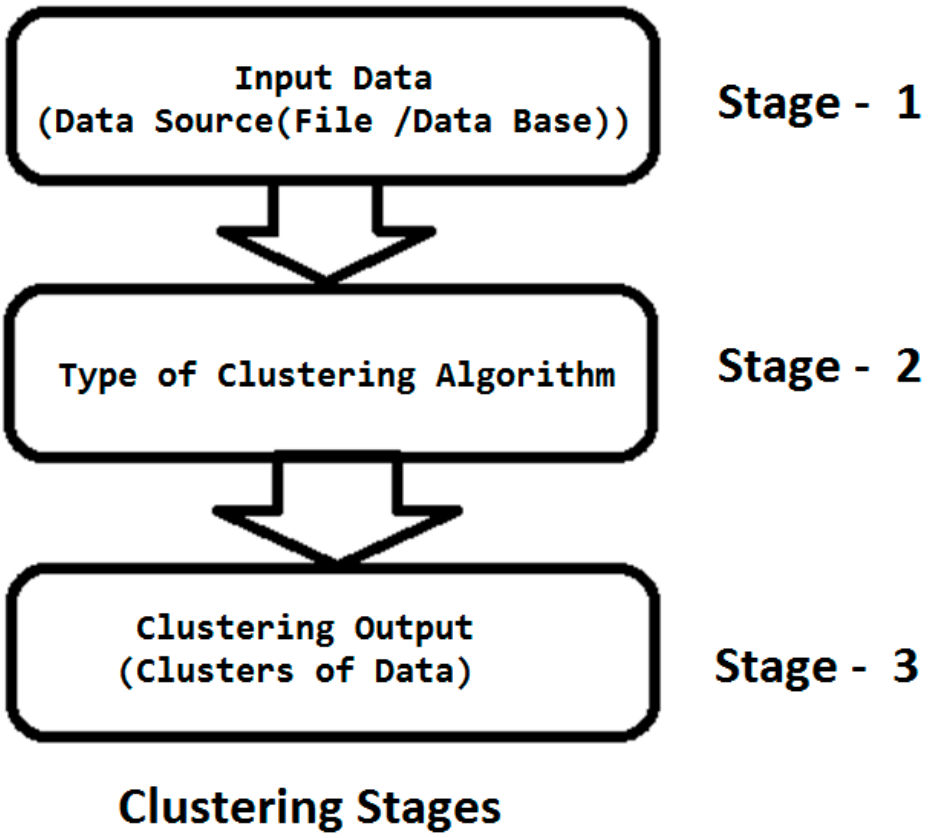

Figure 6. Stage of clustering.

Note:

1. In the stage one, Input Data to clustering algorithm is collected from a file or a data base.
2. In the stage two, Different Types of clustering algorithm are utilized to process the stage 1data.In the current ERA different types of clustering algorithms are available in the market like Constraint Based Method, Soft Computing, Partitioning, Hierarchical, Density Based, Grid Based, Model Based, Bi-clustering, Graph Based, e.t.c.
3. In Knowledge Discovery in Database clustering is a part of it. [17].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 August 2023 doi:10.20944/preprints202308.1676.v1

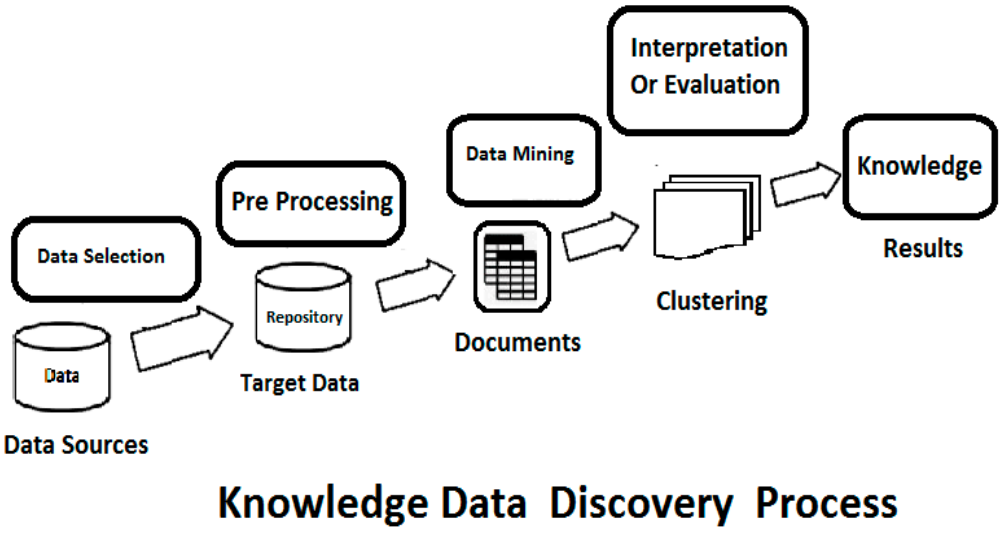

Figure 7. Knowledge Data Discovery Process.

4. Data Preprocessing Techniques [18]

It includes Data Reduction, Cleaning and Transformation.

Table 6. Techniques of Data preprocessing.

Techniques of Data preprocessing	Handle Operations
Data Cleaning [19]	Missing Data, Noisy Data
Data Reduction[20]	1. Attribute Subset Selection (Attributes). 2. Numerosity Reduction (Reduces data by replacing original data by smaller form of data representation). 3. Dimensionality Reduction (Compression the data).
Data Transformation[21]	1. Smoothing (Remove Noise). 2. Aggregation (Generates Attributes summary). 3. Generalization (Converts Low level data to high level data). 4. Normalization (Scales the Attributes). 5. Attribute Construction (Create New Attributes)

II. Literature Survey

In the market Different types clustering methods were there proposed by different researcher's persons. For each clustering method there will be one or more sub clustering Algorithms. Each sub clustering algorithm will have its own constraints. The major clustering methods available in the market were

Table 7. Types of Clustering.

S. No	Clustering Type	Details	Sub Clustering Methods

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 August 2023 doi:10.20944/preprints202308.1676.v1

1.	Partitioning	It is used to group data by moving objects from one group to another using relocation technique [22].	1. CLARA. 2. CLARANS. 3.EMCLUSTERING 4. FCM. 5. K MODES. 6. KMEANS. 7. KMEDOIDS. 8. PAM. 9. XMEANS
2.	Hierarchical	It is used to create clusters based on a particular similarity in the objects [23].It is of two types. They were Agglomerative and Divisive. Agglomerative clustering clusters the data based on combining clusters up. Divisive clustering clusters the data based on merging clusters down. Other names for hierarchical clustering are hierarchical cluster analysis or HCA.	1. AGNES. 2. BIRCH. 3. CHAMELEON. 4. CURE. 5. DIANA. 6. ECHIDNA 7. ROCK.
3.	Density Based	It is used radius as a constraint to group the data. Here data points within a particular radius are considered as a group and remaining points are considered as noise [24].	1. DBSCAN. 2. OPTICS. 3. DBCLASD 4. DENCLUE. 5. CENCLUE.
4.	Grid Based	Grid Based calculates the density of cells and based on the cells densities values clustering is done [25].	1. CLIQUE. 2. OPT GRID. 3. STING. 4. WAVE CLUSTER.
5.	Model Based	It uses a statistical approach for clustering the data where each object is assigned a weight (probability distribution) which is used to cluster the data [26].	1. EM. 2. COBWEB. 3. CLASSIT. 4. SOMS.
6.	Soft Computing	In Soft Computing, clusters individual data points are assigned to more than one cluster and after clustering the clusters will have minimum similarity [27].	1. FCM. 2. GK. 3. SOM. 4. GA Clustering
7.	Biclustering	It is the clustering of the rows and columns of a matrix simultaneous using a data mining technique [28].	1. OPSM. 2. Samba 3. JSa
8.	Graph Based	Graph is a collection of nodes (vertices). Clustering of these nodes of a graph based on certain weights assigned to the nodes is called as Graph Based Clustering [29].	1. Click
9	Hard Clustering	In Hard Clustering individual data points are assigned to a unique cluster and after clustering clusters will have maximum similarity [30].	1. KMEANS

Each Clustering method calculates different types of parameters for doing the Clustering on a given data set. The time and space complexity of the algorithms are different for different clustering algorithms.

Similarity Measures Used by Different Clustering Methods [50]

Similarity measures are used to identify the good clusters in the given data set. There are so many Similarity measures used in the current market. They were Average Distance, Canberra Metric, Chord, Clustering coefficient, Cosine, Czekanowski Coefficient, Euclidean distance,

Index of Association, Kmean, KullbackLeibler Divergence, Mahalanobis, Manhattan distance or City blocks distance, Mean Character Difference, Minkowski Metric, Pearson coefficient,

Weighted Euclidean e.t.c

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 August 2023 doi:10.20944/preprints202308.1676.v1

Table 8. Similarity measures in Clustering.

S.No	Similarity measures Name	Details
1	Average Distance	It is the Euclidean distance but a modified version[31]. $\text{Average Distance} = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2\right)^{\frac{1}{2}}$ Here x, y are data points in n-dimensional space
2	Weighted Euclidean	It is the modified version of Euclidean distance [32]. $\text{Weighted Euclidean Distance} = \left(\sum_{i=1}^{n} w_i(x_i - y_i)^2\right)^{\frac{1}{2}}$
3	Chord	It is the length calculated between two points which are normalized within a hypersphere of radius one [33].
4	Mahalanobis	It is the distance sample point (outlier) and a distribution [34].
5	Mean Character Difference	It is calculated using all points in the given space [35]. $\text{Mean Character Difference} = \frac{1}{n}\sum_{i=1}^{n}\lvert x_i - y_i\rvert$
6	Index of Association	It is calculated using all points in the given space [36]. $\text{Index of Association} = \frac{1}{n}\sum_{i=1}^{n}\left\lvert \frac{x_i}{\sum_{i=1}^{n} x_i} - \frac{y_i}{\sum_{i=1}^{n} y_i}\right\rvert$
7	Canberra Metric	It is calculated using all points in the given space [37]. $\text{Canberra Metric} = \sum_{i=1}^{n}\frac{\lvert x_i - y_i\rvert}{(x_i + y_i)}$
8	Czekanowski Coefficient	It is calculated using all points in the given space [38]. $\text{Czekanowski coefficient} = 1 - \frac{2\sum_{i=1}^{p}\min(x_i, y_i)}{\sum_{i=1}^{p}(x_i + y_i)}$
9	Pearson coefficient	It is calculated using all points in the given space [39]. $\text{Pearson coefficient} = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}}$
10	Minkowski Metric	Minkowski Distance / metric are the distance between two vectors and it's a generalization of both the Manhattan and Euclidean distance [40].
11	Manhattan distance or City blocks distance	It is the distance between vectors. It is equal to the one-norm of the distance between the vectors [41].
12	Euclidean distance	Euclidean distance is also known as Pythagorean distance. It is the distance between any two points (Cartesian coordinates) in the Euclidean space [42].
13	KullbackLeibler Divergence	KullbackLeibler Divergence is used to calculate the distance between two independent discrete probability distributions (data and cluster center point). It is used to create cluster group by combining multiple Fuzzy c-means clustering's results [43].
14	Clustering coefficient	It is used to calculate how the nodes of a graph are connected along with the degree [44].
15	Cosine	It is the replacement of Euclidean distance with cosine function [45].
16	Kmean	It is the mean of all the coordinates or points in the in the Euclidean space [46].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 August 2023 doi:10.20944/preprints202308.1676.v1

Inputs and Outputs in Clustering Process

In every clustering algorithm the user gives so many parameters as inputs to the clustering algorithm and gets the outputs.

Table 9. Inputs and Outputs in clustering process.

S.NO	Inputs and Outputs	Details
1	Number of Inputs for the clustering process	Clustering Algorithm, Algorithm Constraints, Number of Levels and clusters per each level.
2	Number of Levels	In the entire clustering.
3	Number of clusters	At each stage
4	Sum of Square Error (SSE) or other errors	It is a measure of difference between the data obtained by the prediction model that has been done before. [47]
5	Likelihood of Clusters	It is the similarity of clusters in the data points.[48]
6	Unlikelihood of Clusters	It is the dissimilarity of clusters in the data points.
7	Number of variable parameters at each level	These are the input parameters which are changed during the running of the algorithm like threshold.
8	outlier	In the clustering process any object doesn't belong to any cluster it is called as a outlier.[48]
9	Output	Clusters

III. Proposed Algorithm

Buddy Memory Allocation Technique invented by Harry Markowitz for Memory Allocation Technique in the year 1963. Buddy memory allocation technique is used to divide memory into 2 halves and gives a best fit and is easy to implement it [51].Here Modified Buddy system is used for clustering on Alpha numeric weighted based sorted records for clustering.

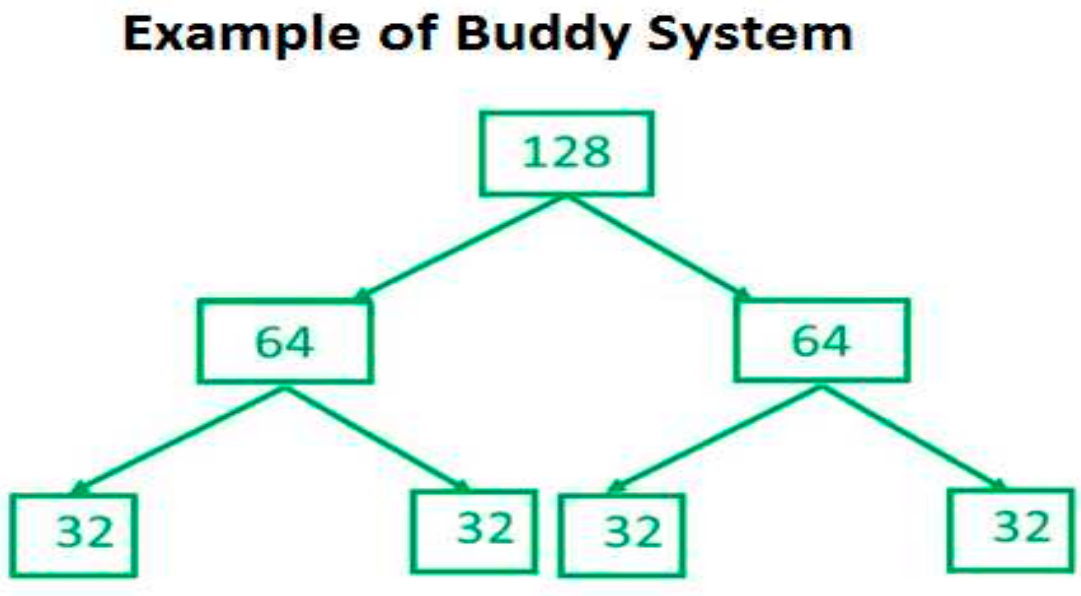

Figure 8. Example of Buddy System.

The problem with the buddy system is it works for the even number of records. But it is difficult to work with odd number of records. So I our algorithm it works for both even and odd record count.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 August 2023 doi:10.20944/preprints202308.1676.v1

Buddy System with Alpha Numeric Weighted Based Clustering Algorithm

1. Take a Sample data set.
2. Calculate The Weight of Individual Object by Position of an attribute of a particular column for all records.

a. Object individual position is calculated using ASCII Character Binary Table.
b. Formula for calculating the **Alpha Numeric Weight based Object Positional Value for a Term** / Field (**ANWAPVT**) of a record.
ANWAPVT= (First Char) ASCII value* n+ (second Char) ASCII value *(n-1) ------- (Last char-3) + (Last char-2) ASCII value*3+ (Last char-1) ASCII value*2+Last char*ASCII value*1
Example: Term is AB.AB=65*2+66*1=130+66=196
Here "A" ASCII value is 65 and "B" ASCII value is 66.

3. Sort the records / data in ascending order as per the Alpha Numeric Weight based Object individual position values. For that call the Sort Function or write a sort function to sort the records.
4. Compute the number of records (N) in the Data source (data base / set / File). Specify Number of levels (L) that should be generated in the clustering process which should be always $\mathbf{2^L<N}$.
5. Use Modified buddy system for level wise cluster generation. I.e. At each level, every cluster is splits into two sub clusters and adds computed clusters to a list. Here list index indicates the cluster number.

Table 10. Example of Modified buddy system.

If No. of elements (N)	Elements in Cluster1	Elements in Cluster2
Even	N /2	N /2
Odd	N /2	N /2 + 1

I.e. cluster1 is divided into two clusters (cluster2 and cluster3).

6. Repeat the step5 till $\mathbf{2^L< N}$. Where

L = Number of levels that has to be generated using buddy system.
N = Number of records in the data set.

7. Assign the elements to each cluster in a based on the list from the sorted data set.
8. Use the list to generate a pie circle chart.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 August 2023 doi:10.20944/preprints202308.1676.v1

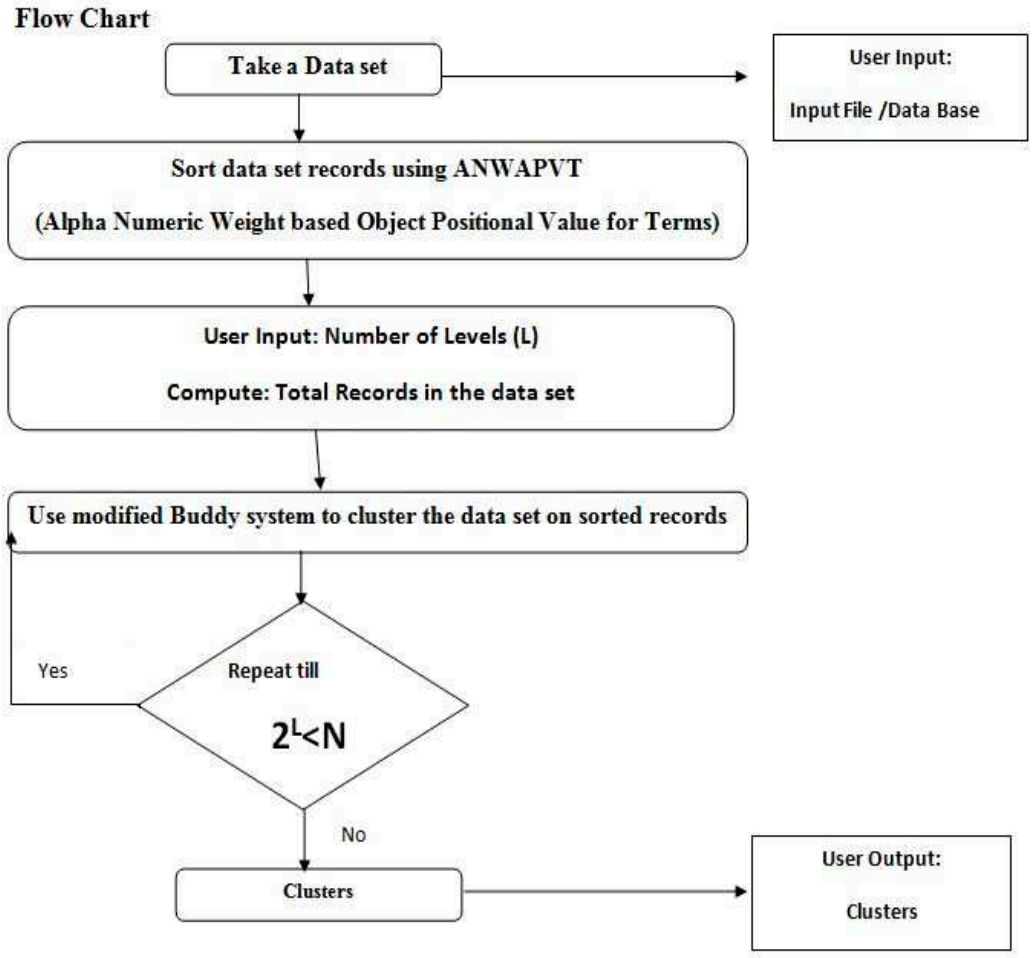

Figure 9. Buddy System Based Alpha Numeric Weight Based Clustering Algorithm with User Threshold.

Note

1. The number of elements in each cluster can be known using algorithm.
2. Data set can be collected / downloaded from freely available public repositories. i.e The data set which we have used is twitter data set.
3. At each level the numbers of clusters are equal to 2^L (Where number of levels L is required by the user.)
4. Data preprocessing techniques applied on the collected data set. This data set will be the input for the proposed Algorithm [52].
5. Clustering output will be saved in the output file / Data base.
6. Data preprocessing has to be done on the data set before clustering algorithm starts [53].
7. Data preprocessing can also be done on multiple data sources to get required data for clustering algorithm [54].
8. Formatted data is given as input for the clustering process and output is patterns [55].
9. Data mining output is the input for the clustering algorithm input.
10. Each clustering algorithm will be associated with a time complexity [56].
11. Patterns can be explored and filtered [57].
12. After data clustering the data is used for visualization and interpretation of results [58].
13. Clustering is used in Association rules [59] and classification [60].

IV. Results

Results are being generated using python.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 August 2023 doi:10.20944/preprints202308.1676.v1

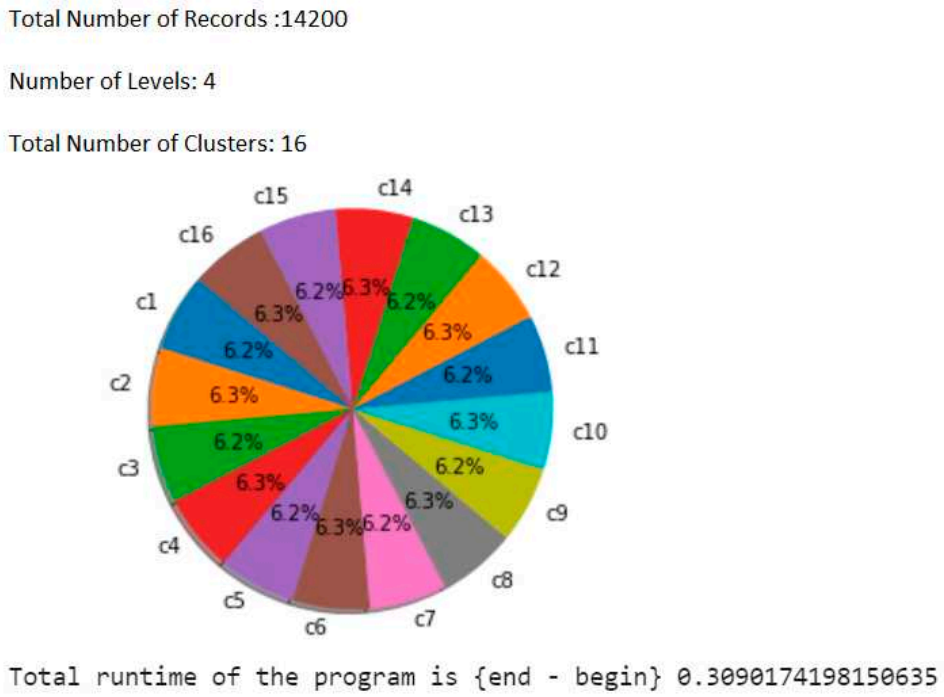

Figure 10. Clustering Result.

V Conclusion

Here we are going to implement Buddy system with Alpha numeric weighted based clustering Algorithm with user preferences to get good clusters. So the efficiency of the clustering algorithm depends on the metrics (Buddy system, Alpha numeric weight of the object with user preferences and number of levels) used in the clustering algorithm.

References

[1]. Kamalpreet Bindra and Anuranjan Mishra, "A detailed study of clustering algorithms", Digital Object Identifier 978-1-5090-3012-5/17/$31.00 ©2017 IEEE.

[2]. Uday Kumar Rai and Dr.Kanika Sharma, "An agglomerative hierarchical clustering algorithm based on global distance measurement", 978-1-4673-8302-8/15 $31.00 © 2015 IEEE, Digital Object Identifier DOI 10.1109/ITME.2015.104.

[3]. R. Sathya and Annamma Abraham, "Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification.

[4]. Sunita V. Lahane, M.U. Kharat and Prasad S. Halgaonkar, "Divisive approach of Clustering for Educational Data", 978-0-7695-4884-5/12 $26.00 © 2012 IEEE, Digital Object Identifier DOI 10.1109/ICETET.2012.55.

[5]. Masoud Makrehchi, "Hierarchical Agglomerative Clustering Using Common Neighbours Similarity", 978-1-5090-4470-2/16 $31.00 © 2016 IEEE, Digital Object Identifier DOI 10.1109/WI.2016.92.

[6]. IlungPranata and Geoff Skinner, "Segmenting and Targeting Customers through Clusters Selection & Analysis", Digital Object Identifier 978-1-5090-0363-1/15/$31.00 ©2015 IEE.

[7]. Mohiuddin Ahmed and Abdun Naser Mahmood, "A Novel Approach for Outlier Detection and Clustering Improvement", Digital Object Identifier 978-1-4673-6322-8/13/$31.00 c 2013 IEEE.

[8]. Vaishali Madaan and Rakesh Kumar, "An Improved Approach for Web Document Clustering", ISBN: 978-1-5386-4119-4/18/$31.00 ©2018 IEEE.

[9]. Huijuan Shen and Zhenjiang Duan, "Application Research of Clustering Algorithm Based on K-Means in Data Mining",978-1-7281-9837-8/20/$31.00@2020 IEEE, Digital Object Identifier 10.1109/CIBDA50819.2020.00023.

[10]. Ying Chen, Jiwon Kim and Hani S. Mahmassani, "Pattern recognition using clustering algorithm for scenario definition in traffic simulation-based decision support systems", Digital Object Identifier 978-1-4799-6078-1/14/$31.00 ©2014 IEEE

[11]. G.B. Coleman and H.C. Andrews, "Image segmentation by clustering", Digital Object Identifier 0018-9219/79/0S00-0773$00.75 0 1979 IEEE.

[12]. Manish Sharma; Durga Toshniwal, "Pre-Clustering Algorithm for Anomaly Detection and Clustering that uses variable size buckets", Digital Object Identifier 978-1-4577-0697-4/12/$26.00 ©2012 IEEE.

[13]. Yu Zhan, Haiwei Pan, Qilong Han, Xiaoqin Xie, Zhiqiang Zhang and Xiao Zhai, "Medical image clustering algorithm based on graph entropy", Digital Object Identifier 978-1-4673-7682-2/15/$31.00 ©2015 IEEE.

[14]. Manne Suneetha, S Sameen Fatima and Shaik Mohd. Zaheer Pervez, "Clustering of web search results using Suffix tree algorithm and avoidance of repetition of same images in search results using L-Point Comparison algorithm", Digital Object Identifier 978-1-4244-7926-9/11/$26.00 ©2011 IEEE.

[15]. J. Prabhu, M. Sudharshan, M. Saravanan and G. Prasad, "Augmenting Rapid Clustering Method for Social Network Analysis",978-0-7695-4138-9/10 $26.00 © 2010 IEEE, Digital Object Identifier 10.1109/ASONAM.2010.55.

[16]. Ioannis P. Panapakidis, Minas C. Alexiadis and Grigoris K. Papagiannis, "Three-Stage Clustering Procedure for Deriving the Typical Load Curves of the Electricity Consumers.

[17]. S. Iiritano and M. Ruffolo, "Managing the Knowledge Contained in Electronic Documents: a Clustering Method for Text Mining", Digital Object Identifier 1529-4188/01$10.00 0 2001 IEEE.

[18]. Sanjay Kumar Dwivedi and Bhupesh Rawat, "AReview Paper on Data Preprocessing: A Critical Phase in Web Usage Mining Process", Digital Object Identifier 978-1-4673-7910-6/15/$31.00 c 2015 IEEE.

[19]. Ratnadeep R. Deshmukh and VaishaliWangikar, "Data Cleaning: Current Approaches and Issues.

[20]. Thippa Reddy Gadekallu, Praveen Kumar Reddy, KuruvaLakshman and Rajesh Kaluri, "Analysis of Dimensionality Reduction Techniques on Big Data".

[21]. Benjamin Schelling & Claudia Plant, "Dataset-Transformation: improving clustering by enhancing the structure with DipScaling and DipTransformation".

[22]. A.Dharmarajan and T. Velmurugan, "Applications of Partition based Clustering Algorithms: A Survey", Digital Object Identifier 978-1-4799-1597-2/13/$31.00 ©2013 IEEE.

[23]. Arpita Nagpal1, Aman Jatain2 and Deepti Gaur, "Review based on data clustering algorithms", Digital Object Identifier 978-1-4673-5758-6/13/$31.00 © 2013 IEEE.

[24]. Pradeep Singh and Prateek A. Meshram, "Survey of Density Based Clustering Algorithms and its Variants", 978-1-5386-4031-9/17/$31.00 ©2017 IEEE.

[25]. AminehAmini, Teh Ying Wah, Mahmoud Reza Saybani and Saeed Reza AghabozorgiSahafYazdi, "A study of density-grid based clustering algorithms on data streams", Digital Object Identifier 978-1-61284-181-6/11/$26.00 ©2011 IEEE.

[26]. Yingjie Tian and Dongkuan Xu, "A Comprehensive Survey of Clustering Algorithms.

[27]. Srutipragayn Swain and Manoj Kumar DasMohapatra, "A review paper on soft computing based clustering algorithm.

[28]. Sara C. Madeira and Arlindo L. Oliveira, "Biclustering Algorithms for Biological Data Analysis: A Survey", 1545-5963/04/$20.00 2004 IEEE.

[29]. Zhou Mingqiang, Huang Hui and Wang Qian, "A graph-based clustering algorithm for anomaly intrusion detection", Digital Object Identifier 978-1-4673-0242-5/12/$31.00 ©2012 IEEE.

[30]. Christina J.and K. Komathy. "Analysis of hard clustering algorithms applicable to regionalization." 2013 IEEE conference on information & communication technologies. IEEE, 2013, 978-1-5090-3012-5/17/$31.00 ©2017 IEEE.

[31]Goran Andonovski and Igor Škrjanc, "Evolving clustering algorithm based on average cluster distance CAD" Conference Larnaca,Cyprus, DOI: 10.1109/EAIS51927.2022.9787746.

[32]. Ming Lei, Zhen-Hua Ling, Li-Rong Dai, "Minimum Generation Error Training With Weighted Euclidean Distance On Lsp For Hmm-Based Speech Synthesis", 978-1-4244-4296-6/10/$25.00 ©2010 IEEE.

[33]. Yun Yang , Shi-ze Guo, Gu-yu Hu and Hua-bo Li4, "An improved Hybrid P2P Control Model Based on Chord", 978-0-7695-4935-4/12 $26.00 © 2012 IEEE.

[34]. H. M. Abdul Fattah, MD Masum Al Masba and K. M Azharrul Hasan, "Sentiment Clustering By Mahalanobis Distance", 978-1-5386-8279-1/18/$31.00@2018 IEEE.

[35]. Zann Koh, Yuren Zhou, Billy PikLik Lau, Ran Liu, Keng Hua Chong and Chau Yuen , "Clustering and Analysis of GPS Trajectory Data Using Distance-Based Features", DOI: 10.1109/ACCESS.2022.3225646.

[36]. Hongyan Cui1, Kuo Zhang, Yajun Fang, Stanislav Sobolevsky, Carlo Ratti And Berthold K. P. Horn, "A Clustering Validity Index Basedon Pairing Frequency ", DOI: 10.1109/ACCESS.2022.3225646.

[37]. Mostafa Raeisi and Abu B. Sesay, Carlo Ratti And Berthold K. P. Horn, "A Distance Metric for Uneven Clusters of Unsupervised K-Means Clustering Algorithm ", DOI: 10.1109/ACCESS.2022.3198992.

[38]. Chun Guan , Kevin Kam Fung Yuen , Frans Coenen, "Particle swarm Optimized Density-based Clustering and Classification: Supervised and unsupervised learning approaches", DOI: https://doi.org/10.1016/j.swevo.2018.09.008.

[39]. Fei xue Huang and Xin Zhao Cheng Li, Frans Coenen, "Clustering Effect of Style based on Pearson correlation", 978-1-4244-5143-2/10/$26.00 ©2010 IEEE.

[40]. L. Svetlova, Moscow, B. Mirkin, Birkbeck and H. Lei, University of Texas, "MFWK-Means: Minkowski Metric Fuzzy Weighted K-Means for high dimensional data clustering", 978-1-4799-1050-2/13/$31.00 ©2013 IEEE.

[41]. H Roopa and T Asha, "Segmentation of X-ray image using city block distance measure", 978-1-5090-5240-0/16/$31.00 ©2016 IEEE.

[42]. Noureddine Bouhmala, "How Good Is The Euclidean Distance Metric For The Clustering Problem", 978-1-4673-8985-3/16 $31.00 © 2016 IEEE, DOI 10.1109/IIAI-AAI.2016.26.

[43]. Allan de M. Martins, Adrilo D. D. Neto and Jos Alfred0 F. Costa, "Clustering using neural networks and Kullback-Leibler divergency", Digital Object Identifier 0-7803-8359-1/04/$20.00 02004 IEEE.

[44]. Shuhua Ren and Alin Fan, "K-means clustering algorithm based on coefficient of variation", Digital Object Identifier 978-1-4244-9306-7/11/$26.00 ©2011 IEEE.

[45]. Shraddha K. Popat, Pramod B. Deshmukh and Vishakha A. Metre, "Hierarchical document clustering based on cosine similarity measure", Digital Object Identifier 978-1-5090-4264-7/17/$31.00 ©2017 IEEE.

[46]. Dianwei Chi, "Research on the Application of K-Means Clustering Algorithm in Student Achievement", Digital Object Identifier 978-1 -7281 -8319-0/21/$31.00 ©2021 IEEE.

[47]. Jaskaranjit Kaur and Harpreet Singh, "Performance Evaluation of a Novel Hybrid Clustering Algorithm using Birch and K-Means", Digital Object Identifier 978-1-4673-6540-6/15/$31.00 ©20 15 IEEE.

[48]. Uday Kumar Rai and Dr.Kanika Sharma, "Maximum Likelihood Estimation based Clustering Algorithm on Wireless Sensor Network-A Review", Digital Object Identifier 978-1-5386-1887-5/17/$31.00 ©2017 IEEE.

[49]. Rajendra Pamula, Jatindra Kumar Deka and Sukumar Nandi, "An Outlier Detection Method based on Clustering", Digital Object Identifier 978-0-7695-4329-1/11 $26.00 © 2011 IEEE.

[50] "Clustering algorithms and validity measures", M. Halkidi, Y. Batistakis and M. Vazirgiannis.

[51] "A High-Performance Memory Allocator for Object-Oriented Systems", J. Morris Chang and Edward F. Gehringer.

[52] "Implementation of Data Preprocessing Techniques on Distributed Big Data Platforms", Oguzcelik, muruvvetHasanbasoglu and Mehmet S.Aktas.

[53] "An Analytical approach for Data Preprocessing", P.SreenivasandDr.C.V.Srikrishna.

[54] "ETL Preprocessing with Multiple Data Sources for Academic Data Analysis", Gant GawWuttMhonandNang Saing Moon Kham.

[55] "Web Mining: Information and Pattern Discovery on the World Wide WebR. Cooley, B. Mobasher and J. Srivastava.

[56] "Comparative study of Data Mining Clustering algorithms.", IyerAurobindVenkatkumar and Sanat kumar Jayanti bhai Kondhol Shardaben.

[57] "Investigating and reflecting on the integration of automatic data analysis and visualization in knowledge discovery", Enrico Bertini and Denis Lalanne

[58]. "Comparative study of Data Mining Clustering algorithms."Qi Luo.

[59]. "Association Rules for Clustering Algorithms for Data Mining of Temporal Power Ramp Balance",NursedaYildirim and BahriUzunoglu

[60]. "A study on classification techniques in data mining", G.Kesavaraj and Dr.S.Sukumaran.

 doi:10.20944/preprints202308.1676.v1

AUTHOR DETAILS:

Dr. Srikanth Thota received his Ph.D in Computer Science Engineering for his research work in Collaborative Filtering based Recommender Systems from J.N.T.U, Kakinada. He received M.Tech. Degree in Computer Science and Technology from Andhra University. He is presently working as an Associate Professor in the department of Computer Science and Engineering, School of Technology, GITAM University, Visakhapatnam, Andhra Pradesh, India. His areas of interest include Machine learning, Artificial intelligence, Data Mining, Recommender Systems, Soft computing.

Mr. MaradanaDurgaVenkata Prasad received his B.TECH (Computer Science and Information Technology) in 2008 from JNTU, Hyderabad and M.Tech. (Software Engineering) in 2010 from Jawaharlal Nehru Technological University, Kakinada, He is a Research Scholar with Regd No: 1260316406 in the department of Computer Science and Engineering, Gandhi Institute Of Technology And Management (GITAM) Visakhapatnam, Andhra Pradesh, India. His Research interests include Clustering in Data Mining, Big Data Analytics, and Artificial Intelligence. He is currently working as an Assistant Professor in Department of Computer Science Engineering, CMR Institute of Technology, Ranga Reddy, India.