Journal of Innovative Engineering and Research (JIER)
Vol. - 7, Issue -2, October, 2024, pp. 1-8 (8 pages)

A Comprehensive Review of Clustering Techniques: Methods, Applications, and Challenges

Dinesh Bhardwaj and Dr. Sonawane Vijay Ramnath

Department of Computer Science & Engineering, Dr. A. P. J. Abdul Kalam University
Indore (M.P.) – 452010, India
Corresponding Author Email: dkbh28@gmail.com

Abstract: ***Clustering is a fundamental unsupervised learning technique widely used for data analysis and pattern recognition. Many subsequent studies rely on identifying operational taxonomic units, and hierarchical clustering is one of the most popular methods for doing so. Because of their quadratic space and computing difficulties, most known methods are limited in their applicability to situations of moderate size or less. To solve the space and computational bottlenecks of existing solutions, we offer a novel online learning-based technique. It examines the strengths, limitations, and computational complexities of each technique and highlights their suitability for different types of datasets. Additionally, emerging trends, such as deep clustering and hybrid methodologies, are discussed to illustrate the evolving nature of clustering research. The review also addresses challenges, such as scalability, high-dimensional data, and interpretability, while outlining future directions for research. This work aims to serve as a valuable resource for researchers and practitioners seeking to understand and apply clustering techniques effectively.***

Keywords: ***Clustering algorithms, data streams clustering, hierarchical clustering, parameter selection.***

I. INTRODUCTION:

Clustering is a powerful method for dissecting microarray genomic data and learning about the underlying patterns of interest. This type of uncertain learning approach is frequently used in larger genomic expression datasets to discover closely related but obscured similarities and likelihoods.. Clustering, also known as unsupervised learning, is one of the most important topics in machine learning since it may be used to extract or summarize fresh information. Data mining, pattern recognition, and statistics are just a few of the numerous fields that benefit from it. This research looks at cluster analysis in the context of data mining. When it comes to solving problems in the information technology (IT) industry, clustering is one of the most important areas of data mining. When the dataset dimensions are high, clustering the data is a computationally intensive and time-consuming process.

Items inside the same cluster "have great resemblance to each other, whereas objects in distinct classes are more dissimilar," as described in. Unsupervised classification includes methods such as clustering. Classification is the process of allocating data items to a taxonomy of categories. When categorizing data items, "un-supervised" clustering does not depend on previously established classes or training samples. Since the goals of pattern recognition and the statistical subfields of discriminant analysis and decision analysis are to develop rules for classifying things, clustering stands in contrast to these other methods.

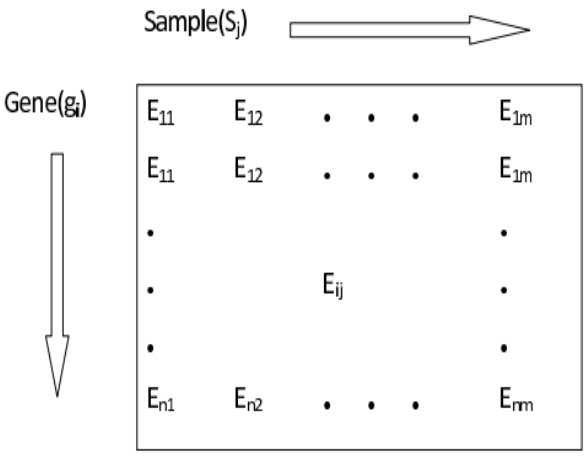

Figure 1: Gene Expression Matrix/Intensity Matrix

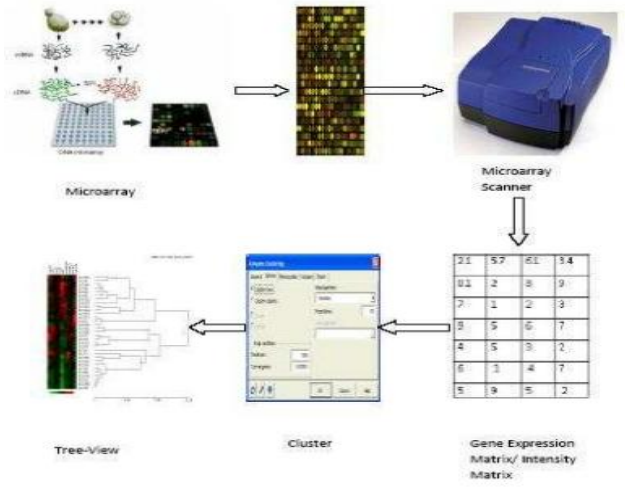

Figure 2: Complete Overview of Gene Expression Analysis

1.1 DATA MINING

Data mining is a method used to evaluate enormous data sets for hidden insights. Analyzing enormous document databases is the subject of text mining, a subset of data mining. It's a technique for deducing meaning from vast volumes of unstructured or semi-structured textual data. This depends on the synergy between human language abilities and the processing speed of computers. In order to communicate effectively, one must be able to recognize and use spelling variants, sort relevant information from irrelevant material, recognize and use synonyms and abbreviations, and discover meaning in context.

The statistical and probabilistic processing of massive volumes of data at rapid speed is a feature of computers' computing capacity. Text mining may be put to use in a variety of contexts, including data extraction, topic tracking, summarization, classification, clustering, linking concepts, visualizing data, and more. Text mining methods may frequently be divided down into supervised learning and unsupervised learning categories. In supervised learning, the learning process is directed by the expert knowledge already existing in the system, allowing for more accurate prediction of the goals. Predictor and target attribute association discovery is the next step in this technique [7]. .

1.2. DATA CLUSTERING

In a cluster, objects are most similar to those already within the cluster and most dissimilar to those outside it. This process is known as data clustering. This data mining method allows for the meaningful categorization of objects into groups, or "clusters," which can then be abstracted from large amounts of data. Which items are placed in the same cluster is determined by the similarity measure. Defining a similarity or distance measure specified across the object feature space is necessary for data clustering since the choice is dependent on the similarity or distance between objects[9]. Whenever a computer is able to classify data into one of a number of predetermined categories, we say that the data has been automatically categorized. However, by using clustering techniques, a machine can determine which partitioning strategy is best for a given dataset. When there is a requirement to classify newly acquired information into a previously established category, categorization is a viable option. When seeking for unseen patterns, clustering is a useful technique to use. Both classification and clustering may provide attractive results on an unknown dataset; the former organizes the data according to a known pattern, while the latter shows that structure. In this thesis, we will examine how to improve the clustering of document datasets. The main objective of document clustering is to maximize similarity between groups while decreasing it overall.

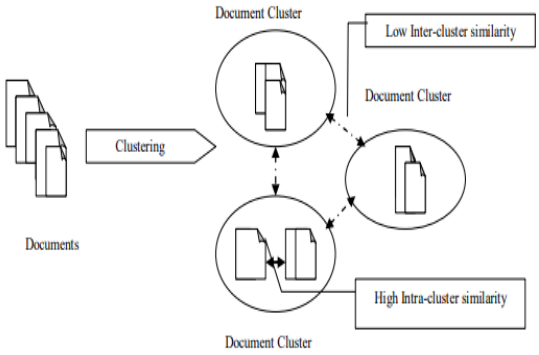

Figure 1: Goal of Document Clustering

Due to their great dimensionality, these datasets provide unique challenges for statistical analysis. Distributed

computing plays a crucial part in the worldwide data mining necessary to make work on Big Data feasible. The emergence of Distributed Data Mining (DDM) as a promising field of study with the potential to address several problems has led to the development of numerous new ideas. The processing of large datasets for scientific and grid applications using machine learning is one of the newest fields of study. Due to the difficulties inherent in the task of extracting previously new information from extremely large centralized real-world datasets, this is one of the major and active fields of study.

1.3. CLUSTERING DATA MINING TECHNIQUES

One way to organize data is by clumping similar pieces together. Though reducing the number of clusters used to represent the data can make it easier to understand, doing so sacrifices some of the data's granularity. A data's clusters are used as a model in this system. Clustering is given context by the history of data modeling, which originates in the fields of mathematics, statistics, and numerical analysis. Clusters, in the context of machine learning, are analogous to covert patterns; cluster discovery is an example of unsupervised learning; and the final system stands in for some kind of data notion. Clustering has a wide variety of practical uses in data mining, including but not limited to: data mining in the sciences, IR/text mining, GIS, Web analysis, CRM, marketing, medical diagnostics, computational biology, and many more rely on this technique.

Several disciplines, including statistics, pattern recognition, and machine learning, are actively investigating clustering. Clustering in data mining is the subject of this review. Complicating matters further, data mining amplifies the difficulties of grouping very huge datasets with exceptionally many features of varying sorts[12]. As a result, the appropriate clustering algorithms face novel computational challenges. Several algorithms have developed in recent years that can meet these requirements, and they have been successfully applied to the problem of data mining in the real world. The survey's focus is on them. This research set out to do just that, providing a comprehensive evaluation of the clustering strategies that may be used to data mining. Cluster analysis is a method of organizing data by identifying and grouping similarities between nodes.

1.4. DATA MINING AND KNOWLEDGE DISCOVERY

For the purpose of extracting useful information that might assist in decision-making, it is becoming more vital to develop powerful tools for analyzing and, possibly, understanding the massive volumes of data stored in files, databases, and other repositories. The definition of data mining, or Knowledge Discovery in Databases, is "the nontrivial process of uncovering legitimate, new, potentially helpful, and ultimately intelligible pattern in data" (KDD). The terms "data mining" and "knowledge discovery in databases" (KDD) are sometimes used interchangeably, however data mining is a subset of KDD.

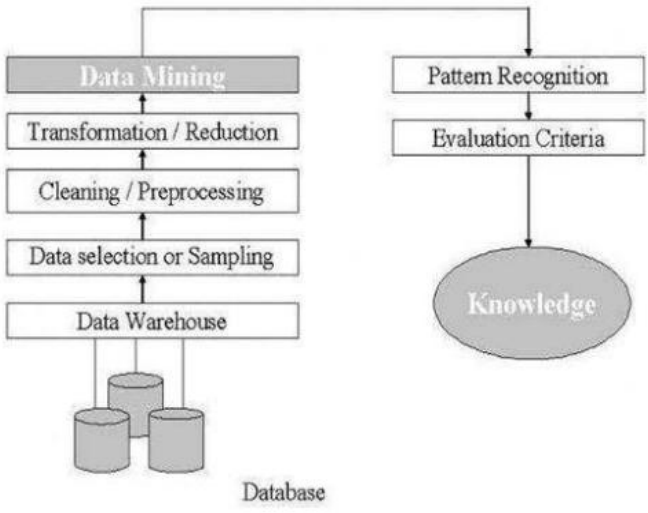

Figure 4: Complete Overview of Knowledge discovery from Databases.

II. MACHINE LEARNING

Machine learning is a subfield of AI that enables machines to "learn" from their experiences and adapt their actions accordingly. In its formal definition, machine learning is "the study that gives computers the ability to learn without being explicitly programmed." The decisions made from the analysis of patterns and relations between data are not the goal of machine learning but rather the purpose of the overall

system. Data Mining is a field that focuses on discovering patterns in large datasets. There is much cross-over between these two fields of study. Machine learning is more concerned with making predictions, whereas data mining is more concerned with uncovering hidden relationships in data. It's common to hear people discuss the distinction between supervised and unsupervised learning. Data labels are known in supervised learning, allowing for the development of a predictive model. On the other hand, in unsupervised learning, similarities must be extracted from the data without any knowledge of the data labels.

III.GRIDDED PORTRAYAL OF HIGH-DIMENSIONAL STANDARDIZED UNIFORM ORGANIZED DATASET

Divides information into understandable or useful groups using a clustering algorithm (clusters). Clusters are most helpful when they faithfully portray the "natural" underlying structure of the data. Cluster analysis has been put to use in a variety of contexts, including the classification of earthquake-prone locations, the identification of functionally similar genes and proteins, and the simplification of document reading. However, there are circumstances in which cluster analysis is only a stepping stone to something more substantial, such as data compression or efficiently pinpointing neighboring locales. Cluster analysis has been used for quite some time in many fields for both theoretical and applied research and development, such as psychology, sociology, biology, statistics, pattern recognition, information retrieval, machine learning, and data mining.

Using data defining the objects or their connections, cluster analysis classifies items (observations, occurrences) into groups. The idea is to have things within a group have certain characteristics while still being distinct from those inside other groups. To achieve optimal clustering, group similarities (or homogeneity) and group differences should be as high as possible. There is no universally accepted definition of a cluster, and in many contexts, clusters are not clearly distinguished from one another. However, the goal of most cluster analyses is to classify the data into distinct, non-overlapping categories. One notable exception is fuzzy clustering which acknowledges that certain objects may have overlapping membership in several categories

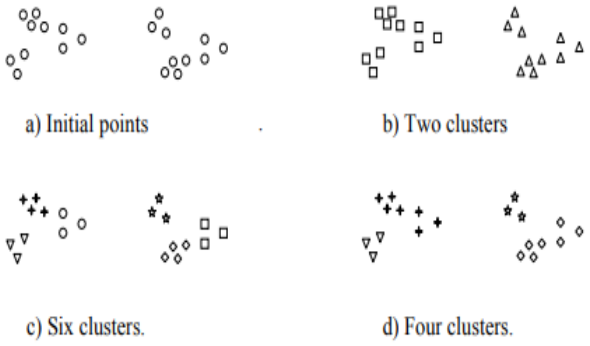

Figure 5: Different clustering's for a set of points.

IV.RESULTS:

4.1. DATA VISUALIZATION

"Data visualization" is a term for the process of making meaningful pictures out of numbers. Discovering and displaying hidden patterns, trends, and correlations in a large data set is its primary goal in order to fulfill a user's needs. The primary advantage of data visualization, as described by John Tukey, the "founder" of exploratory data analysis and one of the earliest pioneers of visualization: "Number values emphasize anticipated values, whereas visual summaries emphasize unexpected 5" One of the primary focuses of visualization is the visual presentation of data, and this field includes both scientific visualization and information visualization.

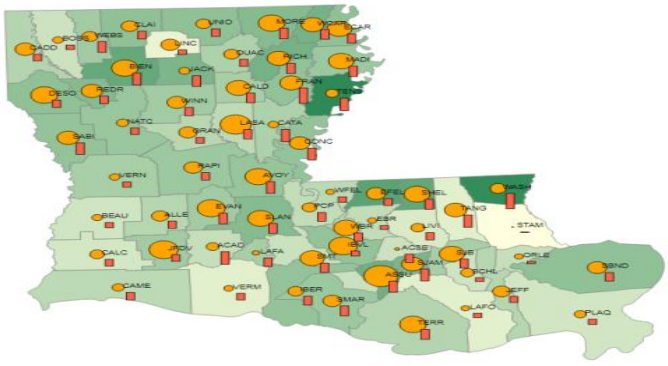

Figure 6: The figure shows the data

Two of the most important subfields of visualization, both concerned with data presentation, are scientific visualization and information visualization. For the geometric representation to make sense, the data must first be transformed into a geometric space, such as a scatterplot, histogram, or parallel coordinates plot. Contrarily, in iconic representations like star plots and Chernoff faces, information is represented in terms of an icon or glyph. The following are some illustrations of the application of the scatterplot matrix and the Chernoff faces.

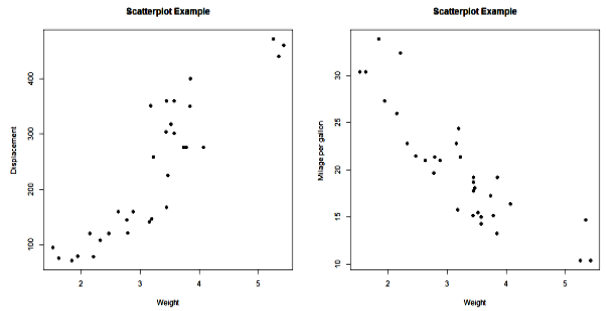

Figure 7: Scatterplot of a bivariate data for a car dataset.

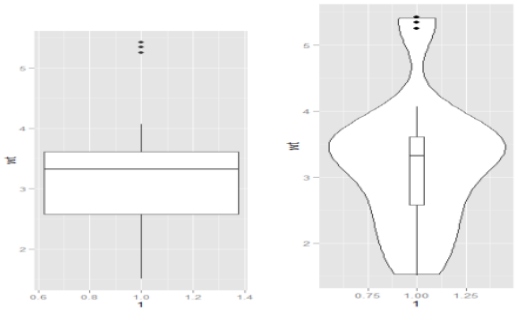

Figure 8: Examples of box and violin plots.

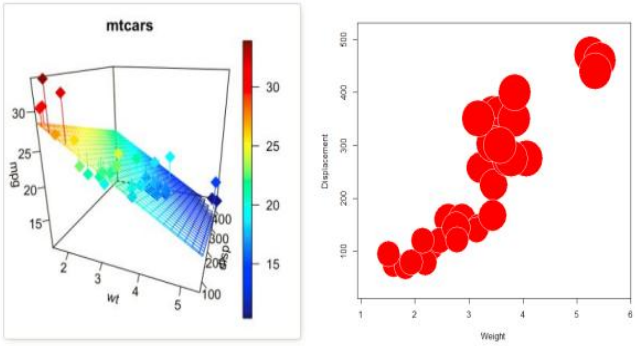

Figure 9: Scatterplot of a trivariate data for a car dataset.

4.2. THE MICROSTRUCTURE OF SYNCHRONY IN CHRONICALLY EPILEPTIC NETWORKS IS MADE UP OF SPATIALLY CLUSTERED NEURONAL ASSEMBLIES

When neural networks synchronize, it can have serious consequences across a wide variety of systems, from the development of healthy circuits (1, 2) to the onset of pathological conditions like epilepsy. (3, 4).Although research into these systems often focuses on the population level of activity, the interplay of individual neurons to generate synchronous network dynamics is still poorly understood. Given that epileptic networks may underpin both typical brain functioning and abnormal activity, studying epilepsy from a networks approach is of particular interest. Several computational studies (5-8) have established a causal relationship between alterations in the network connectivity and the emergence of the hypersynchronous neuronal activity characteristic of this disorder.

In addition, several types of epilepsy are marked by substantial abnormalities in the underlying structure of the neural network, such as cell loss and axon sprouting (9). Temporal lobe epilepsy (TLE) is characterized by changes in the dentate gyrus (DG), such as the death of hilar interneurons and mossy cells, the formation of recurrent connections between granule cells via mossy fiber sprouting, and the emergence of novel cellular shapes. Interictal spikes are short, synchronous events typical of epileptic networks and are typically detected by electroencephalography (EEG) or local field potential recordings in epileptic tissue (17).To further understand how network structure and dynamics are altered in epileptic networks, we use multibeam two-photon calcium imaging in slices from the DG of pilocarpine-treated, chronically epileptic mice to examine functional network structure at the level of single-cell resolution.

4.3. THE CREATION AND OPTIMIZATION OF SPECTRAL CLUSTER MAPS FOR COMBINING CASSIS AND CRISM DATA SETS

It is a major obstacle to the study of our solar system that the high-resolution datasets provided by in situ orbiting spacecraft are usually insufficient, either geographically or

spectrally, or both. Observations of the Martian surface demonstrate the need for high-resolution remote sensing to determine the physicochemical qualities of specific locations. Then, and only then, can things like future landing viability and procedure interpretation be addressed. However, greater detail requires more data but less space. Instruments such as NASA's Mars Reconnaissance Orbiter's (MRO) High-Resolution Imaging Science Experiment (HiRISE) [1, 2] and the Compact Reconnaissance Imaging Spectrometer (CRISM) [3, 4] and the European Space Agency's (ESA) ExoMars Trace Gas Orbiter's (CaSSIS) [5, 6] CaSSIS [3] CaSSIS [5] are all affected by this problem.Despite the fact that these three devices have very high resolution, they cover less than 5% of the Earth's surface between them. Resolution problems, as well as the large number of spectral bands and subclasses, make spectral data difficult to interpret. For this reason, our work employs unsupervised classification, a crucially essential standard approach in geospatial analysis [4] that is particularly useful for evaluating hyperspectral data with inadequate calibration in-field data. In this study, we use a data structure called Spectral Cluster Maps to combine band information with geographical distributions for analysis (SCMs).

High dimensional data is translated to a low latent variable form by directly applying sophisticated algorithms on the entire spectrum itself, and these clusters may be linked to the underlying geochemical composition (compare Gao et al. [5]). Consequently, before implementing different clustering algorithms on the feature space, it is crucial to discover appropriate unsupervised dimensionality reduction strategies to generate correct SCMs. Since principal component analysis (PCA) [6] is the go-to technique for spectral data analysis, we utilize it here as a benchmark against more intricate techniques (see, for example, [7,8]).

Recently, methods like t-SNE [9] have shown promise in research on Machine Learning Networks. In order to better classify data, we reduced the dimensionality of the feature space and focused on its local structures. See, for instance, Pouyet et al. and Song et al. [10,11] or Kohonen's self-organizing maps method for a few examples of these proposed spectral database constructions[12]. In their most recent paper, Gao et al. [5] suggest using an autoencoder approach for spectrum applications on Mars. Currently, spectral data are seldom used by the UMAP technique. Two groups that discuss this issue are Picollo et al. [13] and Wander et al. [14]. The most recent and significant work in this field is by D'Amore and Padovan [15], who utilize UMAP to map reflectance spectra from Mercury.

There are more UMAP publications in the area of biology [16,17], but its speed and resilience make it desirable to use it more widely in planetary science. Creating spectral cluster maps with CRISM and Cassis data has an immediate impact on geologic mapping efforts. Planetary geologic mapping has traditionally relied on more basic geometric and stratigraphic principles due to a lack of available image and topography data. Incorporating a wide range of methodologies, from heuristics to statistics, has been possible because to the proliferation of compositional data in the last several decades [19–22]. We put our method to the test in Coprates Chasma because of the striking mineralogical (color) diversity it exhibits in CRISM and CaSSIS

V. Conclusion:

Clustering techniques are indispensable tools for extracting meaningful patterns and structures from data in a wide range of applications. This review has provided a detailed overview of various clustering methods, including partitioning, hierarchical, density-based, model-based, and grid-based approaches, along with emerging techniques such as deep and hybrid clustering. Each method has its unique strengths and limitations, emphasizing the importance of selecting the right technique based on the characteristics of the dataset and the specific objectives of the analysis.Despite significant advancements, clustering still face challenges such as handling high-dimensional and noisy data, managing large-scale datasets, and ensuring interpretability of results.

Addressing these issues requires ongoing research and innovation, particularly in integrating clustering with emerging technologies like deep learning and cloud computing. Future work should also focus on domain-specific adaptations and developing methods that balance computational efficiency with clustering accuracy.

VI. References

1. G Karypis, E H Han and V Kumar, 'CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modelling' Computer, vol 32, no 8, pp 68-75, 1999
2. T Zhang, R Ramakrishnan, and M Linvy, 'BIRCH: An Efficient data Clustering Method for Very Large Datasets', Data Mining & KD, vol 1, no 2, pp 141-182, 1997
3. Xiong, H., Wu, J., and Chen, J.. K-Means Clustering Versus Validation Measures: A Data-Distribution Perspective. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), Vol. 39, Issue 2, pp. 318–331. (2009)
4. Schaeffer, S. E. Graph clustering. Computer Science Review, Vol.1, Issue 1, pp. 27–64. (2007)
5. Rodriguez, A. and Laio, A. Clustering by fast search and find of density peaks. Science, Vol. 344, Issue 6191, pp. 1492–1496. (2014)
6. McParland, D. and Gormley, I. C. Model based clustering for mixed data: clustMD. Advances in Data Analysis and Classification, Vol. 10, Issue 2, pp. 155–169. (2016)
7. Maier, M., Hein, M., and von Luxburg, U. Optimal construction of k-nearest-neighbor graphs for identifying noisy clusters. Theoretical Computer Science, Vol. 410, Issue 19, pp. 1749–1764. (2009)
8. Liang, J., Zhao, X., Li, D., Cao, F., and Dang, C. Determining the number of clusters using information entropy for mixed data. Pattern Recognition, Vol. 45, Issue 6, pp. 2251–2265. (2012)
9. Li, C. and Biswas, G. Unsupervised learning with mixed numeric and nominal data. IEEE Transactions on Knowledge and Data Engineering, Vol. 14, Issue 4, pp. 673–690. (2002)
10. Huang, Z. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery, Vol. 2, Issue 3, 283–304. (1998).
11. Hsu, C.-C. and Chen, Y.-C. Mining of mixed data with application to catalog marketing. Expert Systems with Applications, Vol. 32, Issue 1, pp. 12–23. (2007)
12. Foss, A., Markatou, M., Ray, B., and Heching, A. A semiparametric method for clustering mixed data. Machine Learning, Vol. 105, Issue 3, pp. 419–458. (2016)
13. Ahmad, A. and Dey, L. A k-mean clustering algorithm for mixed numeric and categorical data. Data & Knowledge Engineering, Vol. 63, Issue 2, pp. 503–527. (2007)
14. M. Mousavi, A. A. Bakar, and M. Vakilian, ''Data stream clustering algorithms: A review,'' Int. J. Adv. Soft Comput. Appl., vol. 7, no. 3, p. 13, 2015.
15. M. Mousavi and A. A. Bakar, ''Improved density based algorithm for data stream clustering,'' J. Teknologi, vol. 77, no. 18, pp. 73–77, Nov. 2015.
16. A. Algergawy, M. Mesiti, R. Nayak, and G. Saake, ''XML data clustering: An overview,'' ACM Comput. Surv., vol. 43, no. 4, pp. 1–41, Oct. 2011.
17. H. L. Nguyen, Y.-K. Woon, and W.-K. Ng, ''A survey on data stream clustering and classification,'' Knowl. Inf. Syst., vol. 45, no. 3, pp. 535–569, 2015.
18. V. Moustaka, A. Vakali, and L. G. Anthopoulos, ''A systematic review for smart city data analytics,''

ACM Comput. Surv., vol. 51, no. 5, pp. 1–41, Jan. 2019.

19. O. Y. Al-Jarrah, P. D. Yoo, S. Muhaidat, G. K. Karagiannidis, and K. Taha, ‘‘Efficient machine learning for big data: A review,’’ Big Data Res., vol. 2, no. 3, pp. 87–93, 2015.
20. D. Xu and Y. Tian, ‘‘A comprehensive survey of clustering algorithms,’’ Ann. Data Sci., vol. 2, no. 2, pp. 165–193, 2015.
21. M. Hahsler and M. Bolaños, ‘‘Clustering data streams based on shared density between micro-clusters,’’ IEEE Trans. Knowl. Data Eng., vol. 28, no. 6, pp. 1449–1461, Jun. 2016.
22. S. Ding, F. Wu, J. Qian, H. Jia, and F. Jin, ‘‘Research on data stream clustering algorithms,’’ Artif. Intell. Rev., vol. 43, no. 4, pp. 593–600, Apr. 2015