International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 6s
DOI: https://doi.org/10.17762/ijritcc.v11i6s.6946
Article Received: 28 March 2023 Revised: 15 May 2023 Accepted: 25 May 2023

Some Clustering Methods, Algorithms and their Applications

S. Velunachiyar[1], Dr. K. Sivakumar[2]
[1]Research Scholar,
Department of mathematics,
Saveetha School of Engineering,
Saveetha Institute of Medical and Technical Sciences (SIMATS),
Saveetha University,
Chennai, Tamilnadu-602105.
velunachiyars41016.sse@saveetha.com
velsam83@gmail.com
https://orcid.org/0000-0001-7609-3627
[2]Professor,
Department of Mathematics,
Saveetha School of Engineering,
Saveetha Institute of Medical and Technical Sciences (SIMATS),
Saveetha University,
Chennai, Tamilnadu-602105.
sivakumarkaliappan.sse@saveetha.com

Abstract:

Clustering is a type of unsupervised learning [15]. When no target values are known, or "supervisors," in an unsupervised learning task, the purpose is to produce training data from the inputs themselves. Data mining and machine learning would be useless without clustering. If you utilize it to categorize your datasets according to their similarities, you'll be able to predict user behavior more accurately. The purpose of this research is to compare and contrast three widely-used data-clustering methods. Clustering techniques include partitioning, hierarchy, density, grid, and fuzzy clustering. Machine learning, data mining, pattern recognition, image analysis, and bioinformatics are just a few of the many fields where clustering is utilized as an analytical technique. In addition to defining the various algorithms, specialized forms of cluster analysis, linking methods, and please offer a review of the clustering techniques used in the big data setting.

Keywords: Clustering, Partitioning Clustering, Hierarchical Clustering, Density-based Clustering, Grid-based Clustering, Fuzzy Clustering.

I. INTRODUCTION

Grouping things of a similar nature together facilitates data analysis [1]. Clustering algorithms are used to break down massive datasets into more manageable chunks by identifying and separating out collections of things that share commonalities. It's a method for making more efficient use of unlabeled data. A method of machine learning called clustering is used to find groups of data that have more features in common than the original dataset. Algorithms that cluster information into useful groups use a wide range of measures of similarity and dissimilarity to group data into relevant categories. Scalability, the ability to work with many attribute types, the discovery of clusters of variable shape, domain expertise in selecting input parameters, robustness to noise, and user-friendliness are among desirable qualities in a strong clustering technique. In cluster analysis, items in the data are categorized simply by the information contained inside the data itself. The objective is to achieve greater internal consistency while assessing external variations. There is no one algorithm for performing cluster analysis; rather, it is a broad problem that can only be solved by a systematic, iterative procedure of learning from experience. Examples of Clustering Techniques include Partitioning, Hierarchical, Density, Grid, Model-Based, and Fuzzy Methods. Cluster analysis has several practical uses, including but not limited to those listed below: market research, pattern identification, data analysis, and image processing. Clustering using K-Means is common practice [2]. Clustering's ability to help uncover interesting patterns and structures in massive data sets with little to no prior knowledge is a major benefit [15].

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 6s
DOI: https://doi.org/10.17762/ijritcc.v11i6s.6946
***Article Received**: 28 March 2023 **Revised**: 15 May 2023 **Accepted**: 25 May 2023*

CLUSTERING

A technique for organizing related data points into distinct groups. The objects that have the most similarities stay together in one group while the other groups share less or none at all.

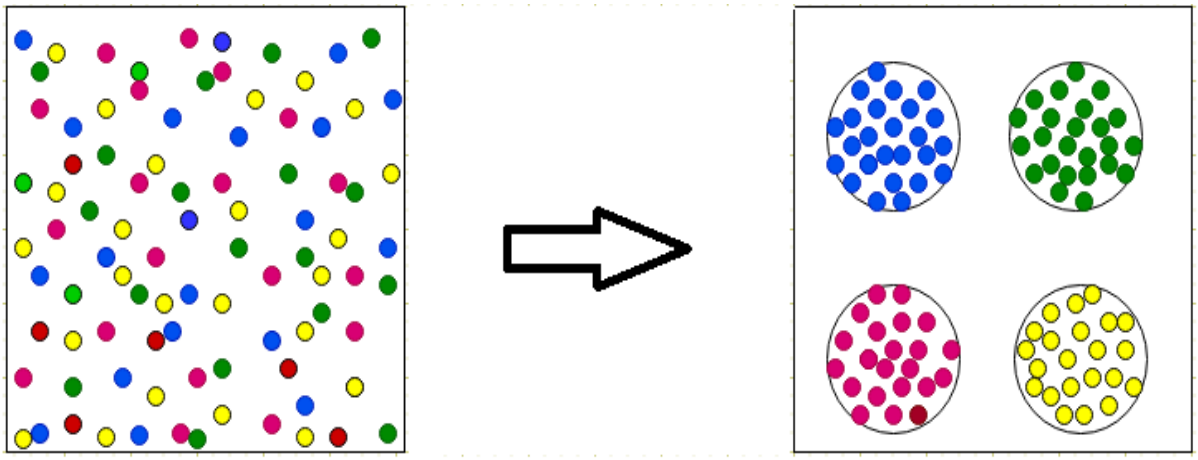

Figure:1 Before Cluster After Cluster

TYPES OF CLUSTERING ALGORITHMS

Multiple clustering methods can be distinguished. Partitioning-based clustering, hierarchical-based clustering, grid-based clustering, model-based clustering, and fuzzy-based clustering are all discussed in this article, along with their respective methods of construction. There is also a summary of the most popular algorithms

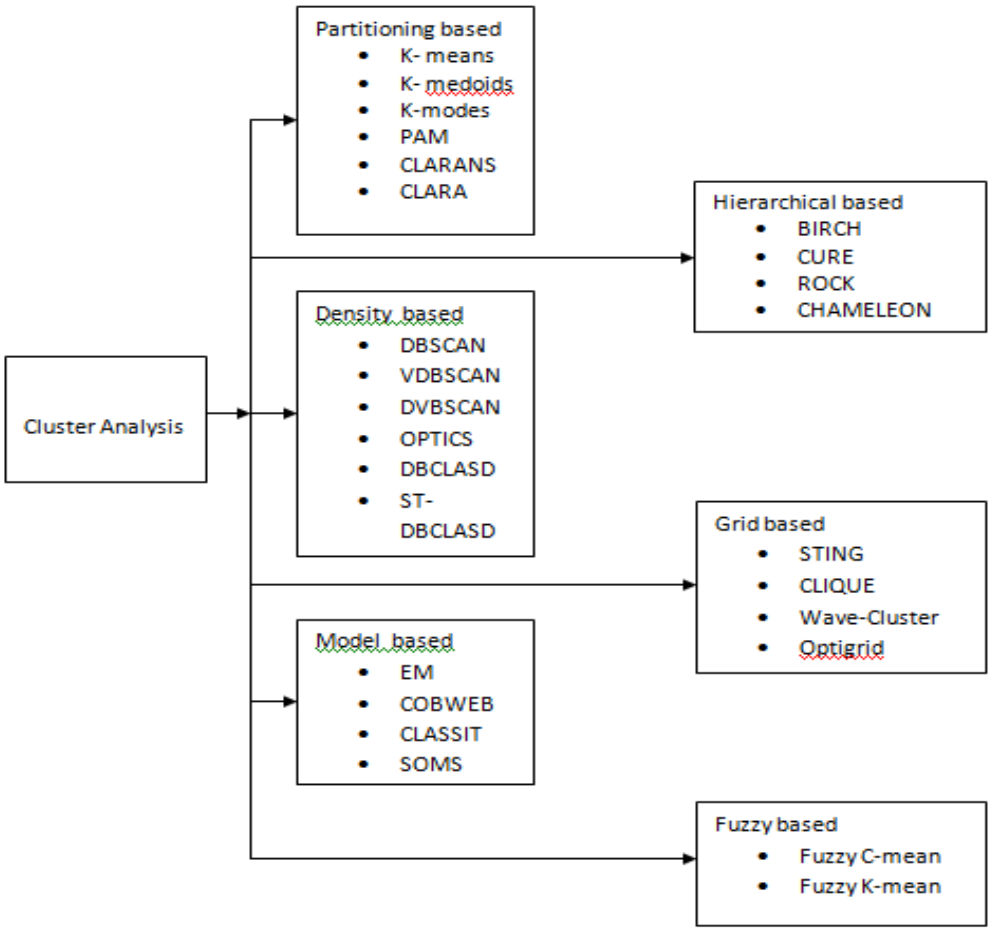

Figure 2: Types of Clustering and its Algorithms

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 6s
DOI: https://doi.org/10.17762/ijritcc.v11i6s.6946
***Article Received**: 28 March 2023 **Revised**: 15 May 2023 **Accepted**: 25 May 2023*

PARTITIONING METHOD

It's a method of cluster analysis that divides information into flat, mutually exclusive categories. In other words, it's a variation on the centroid method. After that, the data set is split into K roughly equivalent pieces. The cluster's center is built to ensure that, within each cluster, the distances between data points are minimized relative to those in neighboring clusters. Although centroid-based algorithms perform well, they are vulnerable to noise and outliers. Partitioning-based clustering methods include K-means, K-medoids, K-modes, and CLARANS (Randomized Search Clustering Algorithm).

- **K-means** is an outstanding, simple, and effective method for clustering data. K-means is a nondeterministic, iterative, numerical, unsupervised method. This method's speed, simplicity, and ability to produce accurate clustering results have made it a popular choice for a wide variety of practical uses. On the other hand, it works wonderfully for producing globular clusters. Academics have made several attempts to boost the k-means algorithm's effectiveness. To make sense of enormous data sets, K-means clustering is often utilised. [24]. K-Means is computationally faster than hierarchical clustering when there are many variables (if K is small). The quality of the resulting clusters is difficult to evaluate objectively. Clusters of different densities and sizes will form depending on the value of k at the outset. The k-center problem is extremely difficult (NP-hard). Given a predetermined number of clusters, determining an appropriate value for k can be challenging. It fails miserably when used to clusters that aren't spherical. Problems can also arise from outliers.

HIERARCHICAL METHOD

In the hierarchical clustering technique known as connectivity-based clustering, it is assumed that objects are connected to their nearest neighbors. Dendrograms are used to create a hierarchical clustering of datasets. The degree of similarity or dissimilarity between any two groups can be seen in a dendrogram [21]. Both Agglomerative Nesting (AGNES) and Divisive Analysis [18] are examples of hierarchical clustering procedures (DIANA). When using an agglomerative method, each observation is initially placed in its own cluster, and as one moves up the hierarchy, clusters with similar characteristics coalesce. All observations are initially placed in one cluster before being divided into smaller groups recursively as the hierarchy is traversed. Typically, a cluster proximity matrix is used to decide whether or not to split or merge a set of nodes. Different types of Hierarchical-based Clustering include CURE (Clustering Using Representatives), BIRCH (Balanced Iterative Reducing and Clustering Using Hierarchies) [57], ROCK (Robust, Clustering Using 1inKs)[44], and CHAMELEON[45].

- The **CURE** Clustering algorithm can cluster datasets of extremely large sizes and recognizes clusters of no spherical shapes with considerable variance. The programme picks out widely spaced points from the dataset and reduces their sizes until they cluster together. It uses a set of points rather than a single centroid to locate groups of similar non-spherical shapes, and it reduces the significance of outliers by moving the points closer together in the center [23]. CURE and ROCK are two examples of agglomerative hierarchical clustering approaches that employ a static model to decide which hierarchical clusters are most similar to one another and should be merged. CURE assesses the similarity of the nearest pair of representative points without taking into account the internal closeness (density or homogeneity) of the two clusters concerned. Instead of individually analyzing each cluster's connectedness within the same data set, ROCK looks at how well linked the two clusters are together using a static interconnectivity model chosen by the user.

DENSITY-BASED CLUSTERING

As a kind of data mining, density-based clustering may help uncover hidden relationships within datasets. This programme accomplishes this by finding different clusters within the dataset and connecting the high-density regions into clusters. Sparse areas separate the moredense portions of the data space. Several methods for density-based clustering have already been used, including DBSCAN (Density Based Spatial Clustering of Applications with Noise) [56], OPTICS (Ordering Points to Identify the Clustering Structure), DENCLUE (Density-based Clustering), VDBSCAN (Varied Density Based Spatial Clustering of Applications with Noise), and DVBSCAN ((Density Variation Based Spatial Clustering of Applications with Noise)). DBSCAN is a method that uses density-based clustering [56].

- Inputs required by the **DBSCAN** algorithm are: The threshold for cluster membership is defined by the required distance between points (in eps). These points are considered to be close by (eps) if their separation is less than or equal to this measurement. We may define a dense zone with a minimum of midpoints. By way of illustration, if midpoint is set to 5, a dense zone must have at least that many points. The term "density" refers to the concentration of points inside a certain area. In graph theory, a "core point" is a node that is surrounded by an invariably large number of other nodes. You can think of these as the nuclei of groups of things. Although a boundary point is quite close to a core point, it lacks the necessary minimum number of points inside the specified radius. Any place that is neither central nor peripheral is considered noise. The most popular density modelling methods are DBSCAN and OPTICS.

GRID-BASED METHOD

Each clustering approach uses grids; however, the grid resolutions are not always the same. All clustering methods require a grid layout in which the object areas are quantized into a predetermined number of cells. This method is fast because it can process data rapidly, despite the fact that the quantized space requires a significant number of cells in each dimension. CLIQUE (Clustering in Quest), STING (Statistical Information Grid), and Wave Cluster [18] are all examples of Grid-based clustering algorithms. CLIQUE is able to locate subspaces inside high-dimensional data that are more suitable for clustering. CLIQUE is a strategy that can be seen in either a dense or grid-based framework. Then, a set number of equally spaced intervals are created throughout all dimensions. Particularly, non-contiguous rectangular cells divide the m-dimensional data

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 6s
DOI: https://doi.org/10.17762/ijritcc.v11i6s.6946
***Article Received**: 28 March 2023 **Revised**: 15 May 2023 **Accepted**: 25 May 2023*

space. A dense unit contains more data points than the output model parameter. Clusters are collections of objects in a subspace that share the most connections among themselves. When STING is running, it divides the globe into squares. There's a step-by-step increase in complexity as you move down the cell levels. At the subsequent level below, each cell from the previous level is split into multiple smaller cells. It is straightforward to calculate the statistics of any higher-level cell by using the parameters of the base cell.

- **Wave Cluster** is an algorithm-based clustering method. Through the use of hat-shaped filters that place more focus on areas where points cluster while suppressing weak information at their borders, it is possible to effectively remove outliers while also achieving multi-resolution, low cost, and efficiency. One method of cluster analysis is called Wave Cluster [31], and it works by imposing a regular two-dimensional grid on the data and then counting how many data points there are in each cell. In this way, the data points are transformed into a series of grayscale values, which is subsequently interpreted as an image. In order to make use of the multi-scaling and noise-reduction capabilities of wavelets, The clustering issue is rephrased as an image segmentation problem. The basic strategy involves establishing a data grid with each data point occupying a cell, processing the data using the wavelet transform, finding related clusters using the average sub-image, and mapping the clusters back to the original space's points. [32].

DISTRIBUTION MODEL-BASED METHOD

The data may be represented by a broad set of distributions, and this clustering method is based on that premise. The number of groups is commonly determined by using data from a normal distribution. The algorithm will not use a data point if it cannot be confidently linked to one of the two inputs. In order to trace the origin of these data points, statistical analysis plays a crucial role. Statistical approaches to Model-based Clustering Algorithms include EM(Expectation-Maximization) and Auto class. Model-based algorithm machine learning techniques such as COBWEB and CLASSIT. The Neural Network technique of Model-based algorithm known as SOM (Self-Organizing Feature Map) is used.

- **COBWEB** is an easy-to-use and widely-recognized approach to gradual conceptual development. Makes a hierarchical grouping like a tree used for categorization. An concept is represented as a node in the network, and each node is accompanied by a probabilistic justification of how it came to be. COBWEB employs the heuristic evaluation measure of category utility to direct tree construction. In order to optimize category utility, it gradually adds objects to a categorization tree. The capability of COBWEB to generate a new class based on an attribute is a key differentiator between it and the K-means approach. COBWEB can execute a two-way search since it allows for the merging and splitting of classes based on their usefulness. It's possible that COBWEB is highly vulnerable to out-of-the-ordinary data. Statistically speaking, the probability distributions for various qualities can be treated as separate entities. The time and space complexity increases when there are more attributes and more values for each attribute, and this is especially true when the attribute values are numerous.

FUZZY CLUSTERING METHOD

The subject of pattern recognition has seen significant growth in the application of fuzzy clustering, a subset of cluster analysis. Only by using fuzzy clustering can groupings be discovered in data that spans the world. Data items' similarities inside a cluster are emphasized, while those between clusters are downplayed, as in a fuzzy clustering approach. The goal function may be computed by using the information provided by the data matrix, the membership matrix, and the prototype clusters. It is a metric for evaluating the level of dissimilarity between data objects contained within the same cluster. Thus, we may find the optimal data-splitting strategy by minimizing the objective function. If you're using fuzzy clustering, there's more than one way to classify a single data point into a cluster. The membership coefficient of a dataset is calculated based on how well it fits into a predefined cluster [25].

- In the **Fuzzy C-Means** Clustering Algorithm, each data point is partitioned into numerous clusters and assigned a membership degree based on its distance from the cluster center. Clustering in this sense includes the Fuzzy C-means method (also known as the Fuzzy k-means algorithm).

SPECIALIZED TYPES OF CLUSTER ANALYSIS

Specific cluster analyses include high-dimensional data clustering, conceptual clustering, consensus clustering, limited clustering, data stream clustering, sequence clustering, and spectral clustering [59].

Data with many qualities, or high-dimensional data, offers unique challenges when attempting to cluster it. The 'curse of dimensionality,' first devised to represent the general rise in complexity of many computer activities as dimensionality develops, is widely acknowledged as the root cause of the inefficiency of traditional clustering algorithms. According to the curse of dimensionality, discriminating between objects becomes more difficult as their number of dimensions grows. New grouping methods are needed since high-dimensional objects look so similar. Thus, in recent years, a lot of work has gone into developing strategies and algorithms for grouping together information in high dimensions. Unanswered questions remain. The process of clustering is a data mining method for automatically grouping data into meaningful categories based on their common features. Similar objects are grouped together into a single cluster, whereas those with significant differences are placed in their own clusters. Without the use of predetermined class labels, clusters can provide a meaningful description of the underlying data structure. There are multiple clustering paradigms that offer various cluster models and computational ways for cluster detection. One thing that all methods have in common is the need for an underlying assessment of similarity between data objects [36].

In the 1980s, a machine learning paradigm called Conceptual Clustering [26] was created for unsupervised categorization. Using conceptual clustering techniques, it is possible to develop a structured hierarchy of categories. A number of other methods, including decision tree learning and mixture model learning, bear parallels to formal concept analysis. The COBWEB system is an

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 6s
DOI: https://doi.org/10.17762/ijritcc.v11i6s.6946
***Article Received**: 28 March 2023 **Revised**: 15 May 2023 **Accepted**: 25 May 2023*

incremental, hierarchical clustering of ideas. Douglas H. Fisher, a professor at Vanderbilt, is credited with developing COBWEB. COBWEB builds a hierarchy of categories out of observational data. A classification tree's label is a probability idea that summarizes the distributions of attribute values for objects in that class. Each node represents a class in the classification tree. An object's class may be inferred using this tree, and any missing attributes can be found.

Consensus clustering aims to combine multiple clusters into a single, more robust cluster than the two that were originally created. By creating a Consensus Matrix at each level, this method makes it possible to safely combine all subclusters into a single supercluster. The ability to individually extract specific partitions, accurate cluster size generation, and improved handling of missing data are all positive aspects.

Constrained clustering, a semi-supervised method for data clustering, includes domain expertise via constraints. Typically, paired statements that define whether or not two things must be in the same cluster make up restrictions. While some constrained clustering algorithms may choose to strictly enforce all constraints in the solution, others may treat the restrictions more as suggestions than requirements [35].

Clustering of data streams is not a novel concept in the fields of data science and statistics. New practical obstacles have emerged in this subject, however, due to the widespread usage of the Internet of Things (IoT), which researchers are working to address. These challenges fall into four classes: high dimensionality, rapid data flow, real-time limitations, and dynamic character. However, there are still caveats to these density-based algorithms for data stream clustering despite their recent proliferation. When using a distance function, the quality of these established clustering techniques drops dramatically [29].

Data mining method that divides a large set of sequences into groups of similar sequences is called sequence clustering [33]. In genomics, an adenine, guanine, cytosine, or thymidine sequence is called a sequence. The fact that this method can be applied to genetic sequences accounts for its extensive adoption in bioinformatics. Due to the fact that the states in the generated sequence models have associated transition probabilities, sequence clustering is likewise a probabilistic approach. Because a low probability of transition denotes a sequence path that is seldom taken and hence likely to be wrong, these models are noise-resistant. This algorithm seems to be a good fit for mining action sequences, which can be collected from event logs, as it accepts sequences as input.

Spectral clustering's adaptability means that it can be used with non-graphical data as well. It makes no assumptions about the underlying cluster geometry. K-Means is only one of several clustering algorithms that makes the assumption that clustered data points are all located on a sphere with their center at the cluster's first node. This is a generalization that may not hold true in all situations. Spectral clustering helps make more precise groups in these situations. It is able to properly group together observations that should be in the same cluster but are split from one another by dimension reduction. The data points in a Spectral Clustering should be linked together, but the borders between them need not be convex. Any spectrum clustering approach may be broken down into three distinct phases. The generation of a similarity matrix constitutes the initial preprocessing step. To perform the second Spectral mapping, Eigen Vectors for the similarity matrix must be generated. Finally, Post-Processing is focused on data clustering. The following are some advantages of spectral clustering algorithms: Avoiding making sweeping assumptions about cluster form, having an aim that ignores local optimums, being statistically consistent, and having a faster run-time are all positives [28].

Techniques used in cluster analysis

Cluster analysis makes use of several methods like Artificial neural networks (ANN), nearest neighbor searches, neighborhood component analyses, and latent class analyses.

SOME APPLICATIONS OF CLUSTERING ALGORITHMS

Clustering is a quick and easy way to perform a high-level analysis of unstructured data. Numerous factors, such as data point density, graphing methods, and the shortest distance, can affect the formation of clusters. By utilizing a metric known as the similarity measure, clustering is able to determine the level of similarity between the objects. To find similarity measurements, you need fewer features. In general, it becomes more challenging to create similarity assessments as the number of features increases.

- A key part of marketing is segmenting your clientele into subsets based on their shared preferences. based on extensive client information including addresses and purchase history;
- Forecasting the stock market and currency, the exchange rate, bank failure, financial risk, futures trading, credit rating forecasting, and credit rating forecasting are all examples of financial tasks that can be predicted.
- Motor insurance policyholders with a high average claim cost may be identified by: frauds; the classifying of plants and animals in biology; the ordering of books in libraries; the detection of insurance fraud in the insurance industry.
- Clustering epicenters of earthquakes in order to pinpoint hazardous areas is a common technique in both city planning and earthquake research.
- Classification of documents on the World Wide Web; clustering of web log data to discover communities of users with similar browsing habits.

METHODS OF LINKING

Commonly known as Linkage Methods, these techniques use distance measurements between clusters to derive clustering rules. The terms "Complete-connection," "Single-linkage," "Average-linkage," and "Centimeter-linkage" all refer to popular types of linkage. While merging clusters, complete-linkage finds the maximum distance between them and uses that as the threshold for merging. When merging clusters, single-linkage finds the path with the fewest hops. This connection might help you spot potentially abnormally high values in your dataset before merging them. Before any merger takes place, average-linkage measures how far apart each cluster typically is. Finally, Centroid-linkage finds the centers of both clusters, then merges them based on the distance between them.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 6s
DOI: https://doi.org/10.17762/ijritcc.v11i6s.6946
Article Received: 28 March 2023 Revised: 15 May 2023 Accepted: 25 May 2023

$$\text{Single linkage: } D_s = \min_{i,j}\{||X_i - X_j||\} \quad ; \quad \text{Complete linkage: } D_{co} = \max_{i,j}\{||X_i - X_j||\}$$

$$\text{Average linkage: } D_a = \sum_{i,j}||X_i - Xj|| \,/\, N_k N_l \quad ; \quad \text{Centroid linkage: } D_{ce} = ||C_k - C_l||$$

CLUSTERING ALGORITHM, DEFINED IN THE CONTEXT OF BIG DATA:

Algorithms are detailed in detail in Table 1, and the 3Vs of Big data (volume, variety, velocity, and value) [56] are used to classify the data in each of the five groups.

When we talk about "volume," we're referring to our capacity to group together massive datasets. The following factors are used to determine a suitable algorithm in terms of volume: i) Dataset size, ii) High dimensionality, and iii) Noisy data with an outlier

"Variety" means that there are a wide range of data items available. The following factors are utilized to identify an acceptable algorithm in terms of Variety: I The Dataset's Genre; and ii) The Form of the Clusters.

A clustering method's speed is measured by how quickly it can sort through data. Consider the following important criteria when picking an effective clustering algorithm for velocity: (i)Time's Complicated Nature.

TABLE-I(BELOW)

TYPES	ALGORITHMS	VOLUME			VARIETY		VELOCITY
		Size of Dataset	High Dimensionality	Noisy Data	Types of Dataset	Clusters Shape	Time Complexity
Partitioning-based[56]	K-means[37]	Large &Small	No	No	Numerical	Non-Convex	O(nkd)
	K-modes[38]	Large	Yes	No	Categorical	Non-Convex	O(n)
	K-medoids[39]	Small	Yes	Yes	Categorical	Non-Convex	$O(n^2dt)$
	PAM[40]	Small	No	No	Numerical	Non-Convex	$O(k(n-k)^2)$
	CLARA[41]	Large	No	No	Numerical	Non-Convex	$O(k(40+k)^2+K(n-k))$
Hierarchical-based[56]	BIRCH[42]	Large	No	No	Numerical	Non-Convex	O(n)
	CURE[43]	Large	Yes	Yes	Numerical	Arbitrary	$O(n^2\log n)$
	ROCK[44]	Large	No	No	Categorical	Arbitrary	$O(n^2+nmmma+n^2\log n)$
	Chameleon[45]	Large	Yes	No	All type	Arbitrary	$O(n^2)$
Density-Based[56]	DBSCAN[46]	Large	No	No	Numerical	Arbitrary	O(n logn)
	OPTICS[47]	Large	No	Yes	Numerical	Arbitrary	O(n logn)
	DENCLUE[48]	Large	Yes	Yes	Numerical	Arbitrary	O(logIDI)
Grid-based[56]	WaveCluster[49]	Large	No	Yes	Spatial	Arbitrary	O(n)
	STING[50]	Large	No	Yes	Spatial	Arbitrary	O(k)
	CLIQUE[51]	Large	Yes	No	Numerical	Arbitrary	O(ck+mk)
	OPTIGRID[52]	Large	Yes	Yes	Spatial	Arbitrary	O(nd); O(ndlogn)
Model-based[56]	EM[53]	Large	Yes	No	Spatial	Non-Convex	O(knp)
	COBWEB[54]	Small	No	No	Numerical	Non-Convex	$O(n^2)$
	SOM[55]	Small	Yes	No	Multivariate	Non-Convex	$O(n^2 m)$

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 6s
DOI: https://doi.org/10.17762/ijritcc.v11i6s.6946
***Article Received**: 28 March 2023 **Revised**: 15 May 2023 **Accepted**: 25 May 2023*

II. LITERATURE WORK

GATH AND A. B. GEVA (1989), In this post, we outline the methodology for performing fuzzy classification without presuming a priori how many data clusters there are. The reliability of a cluster is assessed with the use of performance indicators and hypervolume and density standards. Fuzzy maximum likelihood estimate and fuzzy K-means are combined in this novel approach (FMLE). UFP-ONC (unsupervised fuzzy partition-optimal number of classes) works well when there is a lot of variation in the cluster shape, density, and the number of data points in each cluster. It has been the subject of study using both real-world and simulated data [14].

Krishna, K., et al (1999), This study presents a novel hybrid genetic algorithm (GA) for determining the best way to divide up data among a certain number of categories. In order to create viable offspring from the parent chromosomes, GAs used for clustering need either an expensive crossover operator or an expensive fitness function, or both. Instead of using crossover, the K-means operator (a component of the K-means method) is used in GKA. Additionally, they describe a biased distance-based mutation operator that is optimized for clustering. They use the theory of finite Markov chains to show that the GKA converges to the global best solution. Simulations show that GKA converges to the well-known optimum for the given data as expected. Furthermore, it has been discovered that GKA searches for clusters faster than other evolutionary algorithms [3].

Juha Vesanto et al (2000) , During the discovery phase of data mining, the self-organizing map (SOM) is a useful tool for gaining first insights. It maps the input space onto representations of a regular grid in low dimensions in order to efficiently explore and analyze the characteristics of the data. When the number of SOM units is high, it is necessary for statistical analysis of the map and data to group similar SOM units together. A variety of SOM clustering strategies are covered here. Focus is placed on hierarchical agglomerative clustering and -means partitive clustering. When compared to direct clustering of the data, it has been shown that using SOM to develop prototypes, which are then clustered, is more effective and computationally efficient [22].

Alauddin Yousif Al-Omary a* (2006), In this study, the authors offer a clustering algorithm system, a machine-learning method that relies on clustering (CAS). There are tests done on the CAS algorithm to see how well it works and how much of an improvement there is. Some heuristics were offered to aid machine learning authors in producing higher quality work. We analyze the CAS algorithm using the Ministry of Civil Service's InfoBase. In comparison to UNIMEM, COBWEB, and CLASSIT, the CAS algorithm is determined to be the best machine-learning algorithm. The suggested method is a combination of two separate machine learning approaches. Researching previously solved situations is the first method. Exceptions and multiple inheritance are both supported by CAS. Common probability assumptions made during ideation are likewise avoided by CAS. The second strategy is learning through watching. In order to effectively cluster concepts, Conceptual Similarity Analysis (CAS) employs a set of operators that have been shown to be efficient in this setting. They demonstrated how CAS constructs and queries a clustering hierarchy to decide whether or not an item should be included or characterized [34].

Hong Changa and colleagues (2007), They created a technique for path-based spectral clustering that is more noise- and outlier-resistant than each technique by itself. In order to give a consistent path-based spectral clustering approach, the authors of this paper present a robust path-based similarity measure for spectral clustering in both unsupervised and semi-supervised contexts. This strategy relies on the M-estimation technique from robust statistics. They conducted trials using both fictitious and real-world data and compared our strategy to others. The picture segmentation study specifically makes use of color photographs from the Berkeley segmentation data set and benchmark. Their method outperforms other methods in experiments because it is more reliable [21].

Abu Abbas Osama Mahmoud et al (2008), The purpose of this research was to evaluate and contrast various methods of clustering data. Specifically, we're looking at the Expectation Maximization (EM) Clustering algorithm, the Self-Organizing Maps (SOMs) algorithm, the Hierarchical Clustering technique, and the K-means algorithm. Algorithms are compared with respect to data set size, cluster count, data set type, and software utilized. Certain inferences can be made about the efficacy, dependability, and accuracy of the clustering algorithms [4].

Jinghua Zhao, Wenbo Zhang, et.al (2010). In order to better meet the needs of telecom firms, an improved K-Means algorithm for consumer segmentation is presented in this study. Last but not least, the outcomes of the segmentation were assessed after using the sophisticated K-Means approach to the cluster analysis of consumer segmentation models. The research concluded that the collected segmentation results can form the basis for data-driven, individualized customer service, with positive implications for product design and package suggestions [5].

"Oyelade, O. J., et al" (2010), The k-mean clustering approach was utilized for analyzing the pupils' performance data. To evaluate a private university's standing within a subset of Nigerian higher education institutions, this model was combined with a deterministic one [6].

In this study, we use a novel way to clustering algorithms by weighting each feature equally using a decision tree, as stated by Yunus DOAN et al. (2011). Since adopting this strategy, unsupervised learning algorithms' clustering performance has improved dramatically. According to the research, distinct qualities should be prioritized throughout the clustering phase to create the most consistent rule set for a dataset at the end of the process. These rankings are derived using the weighted Euclidean distance calculation and used in experiments. The primary objective is to quantify each trait using a variety of scales[27].

To the Editors: Youguo Li et al (2012), By combining it with the biggest minimum distance approach, the authors of this study provide a more efficient K-Means clustering algorithm. The shortcomings of the conventional K-Means approach might be made up for by this novel approach. The traditional method's (1) dependency on picking the initial focus point and (2) danger of being

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 6s
DOI: https://doi.org/10.17762/ijritcc.v11i6s.6946
***Article Received**: 28 March 2023 **Revised**: 15 May 2023 **Accepted**: 25 May 2023*

caught in a local minimum [7] were both addressed by the modified K-Means technique.

This is supported by the work of Chen Hailong et al (2013), Algorithms specific to each kind of approach—partitioning, hierarchy, density, grid, and model—need to be thoroughly investigated first. Secondly, evaluate the effectiveness of several clustering techniques in terms of high dimensionality, efficiency, sensitivity to "noise," and cluster shape. Last but not least, try out some K-means clustering analysis and hierarchical clustering MATLAB simulations [8].

We thank A. Fahad et al (2014), This study gives a high-level, theoretical and empirical comparison of existing (clustering) approaches. It also presents concepts and algorithms related to clustering. Based on the findings of previous research, we theoretically developed a classification scheme. Using a wide range of real (large) datasets, they compared the best performing algorithms across categories by conducting comprehensive trials based on empirical evidence. The proposed clustering strategies are evaluated based on a wide range of factors, including stability, scalability, runtime, and internal and external validity. The research also highlighted the subset of clustering methods that excels when dealing with massive datasets. [20].

Peerzada H. A. Hamid, et al (2015), Here, we evaluate the effectiveness of four widely used clustering techniques: the Simple K-mean methodology, the Distributed Block Structured Clustering and Analysis Network, the Hierarchical Clustering and Analysis Network, and the Multi-Dimensional Block Clustering and Analysis Method. These four methods are demonstrated and compared using the clustering program Weka. The results are evaluated on several datasets using the WEKA user interface, such as the Abalone, Bank data, Router, SMS, and WebTK datasets. There is a computation of instances, attributes, and the time required to build the model [13].

Together, Pradeep Singh and colleagues (2017), In this research, we examine the DBSCAN algorithm and evaluate its performance in comparison to other density-based approaches. in terms of the input and output that it requires. Some density-based algorithms are more effective and efficient than the DBSCAN approach. Fast-DBSCAN, DENDIS, and GDBSCAN were among the algorithms that performed better than the classic DBSCAN. These algorithms outperformed DBSCAN in terms of performance. New density-based algorithms [11] that can manage clusters of different densities are still required and are resistant to changes in parameters, and have a small footprint.

Rong Zhou, et al. 1, 2, etc. (2018), Machine learning is proposed as a new method for predicting systolic blood pressure (SBP). Clinical and lifestyle characteristics of the model were evaluated (data about age, height, weight, marital status, smoking history, sexual orientation, etc..). Several machine learning approaches were evaluated, together with several training, validation, and testing ratios, in an effort to enhance the model's accuracy. The results were verified in order to make the model more precise and reliable. The American National Standard of the Association for the Advancement of Medical Instrumentation (AAMI) and the British Hypertension Society (BHS) state, our model's performance was comparable to a grade A [17] for SBP estimate.

1 * Among others, Hany Alashwal (2019), The authors of this study look at clustering methods applied to neurodegenerative disease datasets, specifically Alzheimer's (AD). The goal is to shed some insight into the optimal clustering approach for categorizing AD patients according to their shared characteristics. That's important because it gives us more information about the consequences of utilizing clustering methods to treat AD [16], because clustering algorithms may discover patient patterns that are hard to notice by medical staff.

By adapting existing methods to use MapReduce to parallelize Single-Linkage clustering, Joelson and coworkers "Antonio dos Santos et al" (2021) demonstrate why it is wasteful to calculate density-based clustering hierarchies. The original proposal is for a precise and computationally expensive method based on random block parallelization. An alternative parallelization method is offered, which uses a much faster, recursive sampling method to calculate an approximate clustering hierarchy. As a result, hierarchical density-based clustering may be applied efficiently using MapReduce on big datasets. The method relies on data bubbles, a sort of summarization, and is founded on HDBSCAN*, a cutting-edge hierarchical density-based clustering algorithm. Evaluation of several datasets in terms of runtime and approximation quality demonstrates the utility and efficiency of the proposed strategy [30].

Systolic and diastolic blood pressure estimations were computed using a number of regression techniques, several different methods have been shown to be effective, as reported by Ali Farki et al (2022). It was demonstrated that the clustering technique can boost prediction accuracy for each model due to the large variation and numerous trends present in the data and the produced features [9]..

III. CONCLUSION

Data mining and machine learning rely heavily on clustering. It divides the datasets into subsets with common properties, allowing for more accurate forecasts of user behavior. The best possible groupings of data objects can be created using the various clustering techniques described in the article and its three types of big data. Specialized Clustering techniques and Linkages are given. This organized data set serves as a springboard for many possibilities. Quick, trustworthy, and easy to grasp, K-means clustering [2] gets the job done.

REFERENCES

[1] J. A. Hartigan, Clustering Algorithms. New York: Wiley, 1975.

[2] Anju, Preeti Gulia, Clustering In Big Data: A Review, International Journal of Computer Applications (0975 – 8887) Volume 153 – No3, November 2016.

[3] K. Krishna and M. Narasimha Murty, Genetic K-Means Algorithm, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS, VOL. 29, NO. 3, JUNE 1999

[4] Osama Mahmoud Abu Abbas, Comparisons Between Data Clustering Algorithms, International Arab Journal of Information Technology , Vol. 5, No. 3, 2008, pp.: 320-325.

[5] Jinghua Zhao, Wenbo Zhang, Yanwei Liu, Improved K-Means Cluster Algorithm in Telecommunications Enterprises Customer Segmentation, 978-1-4244-6943-7/10/$26.00 ©2010 IEEE.

[6] Oyelade, O. J, Oladipupo, O. O, Obagbuwa, I. C, Application of k-Means Clustering algorithm for prediction of Students' Academic Performance, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, o. 1, 2010. ISSN 1947-5500.

[7] Youguo Li, Haiyan Wu, A Clustering Method Based on K-Means Algorithm, 2012 International Conference on Solid State Devices and Materials Science, Available online atwww.sciencedirect.com, Physics Procedia 25 (2012) 1104 – 1109, doi:10.1016/j.phpro.2012.03.206.

[8] Hailong Chen, Chunli Liu, Research and Application of Cluster Analysis Algorithm, 2013 2nd International Conference on Measurement, Information and Control, 978-1-4799-1392-3/13/$31.00 m013 IEEE.

[9] Ali Farki, Reza Baradaran Kazemzadeh, and Elham AkhondzadehNoughabi, A Novel Clustering-Based Algorithm for Continuous and Noninvasive Cuff-Less Blood Pressure Estimation, Hindawi Journal of Healthcare Engineering Volume 2022, Article ID 3549238, 13 pageshttps://doi.org/10.1155/2022/3549238.

[10] Rupanka Bhuyan1 ,Samarjeet Borah2, A Survey of Some Density Based Clustering Techniques, https://www.researchgate.net/publication/265381945, DOI: 10.13140/2.1.4554.6887.

[11] Pradeep Singh, Prateek A. Meshram, Survey of Density-Based Clustering Algorithms and its Variants, Proceedings of the International Conference on Inventive Computing and Informatics (ICICI 2017) IEEE Xplore Compliant - Part Number: CFP17L34-ART, ISBN: 978-1-5386-4031-9.

[12] T. Soni Madhulatha, AN OVERVIEW ON CLUSTERING METHODS, IOSR Journal of Engineering Apr. 2012, Vol. 2(4) pp: 719-725. ISSN: 2250-3021.

[13] Peerzada Hamid Ahmad1, Dr. Shilpa Dang2, Performance Evaluation of Clustering Algorithms Using Different Datasets, International Journal of Advance Research in Computer Science and Management Studies, ISSN: 232 7782 1 (Online).

[14] GATH AND A. B. GEVA, Unsupervised Optimal Fuzzy Clustering, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. VOL. II. NO. 7. JULY 1989.

[15] T. Velmurugan and T.Santhanam, A Survey of Partition based Clustering algorithms in Data mining: An Experimental Approach,

[16] Hany Alashwal 1 *, Mohamed El Halaby 2†, Jacob J. Crouse3, Areeg Abdalla2 and Ahmed A. Moustafa4, The Application of Unsupervised Clustering Methods to Alzheimer's Disease, Front. Comput Neurosci 13:31. doi: 10.3389/fncom.2019.00031

[17] Rong Zhou, 1,2 Yong Zhang, 1 Shengzhong Feng,1 and Nurbol Luktarhan3, A Novel Hierarchical Clustering Algorithm Based on Density Peaks for Complex Datasets, Hindawi, Complexity Volume 2018, Article ID 2032461, 8 pageshttps://doi.org/10.1155/2018/2032461.

[18] Rui Xu, Student Member, IEEE and Donald Wunsch II, Fellow, IEEE, Survey of Clustering Algorithms, IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 16, NO. 3, MAY 2005.

[19] Shadi Banitaan, Ali Bou Nassif, Mohammad Azzeh, Class Decomposition using K-means and Hierarchical Clustering, 2015 IEEE 14th International Conference on Machine Learning and Applications, 978-1-5090-0287-0/15 $31.00 © 2015 IEEE, DOI 10.1109/ICMLA.2015.169.

[20] A. Fahad, N. Alshatri, Z. Tari, Member, IEEE, A. Alamri, I. Khalil A. Zomaya, Fellow, IEEE, S. Foufou, and A. Bouras, A Survey of Clustering Algorithms for Big Data: Taxonomy & Empirical Analysis, 10.1109/TETC.2014.2330519, IEEE Transactions on Emerging Topics in Computing.

[21] Hong Changa, Dit-Yan Yeungb, ∗, Robust path-based spectral clustering, Pattern Recognition 41 (2008) 191 – 203, 0031-3203/$30.00 2007 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved. doi:10.1016/j.patcog.2007.04.010.

[22] JuhaVesanto and EsaAlhoniemi, Student Member, IEEE, Clustering of the Self-Organizing Map,IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 11, NO. 3, MAY 2000.

[23] CURE: AN EFFICIENT CLUSTERING ALGORITHM FOR LARGE DATABASES+ ,SUDIPTO GUHA1, RAJEEV RASTOGI2, and KYUSEOK SHIMS3, Information Systems Vol. 26, No. 1, pp. 35-58,ZOOl 8 2001 Published by Elsevier Science Ltd. Printed in Great Britain 0306-4379/01.

[24] Research on k-means Clustering Algorithm an Improved k-means Clustering Algorithm, Third International Symposium on Intelligent Information Technology and Security Informatics, 978-0-7695-4020-7/10 $26.00 © 2010 IEEE, DOI 10.1109/IITSI.2010.74.

[25] Clustering with fuzzy supervised algorithm Fong-Jhu Yih1 , Yuan-Horng Lin2 and Jeng-Ming Yih3,a, MATEC Web of Conferences 119, 01007 (2017) DOI: 10.1051/matecconf/201711901007 IMETI 2016.

[26] Christopher M. Bishop (2006) Pattern Recognition and Machine Learning, Springer ISBN 0-387-31073-8.

[27] Yunus DOĞAN, Derya BİRANT, Alp KUT, A New Approach for Weighted Clustering Using Decision Tree, 978-1-61284-922-5/11/$26.00 2011 IEEE.

[28] Depa Pratima, Nivedita Nimmakanti, Pattern Recognition Algorithms for Cluster Identification Problem, International Journal of Computer Science & Informatics (IJCSI), ISSN (PRINT) : 2231–5292, Vol.- 2, Issue-3.

[29] MUSTAFA TAREQ1, ELANKOVAN A. SUNDARARAJAN2, AN EVOLVING APPROACH TO DATA STREAMS CLUSTERING BASED ON CHEBYCHEV WITH FALSE MERGING, Journal of Theoretical and Applied Information Technology, 15th May 2021. Vol.99. No 9 © 2021 Little Lion Scientific.

[30] Joelson Antonio dos Santos, Talat Iqbal Syed, Murilo C. Naldi ^, Ricardo J. G. B. Campello, and Joerg Sander,

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 6s
DOI: https://doi.org/10.17762/ijritcc.v11i6s.6946
***Article Received**: 28 March 2023 **Revised**: 15 May 2023 **Accepted**: 25 May 2023*

Hierarchical Density-Based Clustering Using MapReduce, IEEE TRANSACTIONS ON BIG DATA, VOL. 7, NO. 1, JANUARY-MARCH 2021,Digital Object Identifier no. 10.1109/TBDATA.2019.2907624.

[31] GholamhoseinSheikholeslami, Surojit Chatterjee, Aidong Zhang, WaveCluster: a wavelet-based clustering approach for spatial data in very large databases, The VLDB Journal c Springer-Verlag 2000.

[32] Fionn Murtagh (1, 2) and Pedro Contreras (2), Methods of Hierarchical Clustering, arXiv:1105.02121v1[cs.IR] 30 Apr 2011.

[33] (Tang 2005) Tang, Z., and MacLennan, J.: Data Mining with SQL Server 2005. Wiley Publishing, Inc., Indianapolis, Indiana, USA (2005) Chapter 8

[34] Alauddin Yousif Al-Omary a*, Mohammad Shahid Jamil b, A new approach of clustering based machine-learning algorithm, Knowledge-Based Systems 19 (2006) 248–258, www.elsevier.com/locate/knosys.

[35] Encyclopedia of Machine Learning pp 220–221

[36] Ira Assent⁎, Clustering high dimensional data, WIREs Data Mining and Knowledge Discovery, c 2012 John Wiley & Sons, Inc. DOI: 10.1002/widm.1062 .

[37] MacQueen J (1967) some methods for classification and analysis of multivariate observations. ProcFifth Berkeley Symp Math Stat Probab 1:281–297.

[38] Park H, Jun C (2009) A simple and fast algorithm for K-medoids clustering. Expert Syst Appl36:3336–3341.

[39] Kaufman L, Rousseeuw P (1990) Partitioning around medoids (program pam). Finding group's indata: an introduction to cluster analysis. Wiley, Hoboken.

[40] Kaufman L, Rousseeuw P (2008) Finding groups in data: an introduction to cluster analysis, vol 344.Wiley, Hoboken. Doi: 10.1002/9780470316801 .

[41] L. Kaufman and P.J. Rousseeuw. (1990) Finding Groups in Data:an Introduction to Cluster Analysis, John Wiley & Sons.

[42] Zhang T, Ramakrishnan R, LivnyM (1996) BIRCH: an efficient data clustering method for very large databases. ACM SIGMOD Rec 25:103– 104.

[43] Guha S, Rastogi R, Shim K (1998) CURE: an efficient clustering algorithm for large databases. ACMSIGMOD Rec 27:73–84

[44] Guha S, Rastogi R, Shim K (1999) ROCK: a robust clustering algorithm for categorical attributes. In: Proceedings of the 15th international conference on data engineering, pp 512-521.

[45] Karypis G, Han E, Kumar V (1999) Chameleon: hierarchical clustering using dynamic modeling.Computer 32:68–75.

[46] Ester M, Kriegel H, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the second ACM SIGKDD international conference on knowledge discovery and data mining, pp 226–231.

[47] AnkerstM, BreunigM, Kriegel H, Sander J (1999) OPTICS: ordering points to identify the clusteringstructure. In: Proceedings on 1999 ACMSIGMOD international conference on management of data, vol 28, pp 49–60.

[48] Hinneburg A, Keim D (1998) An efficient approach to clustering in large multimedia databases with noise. In Proceedings of the 4th ACM SIGKDD international conference on knowledge discovery and data mining 98: 58–65.

[49] Das, S. K. ., Pani, S. K. ., Padhy, S. ., Dash, S. ., & Acharya, A. K. . (2023). Application of Machine Learning Models for Slope Instabilities Prediction in Open Cast mines. International Journal of Intelligent Systems and Applications in Engineering, 11(1), 111–121. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2449.

[50] Hartuv E, Shamir R (2000) A clustering algorithm based on graph connectivity. Inf Process Lett76:175–181.

[51] Wang W, Yang J, Muntz R (1997) STING: a statistical information grid approach to spatial datamining. In VLDB, pp 186–195.

[52] Agrawal R, Gehrke J, Gunopulos D, Raghavan P (1998) Automatic subspace clustering of high dimensional data for data mining applications. In: Proceedings 1998 ACM sigmodinternationalconference on management of data, vol 27, pp 94–105.

[53] Hinneburg, A., & Keim, D. A. (1999). Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering.

[54] Wu, C. J. (1983). On the convergence properties of the EM algorithm. The Annals of statistics, 95-103.

[55] FisherD(1987) Knowledge acquisition via incremental conceptual clustering. Mach Learn 2:139–172.

[56] Kohonen, T. (1998). The self-organizing map. Neurocomputing, 21(1-3), 1-6.

[57] AblaChouni Benabella1*, Asmaa Benghabrit2, Imane Bouhaddou3, A survey of clustering algorithms for an industrial context, Second International Conference on Intelligent Computing in Data Sciences (ICDS 2018), Available online at www.sciencedirect.com, Procedia Computer Science 148 (2019) 291–302, 10.1016/j.procs.2019.01.022.

[58] Tian Zhang, Raghu Ramakrishnan,MironLivny", BIRCH: An Efficient Data Clustering Method for Very Large Databases, SIGMOD '96 6/96 Montreal, Canada IQ 1996 ACM 0-89791 -794-4/96/0006 ...$3.5.

[59] SI:~IPTO GUHA~, RAJEEV RASTOGI', and KYUSEOK SHIMS, ROCK: A R.OBUST CLUSTERING ALGORITHM FOR, CATEGORICAL ATTRIBUTES+, InformafiotlSysrems Vol. 25, No. 5, pp. 345-366, 2000 0 2000 Published by Elsevier Science Ltd. Printed in Great Britain 0306-4379100 $20.00.

[60] G. Naga Rama Devi, Comparative Study on Machine Learning Algorithms using Weka, International Journal of Engineering Research & Technology (IJERT) IJERT www.ijert.org NCDMA - 2014 Conference Proceedings ISSN: 2278-0181.