An Unsupervised Machine Learning Algorithms: Comprehensive Review

Article *in* International Journal of Computing and Digital Systems · April 2023

DOI: 10.12785/ijcds/130172

CITATIONS
94

READS
17,968

4 authors:

Samreen Naeem
Islamia University of Bahawalpur
33 PUBLICATIONS **477** CITATIONS
SEE PROFILE

Aqib Ali
Concordia College Bahawalpur
51 PUBLICATIONS **762** CITATIONS
SEE PROFILE

Sania Anam
Govt Associate College for Women Ahmadpur East, Bahawalpur, Pakistan
32 PUBLICATIONS **531** CITATIONS
SEE PROFILE

Munawar Ahmed
Islamia University of Bahawalpur
16 PUBLICATIONS **177** CITATIONS
SEE PROFILE

International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys.13, No.1 (Apr-23)

http://dx.doi.org/10.12785/ijcds/130172

An Unsupervised Machine Learning Algorithms: Comprehensive Review

Samreen Naeem[1], Aqib Ali[1], Sania Anam[2] and Muhammad Munawar Ahmed[3]

[1]*College of Automation, Southeast University, Nanjing 210096, China.*
[2]*Department of Computer Science, Govt Associate College for Women Ahmadpur East, Bahawalpur, Pakistan.*
[3]*Department Information Technology, The Islamia University of Bahawalpur, Bahawalpur, Pakistan.*

Received 25 May. 2022, Revised 19 Dec. 2022, Accepted 6 Feb. 2023, Published 16 Apr. 2023

Abstract: Machine learning (ML) is a data-driven strategy in which computers learn from data without human intervention. The outstanding ML applications are used in a variety of areas. In ML, there are three types of learning problems: Supervised, Unsupervised, and Semi-Supervised Learning. Examples of unsupervised learning techniques and algorithms include Apriori algorithm, ECLAT algorithm, frequent pattern growth algorithm, clustering using k-means, principal components analysis. Objects are grouped based on their same properties. The clustering algorithms are divided into two categories: hierarchical clustering and partition clustering. Many unsupervised learning techniques and algorithms have been created during the last decade, and some of them are well-known and commonly used unsupervised learning algorithms. Unsupervised learning approaches have seen a lot of success in disciplines including machine vision, speech recognition, the creation of self-driving cars, and natural language processing. Unsupervised learning eliminates the requirement for labeled data and human feature engineering, making standard machine learning approaches more flexible and automated. Unsupervised learning is the topic of this survey report.

Keywords: Machine Learning, Unsupervised Learning, Clustering, Unsupervised Algorithms.

1. INTRODUCTION

Because of its inherent concepts, machine learning can be considered a sub field of Artificial Intelligence (AI). Allows for prediction; its core components are algorithms. ML enables systems to learn independently without having to be explicitly programmed to do so, resulting in more intelligent behavior. Develop models that detect trends in historical data and utilize those models to make forecasts to generate data-driven predictions [1], [2], [3]. Figure 1 depicts the overall architecture of machine learning, which is made up of multiple steps: Modeling (which includes functionality engineering, model training, and assessment), Data Acquisition and Understanding (data collecting and understanding), and Business Comprehending (understanding and understanding the domain) as well as a deployment (deploy the model to the cloud). Unsupervised Learning (UL) is a machine learning approach for detecting patterns in datasets with unlabeled or unstructured data points. In this learning approach, an artificial intelligence system gets just the input data and not the associated output data. Unsupervised machine learning, unlike supervised learning, does not need the presence of a person to oversee the model [4]. The data scientist enables the system to learn on its own by looking at the data and identifying patterns. To put it another way, this sub-category of machine learning allows a system to operate on particular data without external instruction. Unsupervised learning approaches are essential to constructing human artificial intelligence systems [5]. Because intelligent machines must be able to make (independent) conclusions based on enormous amounts of unlabeled data, this is the case. UL algorithms are better at solving complicated tasks than supervised learning algorithms. However, supervised learning models provide more accurate results because a programmer explicitly teaches the system what to search for in the data presented. Unsupervised learning, on the other hand, may be highly unexpected. Unsupervised learning may be the foundation for artificial neural networks, making deep learning possible [6]. While this is true, supervised learning techniques for neural networks may also be used if the intended output is already known. Learning without supervision may be a goal in and of itself. UL models, for example, may be used to uncover hidden models in massive amounts of data and categorize and label data points. The similarities and contrasts between unsorted data points are used to group them [7].

Here are a few reasons why unsupervised learning is so important.

- There is a lot of unlabeled data.

E-mail address: samreencsit@gmail.com, aqibcsit@gmail.com, sania.anam7@gmail.com, munawar69@iub.edu.pk

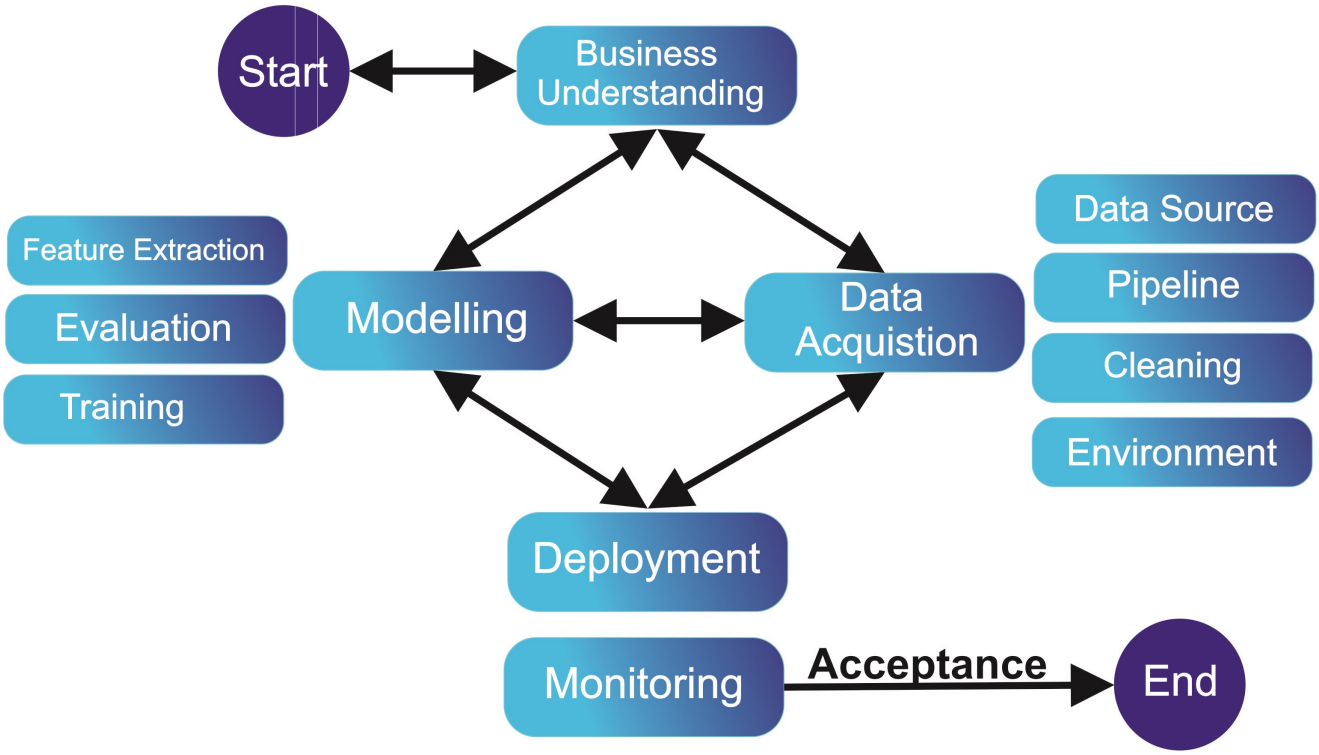


Figure 1. Basic Architecture of ML

- Data tagging is a time-consuming operation that necessitates human intervention.
- However, ML may be used to drive the same process, making coding easier for everyone involved.
- It can be used to investigate unknown or unprocessed data.
- It comes in handy when dealing with massive data sets and pattern detection.

A. *Literature Review*

Many researchers have surveyed unsupervised learning (UL) techniques. The reference [8] surveyed unsupervised learning literature. Their study included 49 studies. UL models are equivalent to supervised learning (SL) models, and Fuzzy C-means and Fuzzy SOMs perform best among UL methods. Their work focused on UL models for software fault prediction. The reference [4] analyzed supervised and UL studies using a literature scan. They prioritized research works published between 2015 and 2018 that address or use supervised and unsupervised ML approaches. This survey only included k-means, hierarchical clustering, and PCA. The reference [9] surveyed UL multiway models, algorithms, and their applications in chemometrics, neurology, social network analysis, text mining, and computer vision. Their poll exclusively analyzed unsupervised multiway data. The reference [10] surveyed the literature—this study surveys time-series clustering techniques. The uniqueness and limitations of earlier studies are also explored, along with prospective research areas. Time series clustering applications are also listed. This literature review focuses on time series clustering approaches. The reference [11] surveys unsupervised and semi-supervised clustering that describe clustering techniques and methodologies. The authors gave external and internal clustering validity measures. Their work helps researchers, although their literature review is limited to algorithms and clustering.

B. *Motivation and Contribution*

Unsupervised algorithms are extensively employed to complete data mining jobs; they are discussed alone or in groups based on learning needs. Literature studies on supervised algorithms tend to focus little on unsupervised. The authors analyzed 35 papers between 2018 and 2022 and found that majority focused on unsupervised learning techniques. This review focused on unsupervised machine learning techniques developed between 2018 and 2022.

2. UNSUPERVISED LEARNING

In supervised learning, a data scientist offers labeled data to the system, such as photographs of cats tagged as cats, so that it may learn by example. In unsupervised learning, a data scientist merely gives photos, and it is up to the system to examine the data and determine whether or not they are cat images. Large amounts of data are required for unsupervised machine learning [12]. In most

circumstances, supervised learning works similarly, with the model becoming more accurate as more examples are added. When data scientists use datasets to train algorithms, the unsupervised learning process begins. These datasets include no labeled or classed data points. The purpose of learning the algorithm is to find patterns in the dataset and rate the data points according to those patterns. The clustering, association, anomaly detection, and autoencoder issues are four types of unsupervised learning challenges, as shown in Figure 2.

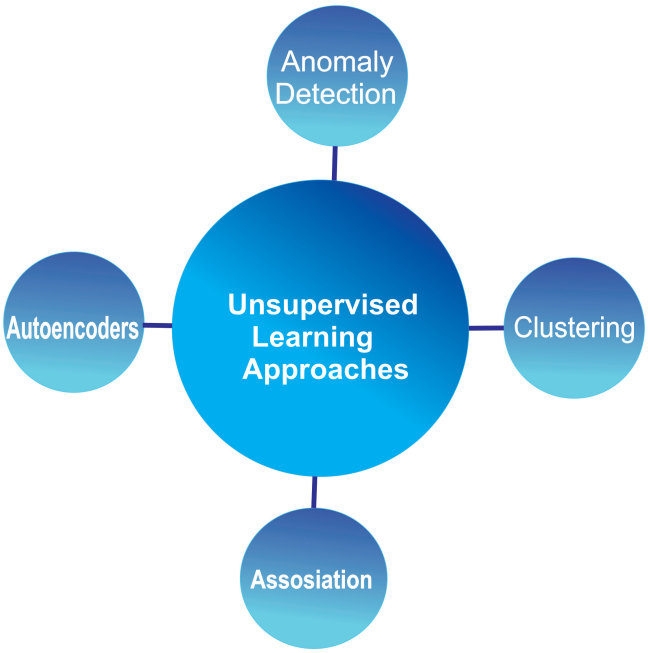


Figure 2. Types of Unsupervised Learning

In the case of cat photos, the unsupervised learning system may learn to recognize distinguishing features like whiskers, long tails, and retractable claws. Unsupervised learning is how humans learn to identify and classify objects they think about. Let's pretend you've never had ketchup or spicy sauce before [13]. You will be able to discern the difference between two "unlabeled" bottles of ketchup and chili sauce if you are given two "unlabeled" bottles of each and asked to taste them. Even if you don't know the names of either sauce (one sour, the other spicy), you'll be able to recognize its characteristics. You can get a better sense of the flavor by tasting them more times. Soon, you'll be able to categorize foods depending on the amount of sauce they contain just by tasting them. Specific qualities that distinguish the two sauces and the group's nutrition can be discovered by studying the flavor. To categorize them, you don't need to know the names of the sauces or the foods. Alternatively, you might refer to one as sweet sauce and the other as a spicy sauce. Machines use unsupervised learning to find patterns and similarly classify data points. In the same way, supervised learning would entail someone informing you of the sauce names and flavors ahead of time [14], [15]. The workflow of the clustering algorithm is shown in Figure 3.

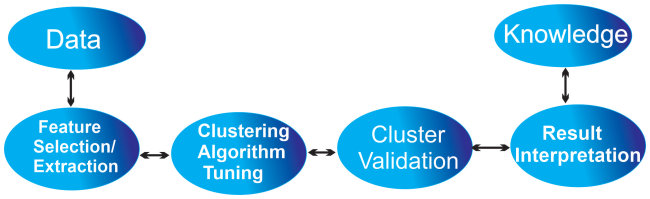


Figure 3. Workflow of clustering in unsupervised ML

A. *Types of Unsupervised Learning*

1) *Clustering*

The practice of classifying items into groups is known as clustering or cluster analysis. Clustering may be divided into several forms, including partitioning, hierarchical, overlapping, and probabilistic. Data is partitioned so that each piece of information may only belong to one cluster [16]. Exclusive pooling is another term for it. K-means exemplify partitioning. Each data point is a cluster in hierarchical clustering. The number of clusters is reduced via iterative connections between the two closest clusters. It is used to organize data in overlapping fuzzy sets [17], [18] . Figure 4 shows the example of clustering. Each point can be

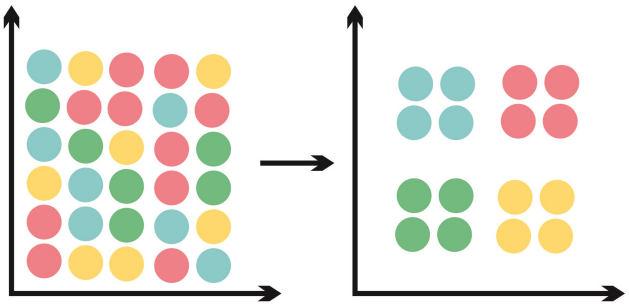


Figure 4. Example of Clustering

assigned to two or more categories with varying degrees of affiliation. In this case, the data will be paired with a suitable membership value, such as K-Means clustering. Finally, the probability distribution is used to generate clusters in probabilistic [19]. Clustering may be divided into three categories according to how they operate, as shown in Figure 5.

Clustering may be divided into three categories according to how they operate, as shown in Figure 5.

- **Exclusive Clustering:** As the name implies, complete clustering states that a data item or object may only exist within a cluster [14].
- **Hierarchical Clustering**: Hierarchical clustering [20] seeks to establish a hierarchy of clusters. Agglomerated and divided hierarchical grouping are the two forms of the hierarchical set. Agglomeration uses a

Figure 5. Types of Clustering

bottom-up method, first considering each data point as a single group, then merging the pairs of groups as they advance up the hierarchy. Agglomerate is the polar opposite of divisive. As you move down the structure, each data point begins in a single group and then divides.

- **Overlapping Clustering:** You may use overlapping grouping to divide a data point into two or more groups [21].

2) Association

The unsupervised learning approach of Association Rule Learning (ARL) is used to uncover associations between variables in massive datasets. ARL can accept non-numeric data points, unlike specific machine learning methods. In a nutshell, ARL is concerned with how particular variables are linked. People who purchase a motorcycle, for example, are more likely to get a helmet [22]. It's possible to make money by forming these kinds of connections. Suppose consumers who purchase product X also purchase product Y, and an online shop can suggest product Y to anybody who buys product X. Internally. In that case, statements are used to learn the laws of association. These assertions may highlight connections between disparate data sets. Support and trust are used if/when patterns or relationships are identified. The media determines the frequency with which the if / then connection appears in the database. The number of times the if/then the relationship has been determined to be legitimate is called confidence. The association rule allows shopping cart analysis and online use mining [23].

3) Anomaly Detection

Any procedure that discovers outliers in a data set is known as anomaly detection. These anomalies might suggest unusual network activity, a faulty sensor, or data that has to be cleaned before analysis. When data models go beyond or diverge from usual models, this is an anomaly. An unusual network traffic pattern, for example, indicates that the compromised system is transferring sensitive data to an unauthorized server. Anomalies are identified or predicted by finding or forecasting data points that differ from the standard model [24], [25]. Intrusion detection, insurance, fraud detection, and military surveillance are just a few of the uses for anomaly detection.

4) Autoencoders

Autoencoders are an unsupervised learning approach that uses neural networks to do representation learning. We'll create a neural network architecture with a bottleneck that compels the network to use a compressed knowledge representation of the original input [26]. This compression and subsequent reconstruction would be complicated if the input properties were unrelated. If the data has some structure (for example, correlations between input attributes), this structure may be learned and utilized by driving the input through the network bottleneck. The presence of an information bottleneck is a fundamental feature of our network design; without it, our network might quickly learn to store input data by passing them through the web [27]. Autoencoder algorithm components are shown in Figure 6.

Figure 6. Autoencoder Components

3. REVIEW ON ML UNSUPERVISED ALGORITHM

We thoroughly analyze the literature on unsupervised learning methodologies and algorithms and performance measures used in unsupervised learning. The benefits and drawbacks of various unsupervised learning research in this paper. This research will help academics go in a new direction by identifying new research areas and filling a research vacuum in unsupervised learning [28]. Researchers will be able to compare the efficiency and efficacy of unsupervised learning algorithms as part of this project. Algorithms are used to implement both clustering and association rule learning. Some of the most significant algorithms used to apply the association rule include the a priori method, the ECLAT algorithm, and the frequent model growth (FP) algorithm. Algorithms like k-mean clustering and principal component analysis make clustering possible (PCA), as shown in Figure 7 [29], [30].

A. Apriori Algorithm

The Apriori algorithm was created with data mining in mind. It may be used to extract data from databases with many transactions, such as a database containing a list of things purchased by supermarket customers. It's used to detect the collection of things that customers are more likely to buy together in shopping cart analysis and to identify the adverse effects of medications [31].

The stages for the Apriori algorithm are as follows:

Figure 7. Widely used Unsupervised Learning Algorithms

- **Step 1:** Determine the transactional database's support for the article sets and select the slightest degree of support and reliability.
- **Step 2:** Select any available media with a more excellent support value than the minimum or specified support value.
- **Step 3:** Find all the rules in these subgroups with a higher confidence value than the threshold or minimal confidence.
- **Step 4:** Sort the rules in ascending order by their elevation.

1) Advantages

- This is a straightforward algorithm.
- On big datasets, the algorithm's join and prune phases are simple to implement [32].

2) Disadvantages

- In comparison to other algorithms, the a priori algorithm is sluggish.
- Because it checks the database many times, overall performance may suffer.
- The a priori algorithm's spatial and temporal complexity is O (2D), which is exceptionally high. The horizontal breadth of the database is represented by D [33].

3) Applications

- Extraction of association rules in data mining of admitted students based on features and specializations in the educational area [34].
- In the medical profession, for example, patient database analysis.
- In forestry, data from forest fires is used to analyze the frequency and intensity of forest fires.
- Many firms employ it a priori, such as Amazon in their referral system and Google in their autocomplete feature.

B. ECLAT Algorithm

ECLAT (Equivalence Class Clustering and Bottom-up Lattice Traversal) is a data mining technique used to acquire element set mining and locate frequent items. Because the a priori technique utilizes a horizontal data structure, it must scan the database numerous times to find frequently occurring objects. On the other hand, ECLAT takes a vertical approach and is faster in general since it only has to scan the database once [35].

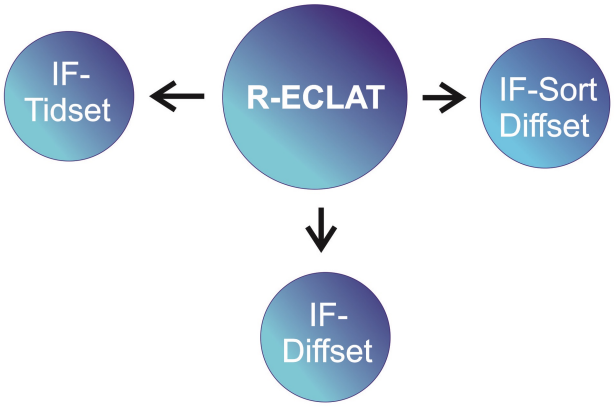


Figure 8. ECLAT Algorithm Model

The stages for the ECLAT algorithm are as follows:

- **Step 1:** For each item in the database, get a list of the transaction ID. We scan the entire database in this step. The list of transactions containing element an is the t transaction ID list of element a.
- **Step 2:** Create a new list of transactions whose members are transactions involving elements a and b by intersecting the list of element titles a with the list of element titles b.
- **Step 3:** Apply conditional a to other items in the database and repeat step 1.
- **Step 4:** Repeat the preceding procedures for the remaining items.

1) Advantages

- The Eclat method utilizes less memory than the Apriori algorithm since it uses a deep search technique [36].

- Eclat's approach does not require repetitive input scanning to compute individual support values.
- Unlike Apriori, which scans the original dataset, the Eclat algorithm searches the recently created datasets.

2) Disadvantages

- The Eclat algorithm uses more RAM to build intermediate transaction ID sets.

3) Applications

- In the medical profession, for example, patient database analysis.
- In forestry, data from forest fires is used to analyze the frequency and intensity of forest fires [36].

C. Frequent Pattern Growth Algorithm

The Apriori algorithm has been improved with the Frequent Pattern (FP) Growth algorithm. This algorithm represents the database in the form of a pattern or frequent tree (FT) structure. The most common patterns are extracted using this regular tree. The Apriori technique must search the database n + 1 times (where n is the most extended model's length), but the FP growth approach needs two scans [37]. The stages for the Frequent pattern (FP) growth algorithm are as follows:

- **Step 1:** The first step is to run a database scan to see whether there are any occurrences of the item sets. This is the same as the first step in the Apriori method. The supporting count or frequency of one collection of items in the database is the frequency of one set of things.
- **Step 2:** The FP tree is built. Begin by constructing the tree's root. The word null is used to represent the root.
- **Step 3:** Re-scanning the database and going over all the transactions is the next stage. Examine the first transaction to see what items it contains. The highest-counting things are taken first, then the lowest-counting items, etc. It denotes that the tree branch is built up of sets of transaction components in decreasing order of count.
- **Step 4:** The next transaction in the database is examined. The object sets are ordered in ascending order by count. If a group of components from this transaction already exists in the root, this transaction branch will share a common prefix. This signifies that the standard item set is linked to the new node of another item set in this transaction.
- **Step 5:** The item set count increases as transactions are made. As new nodes are established and joined based on transactions, the count of both the familiar and new nodes increases by one.
- **Step 6:** The constructed FP tree must now be extracted. The lowest node, as well as the relationships between the weakest nodes, are evaluated first. The lowest node represents the length of frequency model 1. Then take the way via the FP tree. The conditional model base refers to this path or pathways. The dependent model is based on a secondary database containing prefix pathways in the FP tree that begin at the lowest node (suffix).
- **Step 7:** Count the number of sets of items in the path to create a conditional FP tree. The hanging FP tree considers the collections of items that pass the support criterion.
- **Step 8:** Create a conditional FP tree by counting the number of sets of items in the route. The hanging FP tree considers the groups of things that pass the support threshold.
- **Step 9:** The conditional FP tree generates frequent patterns.

1) Advantages

- This approach only needs to scan the database twice, compared to Apriori, which examines the transactions for each iteration [38].
- This method avoids item matching, which speeds up the process.
- Extraction of long and short frequent patterns is efficient and scalable since the database is compressed in memory.

2) Disadvantages

- FP Tree is bulkier and more complicated to build than Apriori, and it might be rather expensive.
- The approach may not fit in shared memory if the database is extensive.

3) Applications

- Clustering, classification, software issue identification, recommendations, and other problems may all be solved with a Frequent pattern (FP) growth algorithm [38].

D. Clustering using K-Means

In data science, several rounds of the k-means method are commonly utilized. The k-means clustering algorithm divides components into groups based on their similarity. Graphically representation of the K-mean clustering workflow is shown in Figure 9.

The letter k denotes the number of groups. As a result, if k is 3, there will be three groupings [39], [40], [41]. This clustering algorithm divides the unlabeled dataset into unique clusters with comparable qualities for each data

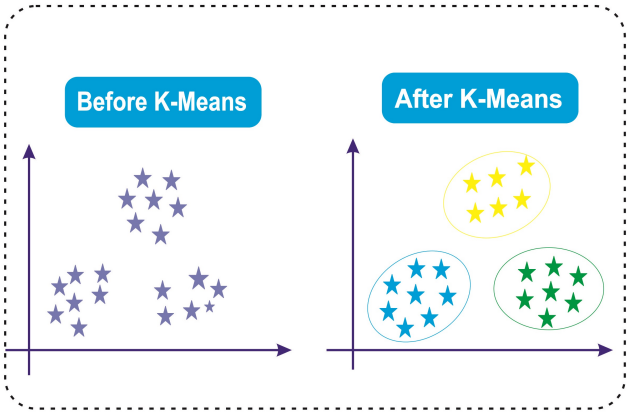


Figure 9. Workflow of K-mean Clustering

point. The trick is to locate cluster centroids, which are K centers. Each group will have a cluster centroid. After presenting a new data point, the algorithm will use metrics like Euclidean distance to identify which cluster the data point belongs to. The centroids are calculated using the K-mean clustering technique, which iterates until the best centroid is obtained. The number of groupings is likely known. Another term for it is the flat clustering algorithm. The number of groups found by the algorithm is denoted by the letter "K" in K-means [42].

The steps below will demonstrate how the K-Means clustering approach works:

- **Step 1:** We must first specify the number of groups (K) that this method should produce.
- **Step 2:** Next, pick K random data points and assign them to one of the groups. In a nutshell, it ranks data according to the number of data points it contains.
- **Step 3:** Now, we'll calculate the cluster centroids.
- **Step 4:** Repeat steps 1-3 until you discover the optimum centroid, i.e., allocating data points to non-varying groups.
  - **4.1:** To begin, compute the total of the squared distances between the data points and the centroids.
  - **4.2:** Now, we must allocate each data point to the closest group to the others (centroid).
  - **4.3:** Finally, by averaging all data points in the cluster, compute the centroids of the groups.

1) Advantages

- It is simple to comprehend and put into practice.
- If we had a high number of variables, K-mean would be quicker than hierarchical clustering.
- When the centroids are recalculated, the cluster of an instance might be modified.
- K-mean clustering provides smaller groups than hierarchical clustering.

2) Disadvantages

- It's impossible to determine the number of clusters or the value of k.
- The number of groups in a network (k value) and other initial inputs is strongly influenced by the output.
- The order in which the data is input significantly impacts the final result.
- It is pretty sensitive to changes in size. The outcome will be dramatically different if we scale our data using normalization or standards—the ultimate result [42].
- If the nests have a complex geometric shape, nesting is not suggested.

3) Applications

- Segmentation of the market
- Document grouping and picture segmentation
- Compression of images
- Segmentation of customers
- Dynamic data trend analysis

E. Principal Components Analysis (PCA)

PCA is a dimensionality reduction approach that reduces the dimensionality of big data sets by converting many variables into a smaller group that still maintains the majority of the data information in the 'large set.' The loss of precision while lowering the number of variables in a dataset is unavoidable, but the answer to reducing dimensionality is to exchange some accuracy for simplicity. Because smaller datasets are easier to study and visualize and there are no unnecessary variables to analyze, data analysis is easier and faster for machine learning algorithms [43]. The graphical representation of the K-mean clustering workflow is shown in Figure 9.

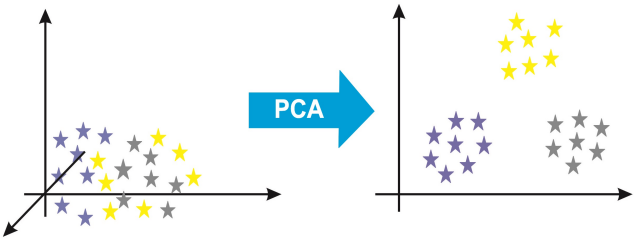


Figure 10. Workflow of PCA

To summarize, PCA's goal is to decrease the number of variables in a data collection while retaining as much information as feasible.

TABLE I. A Comparison of Five Commonly used Unsupervised Algorithms

	Apriori Algorithm	ECLAT Algorithm	Frequent Pattern Growth Algorithm	K-Means Algorithm	Principal Components Analysis Algorithm
Accuracy in General	Satisfactory	Good	Good	Superb	Superb
Speed of Learning	Good	Excellent	Good	Superb	Superb
Speed of Classification	Superb	Superb	Good	Excellent	Superb
Tolerance to Missing Values	Good	Excellent	Superb	Superb	Superb
Tolerance to Irrelevant Attributes	Good	Excellent	Satisfactory	Superb	Good
Tolerance to Redundant Attributes	Excellent	Good	Satisfactory	Excellent	Good
Tolerance to Highly Interdependent Attributes	Satisfactory	Good	Excellent	Excellent	Satisfactory
Tolerance to Noise	Excellent	Good	Satisfactory	Superb	Superb
Dealing with Danger of Overfitting	Good	Good	Superb	Excellent	Excellent
Attempts for Incremental Learning	Superb	Satisfactory	Good	Good	Superb
Transparency of Knowledge/ Classification	Satisfactory	Good	Excellent	Excellent	Good
Support Multi classification	Good	Excellent	Excellent	Superb	Superb

The steps below will demonstrate how the PCA approach works:

- **Step 1:** Obtain the data set
- **Step 2:** Data representation in a structure.
- **Step 3:** Standardization of data
- **Step 4:** Z's covariance is calculated.
- **Step 5:** Eigenvalues and eigenvectors are calculated.
- **Step 6:** The eigenvectors are categorized.
- **Step 7:** New characteristics or primary components are calculated.
- **Step 8:** Remove characteristics from the new dataset that are less significant or irrelevant.

1) Advantages

- PCA helps us to better generalize machine learning models by lowering the dimensionality of the input. This aids us in overcoming the "dimensionality curse" [44].
- The calculation is simple. PCA is based on linear algebra, which computers can solve quickly.
- Other machine learning algorithms will be sped up. Machine learning algorithms are trained on crucial components rather than the original dataset and converge faster.
- Reduces the challenges associated with high-dimensional data. Regression-based algorithms are readily over-adaptable when dealing with high-dimensional data. We avoid overfitting prediction algorithms by utilizing PCA to minimize the size of the training dataset in advance.

2) Disadvantages

- The key components have low interpretability. Principal components are linear combinations of the original data's features, but they're not easy to understand. For example, it's challenging to identify the dataset's most relevant properties after computing the main components.
- The trade-off between dimensionality reduction and information loss. Reduced dimensionality is beneficial, but it comes at a cost. Information loss is an unavoidable aspect of the PCA [44].

3) Applications

- PCA is mainly utilized in artificial intelligence applications such as computer vision, image compression, and a dimensionality reduction approach.
- If the data is vast enough, it may also be utilized to find hidden models. Finance, Data Mining, Psychology, and more areas employ PCA [45].

4. COMPARATIVELY ANALYSIS

So, here's a comparison of the most popular unsupervised classification algorithms. Several strategies have been created, some of which have been addressed in earlier sections. Based on available facts and theoretical studies, Table 1 compares various regularly used unsupervised algorithms. This comparison demonstrates that no single learning algorithm beats the others.

5. CONCLUSION

Unsupervised learning is one of the many types of machine learning. The model is trained on an unlabeled dataset in unsupervised learning. Grouping, association, anomaly detection, and automated encoders are also included. Various techniques for unsupervised learning have been presented throughout the last decade. Unsupervised learning has many applications, from intrusion detection to information retrieval, disease diagnosis, and protein sequence search. This review of the literature focuses on unsupervised learning methodologies and algorithms and the numerous assessment metrics used to evaluate the performance of unsupervised learning models. It also outlines the advantages and disadvantages of each study. This survey report will aid academics in determining which unsupervised learning algorithms or approaches to utilize for issue solving. Also, which study field needs greater attention. The scope of this research is confined to commonly used unsupervised learning techniques. Only research within the last five years should be highlighted. We may operate more algorithms and methodologies in the future to improve targeting.

A. Acknowledgment

The authors would like to thank the referees for their careful reading and for their comments, which significantly improved the paper. Additionally, thanks to Dr. Salman Qadri, (Associate Professor, Chairman Department of Computer Science, MNS University of Agriculture, Multan, Pakistan) and Dr. Farrukh Jamal, (Assistant Professor, Department of Statistics, The Islamia University of Bahawalpur, Pakistan) for his motivational support.

B. Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] J. Alzubi, A. Nayyar, and A. Kumar, "Machine learning from theory to algorithms: an overview," in *Journal of physics: conference series*, vol. 1142, no. 1. IOP Publishing, 2018, p. 012012.

[2] Z. Zeng, Y. Li, Y. Li, and Y. Luo, "Statistical and machine learning methods for spatially resolved transcriptomics data analysis," *Genome biology*, vol. 23, no. 1, pp. 1–23, 2022.

[3] J. G. Greener, S. M. Kandathil, L. Moffat, and D. T. Jones, "A guide to machine learning for biologists," *Nature Reviews Molecular Cell Biology*, vol. 23, no. 1, pp. 40–55, 2022.

[4] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, and A. J. Aljaaf, "A systematic review on supervised and unsupervised machine learning algorithms for data science," *Supervised and unsupervised learning for data science*, pp. 3–21, 2020.

[5] R. Saravanan and P. Sujatha, "A state of art techniques on machine learning algorithms: a perspective of supervised learning approaches in data classification," in *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE, 2018, pp. 945–949.

[6] D. Krotov and J. J. Hopfield, "Unsupervised learning by competing hidden units," *Proceedings of the National Academy of Sciences*, vol. 116, no. 16, pp. 7723–7731, 2019.

[7] H. U. Dike, Y. Zhou, K. K. Deveerasetty, and Q. Wu, "Unsupervised learning based on artificial neural network: A review," in *2018 IEEE International Conference on Cyborg and Bionic Systems (CBS)*. IEEE, 2018, pp. 322–327.

[8] N. Li, M. Shepperd, and Y. Guo, "A systematic review of unsupervised learning techniques for software defect prediction," *Information and Software Technology*, vol. 122, p. 106287, 2020.

[9] E. Acar and B. Yener, "Unsupervised multiway data analysis: A literature survey," *IEEE transactions on knowledge and data engineering*, vol. 21, no. 1, pp. 6–20, 2008.

[10] V. Kavitha and M. Punithavalli, "Clustering time series data stream-a literature survey," *arXiv preprint arXiv:1005.4270*, 2010.

[11] N. Grira, M. Crucianu, and N. Boujemaa, "Unsupervised and semi-supervised clustering: a brief survey," *A review of machine learning techniques for processing multimedia content*, vol. 1, pp. 9–16, 2004.

[12] R. A. Bantan, A. Ali, S. Naeem, F. Jamal, M. Elgarhy, and C. Chesneau, "Discrimination of sunflower seeds using multispectral and texture dataset in combination with region selection and supervised classification methods," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 30, no. 11, p. 113142, 2020.

[13] J. Gao, C. Zhong, X. Chen, H. Lin, and Z. Zhang, "Unsupervised learning for passive beamforming," *ieee communications letters*, vol. 24, no. 5, pp. 1052–1056, 2020.

[14] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 132–149.

[15] M. Akram, S. Siddique, and M. G. Alharbi, "Clustering algorithm with strength of connectedness for m-polar fuzzy network models," *Mathematical Biosciences and Engineering*, vol. 19, no. 1, pp. 420–455, 2022.

[16] M. Z. Rodriguez, C. H. Comin, D. Casanova, O. M. Bruno, D. R. Amancio, L. d. F. Costa, and F. A. Rodrigues, "Clustering algorithms: A comparative approach," *PloS one*, vol. 14, no. 1, p. e0210236, 2019.

[17] S. Ahmadian, A. Epasto, R. Kumar, and M. Mahdian, "Clustering without over-representation," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 267–275.

[18] A. Ali, W. K. Mashwani, M. H. Tahir, S. B. Belhaouari, H. Alrabaiah, S. Naeem, J. A. Nasir, F. Jamal, and C. Chesneau, "Statistical features analysis and discrimination of maize seeds utilizing machine vision approach," *Journal of Intelligent & Fuzzy Systems*, vol. 40, no. 1, pp. 703–714, 2021.

[19] C. S. Saba and N. Ngepah, "Convergence in renewable energy sources and the dynamics of their determinants: an insight from a club clustering algorithm," *Energy Reports*, vol. 8, pp. 3483–3506, 2022.

[20] A. Maji, N. R. Velaga, and Y. Urie, "Hierarchical clustering analysis framework of mutually exclusive crash causation parameters for regional road safety strategies," *International journal of injury control and safety promotion*, vol. 25, no. 3, pp. 257–271, 2018.

[21] H. Ma, "Design of chinese linguistics teaching system based on k-means clustering algorithm," in *The International Conference on Cyber Security Intelligence and Analytics*. Springer, 2022, pp. 424–431.

[22] Y. Zhang, Y. Zhang, Q. Chen, Z. Ai, and Z. Gong, "True-link clustering through signaling process and subcommunity merge in overlapping community detection," *Neural Computing and Applications*, vol. 30, no. 12, pp. 3613–3621, 2018.

[23] M. Li, X. Zhu, and S. Gong, "Unsupervised person re-identification by deep learning tracklet association," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 737–753.

[24] M. V. Prasad, R. Balakrishnan *et al.*, "Spatio-temporal association rule based deep annotation-free clustering (star-dac) for unsupervised person re-identification," *Pattern Recognition*, vol. 122, p. 108287, 2022.

[25] M. A. Kabir and X. Luo, "Unsupervised learning for network flow based anomaly detection in the era of deep learning," in *2020 IEEE Sixth International Conference on Big Data Computing Service and Applications (BigDataService)*. IEEE, 2020, pp. 165–168.

[26] K. Kottmann, P. Huembeli, M. Lewenstein, and A. Acín, "Unsupervised phase discovery with deep anomaly detection," *Physical Review Letters*, vol. 125, no. 17, p. 170603, 2020.

[27] H. Choi, M. Kim, G. Lee, and W. Kim, "Unsupervised learning approach for network intrusion detection system using autoencoders," *The Journal of Supercomputing*, vol. 75, no. 9, pp. 5597–5621, 2019.

[28] J.-H. Seong and D.-H. Seo, "Selective unsupervised learning-based wi-fi fingerprint system using autoencoder and gan," *IEEE Internet of Things Journal*, vol. 7, no. 3, pp. 1898–1909, 2019.

[29] A. I. Károly, R. Fullér, and P. Galambos, "Unsupervised clustering for deep learning: A tutorial survey," *Acta Polytechnica Hungarica*, vol. 15, no. 8, pp. 29–53, 2018.

[30] N. Urs, S. Behpour, A. Georgaras, and M. V. Albert, "Unsupervised learning in images and audio to produce neural receptive fields: a primer and accessible notebook," *Artificial Intelligence Review*, vol. 55, no. 1, pp. 111–128, 2022.

[31] K. La Marca and H. Bedle, "Deepwater seismic facies and architectural element interpretation aided with unsupervised machine learning techniques: Taranaki basin, new zealand," *Marine and Petroleum Geology*, vol. 136, p. 105427, 2022.

[32] P. Edastama, A. S. Bist, and A. Prambudi, "Implementation of data mining on glasses sales using the apriori algorithm," *International Journal of Cyber and IT Service Management*, vol. 1, no. 2, pp. 159–172, 2021.

[33] P. B. I. S. Putra, N. P. S. M. Suryani, and S. Aryani, "Analysis of apriori algorithm on sales transactions to arrange placement of goods on minimarket," *International Journal of Engineering and Emerging Technology*, vol. 3, no. 1, pp. 13–17, 2018.

[34] X. Xie, G. Fu, Y. Xue, Z. Zhao, P. Chen, B. Lu, and S. Jiang, "Risk prediction and factors risk analysis based on ifoa-grnn and apriori algorithms: Application of artificial intelligence in accident prevention," *Process Safety and Environmental Protection*, vol. 122, pp. 169–184, 2019.

[35] L. Jia, L. Xiang, and X. Liu, "An improved eclat algorithm based on tissue-like p system with active membranes," *Processes*, vol. 7, no. 9, p. 555, 2019.

[36] W. Mohamed and M. A. Abdel-Fattah, "A proposed hybrid algorithm for mining frequent patterns on spark," *International Journal of Business Intelligence and Data Mining*, vol. 20, no. 2, pp. 146–169, 2022.

[37] C.-H. Chee, J. Jaafar, I. A. Aziz, M. H. Hasan, and W. Yeoh, "Algorithms for frequent itemset mining: a literature review," *Artificial Intelligence Review*, vol. 52, no. 4, pp. 2603–2621, 2019.

[38] L. Gubu, D. Rosadi *et al.*, "Robust mean–variance portfolio selection using cluster analysis: A comparison between kamila and weighted k-mean clustering," *Asian Economic and Financial Review*, vol. 10, no. 10, pp. 1169–1186, 2020.

[39] S. Park, Y. Hwang, and B.-J. Yang, "Unsupervised learning of topological phase diagram using topological data analysis," *Physical Review B*, vol. 105, no. 19, p. 195115, 2022.

[40] P. An, Z. Wang, and C. Zhang, "Ensemble unsupervised autoencoders and gaussian mixture model for cyberattack detection," *Information Processing & Management*, vol. 59, no. 2, p. 102844, 2022.

[41] M. Rashid, H. Singh, and V. Goyal, "Cloud storage privacy in health care systems based on ip and geo-location validation using k-mean clustering technique," *International Journal of E-Health and Medical Communications (IJEHMC)*, vol. 10, no. 4, pp. 54–65, 2019.

[42] M. Mateen, J. Wen, S. Song, and Z. Huang, "Fundus image classification using vgg-19 architecture with pca and svd," *Symmetry*, vol. 11, no. 1, p. 1, 2018.

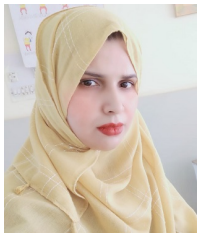
[43] Y. Dong and S. J. Qin, "A novel dynamic pca algorithm for dynamic data modeling and process monitoring," *Journal of Process Control*, vol. 67, pp. 1–11, 2018.

[44] S. Ghani, S. Kumari, and A. Bardhan, "A novel liquefaction study for fine-grained soil using pca-based hybrid soft computing models," *Sādhanā*, vol. 46, no. 3, pp. 1–17, 2021.

[45] D. H. Grossoehme, M. Brown, G. Richner, S. M. Zhou, and S. Friebert, "A retrospective examination of home pca use and parental satisfaction with pediatric palliative care patients," *American Journal of Hospice and Palliative Medicine®*, vol. 39, no. 3, pp. 295–307, 2022.

SAMREEN NAEEM got her Bachelor's Degree in Information Technology (IT) from Sargodha University Pakistan, after that she enrolled and completed her M.Phil Degree in Computer Science from The Islamia University of Bahawalpur, Pakistan. She is also work as Lecturer Computer Science and IT in reputed institutes in Pakistan. Now she is pursuing a PhD at the Southeast University China to complete her education.

AQIB ALI got his Bachelor's Degree in Computer (2017), after that he enrolled and completed his M.Phil. Degree in Computer Science (2020) from The Islamia University of Bahawalpur, Pakistan. He is also working as Lecturer Computer Science and IT in reputed institutes in Pakistan. Now he is pursuing a Ph.D. degree from the Southeast University China to complete his education.



Sania Anam got her Bachelor's Degree in Computer (2007),Master degree in Computer (2009) after that he enrolled and completed his M.Phil. Degree in Computer Science (2016) from The Islamia University of Bahawalpur, Pakistan. Since 2016, she is working as Lecturer in Computer Science in Govt Associate College for Women Ahmad pur East ,Bahawalpur, Pakistan.

MUHAMMAD MUNAWAR AHMED completed his Bachelor's Degree in Computer (2005), after that he completed his MSCS Degree in session (2011-13) from The Islamia University of Bahawalpur, Pakistan. He is also working as Lecturer, department of Information Technology at The Islamia University of Bahawalpur, Pakistan. Currently, he is enrolled in Ph.D. program at The Islamia University of Bahawalpur, Pakistan.