*Grenze International Journal of Engineering and Technology, June Issue*

# A Comprehensive Study on Current Trends in Unsupervised Machine learning Algorithms and Challenges in Real World Applications

Dr. P. Velvadivu[1], Dr. M. Sujithra[2] and R. Priyadharshini[3]
[1-2]Assistant Professor, Department of Computing - Data Science, Coimbatore Institute of Technology
Email: {velvadivu, sujithra}@cit.edu.in
[3]MSC Data Science, Department of Computing, Coimbatore Institute of Technology
Email: devidharshini001@gmail.com

***Abstract*—In this technology fueled world, everyone is using smart devices, electronic gadgets, wireless products etc. and huge amounts of data are being generated, collected, and stored in the databases. To efficiently process and intelligently analyze the huge amount of data, the knowledge about subfield of Artificial Intelligence that is, Particularly Machine learning (ML) is required. There are various types of machine learning and its algorithms have been introduced to handle real world scenarios. This paper discusses the Comprehensive survey based on methodologies, techniques, algorithms, applications, and challenges faced by unsupervised machine learning and how unsupervised learning techniques can be helpful in real world business and environment. Thus, this study's key contribution is explaining the principles of different unsupervised machine learning techniques and their applicability in various real-world application domains, such as cybersecurity systems, smart cities, healthcare, e-commerce, agriculture, and many more.**

***Index Terms*— Machine learning, Unsupervised learning, clustering, feature selection and featureextraction.**

## I. INTRODUCTION

Very large and enormous amount of data and information are collected and stored from various sources like mobile phones, personal computers, sensors, cameras, satellites, log files, health care tracker, bio informatics, human generated data like social media data where enormous number of photos, videos, audios have been uploaded daily on the internet. Every day, 2.5 quintillion bytes of data generated roughly. Intelligently collecting, processing, analyzing huge volumes of data and developing corresponding smart gadgets, automated applications using the knowledge of Artificial Intelligence and Machine learning. Machine learning allows software applications and programs to automatically predict with accurate accuracy without being explicitly programmed. Machine learning is the most important field of Data Science. In real world, through learning capability and experience human tries to learn and Machine works based on human instructions. Machine learning is the one where machine automatically learns by experience as human does. The role of Machine learning is to learn, improve performance by experience and predict things with best accuracy. Machine learning have been classified into supervised, unsupervised and reinforcement learning

*Grenze ID: 01.GIJET.9.2.41*

*A. Steps followed in Machine learning process*

- *Gathering Data*: Raw data either from excel, access or text files. This step (Collecting past data) forms a strong foundation for future learning.
- *Data Preparation*: Raw data from any source contains missing values, irrelevant data, outliers etc. This technique involves data cleaning, normalization, dimensionality reduction, treatment of outliers and methods to remove irrelevant data.
- *Choosing a model*: Before choosing a model, there is a need to identify which type of machine learning the problem statement is. Then choosing a right machine learning model under any of the machine learning category plays an important role for future prediction.
- *Training*: Normally, 70% of data will be used for training part and remaining 30% of dataset will be used for testing or evaluation part. Training the machine learning model helps the model to understand and train by its own with well understanding of dataset.
- *Evaluation*: 30% of testing dataset is used to test the machine learning algorithm and check how well the algorithm is trained based on performance measures like accuracy, precision, recall etc.
- *Hyperparameter tuning*: Hyperparameter tuning is a parameter whose value will be set before actual training process begins.
- *Prediction*:

*B. Machine learning types*

*1) Supervised learning*

In Supervised learning, the model gets trained based on the labelled data. In training phase, the input data will get trained and tested with the target attribute. Supervised learning is also called task oriented because it mainly focusses on task and feed more data to train algorithm until it accurately predicts and perform. Supervised algorithm has been classified into two types, Classification and Regression. Some of the algorithms are K Nearest Neighbor, Random Forest, Decision tree, Support Vector Machine, Logistic Regression etc.

*2) Unsupervised learning*

It mainly focusses on identifying underlying trends, patterns, and insights from the dataset. Here, unsupervised learning models are trained using unlabeled dataset and automatically extract patterns, facts and figures without any supervision. Unsupervised learning mainly classified into two types, Clustering and Association. Several algorithms follow these two types namely Agglomerative hierarchical clustering, K means clustering and FP Growth, Apriori algorithm (Association) problem. Some applications are Recommendation system, Identity management etc.

*3) Reinforcement learning*

Feedback based reinforcement learning agent trains automatically using learning, feedback, and previous experience. The agent will be rewarded if it does right action in the environment, and it will be penalized if it does any wrong actions. Many applications have been developed using reinforcement learning techniques like, Gaming technology, Robotics, Self-driving cars etc.

4) *Difference between machine learning types*

TABLE I. DIFFERENCE BETWEEN MACHINE LEARNING TYPES

| **Supervised learning** | **Unsupervised learning** | **Reinforcement learning** |
|---|---|---|
| Machine learning model or algorithms learns from labeled data. It is also called as Task oriented approach. | Machine learning model or algorithms learns from unlabeled data. It is also called as Data Driven Approach. | Reinforcement learning models are based on reward or penalty. It is also called as Environment Driven Approach. |
| **Types:** Classification or Regression | **Types:** Clustering and Association rules. | **Types:** Classification and Control |
| **Algorithms:** Random Forest, Decision tree, K NearestNeighbor | **Algorithms:** K Means Clustering, DBSCAN Algorithm, Principal Component Analysis | It is formalized using Markov Decision process. |
| **Applications:** Medical Diagnosis, Spam Detection | **Applications:** Recommendation systems, Customer segmentation. | **Applications**: Robotics, Video games. |

## II. UNSUPERVISED MACHINE LEARNING

Unsupervised learning, by the name itself it could be easily understood that it will not guide by any supervision. This type of learning should automatically extract knowledge, underlying hidden patterns, data groupings from the dataset without human intervention. It will group objects or items based on similarities. Unsupervised machine learning will deal with unlabeled dataset where there is no target output tagged with corresponding input. Hence unsupervised learning is helpful in real life scenarios because all real-world problems will not come up with input and output pattern.

### *A. Steps involved in Unsupervised learning techniques*:

- *Gathering Unlabeled data*: In unsupervised machine learning, gathering of data (raw unlabeled data) isthe important part where it finds insights and trends from the data without supervision.
- *Interpretation*: It interprets the raw input data to find out the hidden patterns and trends.
- *Algorithm*: Then will apply suitable algorithms like clustering algorithms or association rules.
- *Processing*: Here, the data points divide into groups called clusters based on similarity which is measured using Euclidean or cosine distance.
- *Output*: When new data point arrives, the algorithm will push the data point into most similar groups andgives the predicted output based on similarity without any supervision.

### *B. Advantages of using Unsupervised learning*

- Unsupervised learning helps to solve problems without human intervention.
- It automatically learns from the data and discover underlying patterns and group items or objects basedon similarities.
- It is less complex when compared with supervised machine learning, because in supervised it involveshuman intervention because one has to understand the input data and label them.

### *C. Disadvantages of using Unsupervised learning*

- It provides less accuracy of results because it has no labeled data and machine must discover automatically the underlying new patterns and relationships hence provides somewhat less accuracy compared to supervised machine learning.
- Evaluating an unsupervised machine learning model is quite difficult when compared to supervised model.

### *D. Types of Unsupervised learning technique*

In unsupervised we have input data and not having corresponding output data. If there is a set of image dataset, then algorithm does not know about the input features and not trained upon images provided. The unsupervised model should try to learn upon their own and perform the task by clustering or grouping the images based on similarities. Unsupervised have been classified into two types,

- Clustering
- Association

### *E. Clustering*

Clustering is grouping of objects based on similarities. It groups the given data points and objects that possess more similarity will remain in same group and objects that possess less similarity will move to other groups of clusters. It can be helpful in marketing sectors or industries where they group customers based on their behavior. Clustering has been used in wide range of applications like e-commerce sites, cybersecurity, health care analytics, behavioral analytics etc. Many clustering algorithms has been introduced, most popular and widely used clustering algorithms that is used in machine learning is,

- K-Means Clustering
- Agglomerative hierarchical clustering
- DBSCAN Clustering

### *F. Association Rules*

Association rules is a type of unsupervised learning method which is used to find the relationship between the objects in the large databases. Association rule mining makes effective marketing strategy. Market basket analysisis a one example for association rule mining since it finds relationship between the items purchased by

the customers. If a person buys product ‘x’, then he/she might buy product ‘y’. If a customer who doesn’t buy product‘y’ followed by ‘x’ then they are said to be typical customers and marketing agents target them and cross sell the items to them. Association rule mining finds frequent items, pairs, associations etc. from relational or transactionalor any kind of databases.

It has been divided into two parts,

- Antecedent
- Consequent

“*If customer purchases the product bread, then he is likely to buy Jam*”

- ***Antecedent:*** It can be found in datasets. It is bread from above statement.
- ***Consequent***: It can be found in combination with Antecedent. It is Jam from above statement.

The relationship can be described in two parameters, “Support” and “Confidence”. Support indicates how many times the if/then relation occurs in the datasets, whereas Confidence refers to number of times these if/then relationships have found to be true.

There are different types of algorithms in association rule mining,

- Apriori algorithm
- FP – Growth algorithm

## III. CLUSTERING ALGORITHM

### *A. K – Means Clustering*

It is widely used clustering algorithms in Unsupervised machine learning. K- Means clustering is an iterative procedure where the data points are grouped into K clusters. Each data point belongs to one single cluster. It groups data points based on similarity, the similarity between the data points can be determined by calculating distance between them. The distance between data points and clusters centroid point (Initially, which is the random data point selected among all data points) can be calculated in many ways,

- Squared Euclidean distance
- Manhattan Distance
- Cosine Distance
- Correlation Distance

Similarity can be measured using any of these distance-based measurement and it is completely applicationspecific.

*1. Working of K – Means clustering*

- Initially, determining the k-value, where ‘k’ is denoted as number of cluster centroids among all datapoints.
- Cluster centroids: Randomly selecting k data points that is, if k=5, then randomly choosing 5 data pointsfrom groups of data.
- After selecting centroids, calculating the distance between each data point with centroid using any of thedistance-based measures.
- The minimum distance between centroid and corresponding data point will form a cluster formation.
- Similarly, calculating distance for all the data points, the data points which is at minimum distance fromone cluster will get joined with that cluster group. Thus, forming k clusters.
- Since, this is an iterative procedure, there are two steps
    - Assigning the data points
    - Updating the clusters.
- Updating clusters will occur again and again until there is no assigning of data points from one cluster toanother.
- Finally, it groups the data points with k clusters based on similarity.
- There are two ways of selecting k value, (a) Elbow Method and (b) Silhouette Method.

*2. Elbow Method*

Elbow method is used to identify the optimal number of k clusters in the dataset. It determines whether the selected k value will provide optimal accuracy of grouping the data points. To understand, Initially, fix k: 1, it forms one single cluster where all data points belong to one cluster. Similarly, fix k value as 2, then data points have been divided into two clusters. When k value increases, the distance between data points and cluster centroid points decreases. It is said to be optimal if k value is above 3 because the distance decreases rapidly due

to increase in k value. When k is 3 or above, the distance between data point and centroid becomes minimum and become stable. So, selecting k points above 3 can be optimal solution for identifying the number of clusters.

*3. Silhouette Method*

It is used to find the accurate separation of k clusters in the dataset. It can be calculated using the formula,

$$s(o) = \frac{b(o) - a(o)}{\text{Max}\{a(o), b(o)\}}$$

s(o) = silhouette coefficient of data point 'o'

b(o) = computing average distance between data point 'o' to all other clusters.

a(o) = computing average distance between data point 'o' to other data points in same cluster. Silhouette coefficient should range from [-1,1]. If silhouette coefficient is,

- 1: It is the best coefficient value and the number of k cluster selected will correctly groups between allthe data points.
- -1: It denotes worst k value selection and it does not groups data points correctly.
- 0: It is said to be overlapping of clusters.

So, the silhouette value should be high as possible and closed to 1. Compared to elbow method, silhouettemethod gives the best separation of clusters between the data points.

*B. Agglomerative Hierarchical Clustering*

Agglomerative hierarchical clustering is a type of hierarchical clustering and widely used clustering algorithm. Initially, all the data points are considered as one single cluster. Calculating Euclidean distance or any kind of distance measure between each single data point or each single cluster. The minimum distance between pair of clusters will be combined as one single cluster. Similarly, calculating each cluster's distance and that with minimum distance merges, thus forming one single cluster or k clusters finally.

*1. Working of Agglomerative hierarchical clustering*

- All the data points are considered as single cluster.
- Computing proximity matrix for all the data points.
- Groups data points based on minimum distance.
- Updating the proximity matrix after merging the clusters.
- Similarly, repeating the process until it attains one single cluster or k cluster.
- Agglomerative hierarchical clustering can be visualized using a chart called "dendrogram"
- Initially if there are 5 data points say A, B, C, D, E, each data points are considered as single cluster.
- Distance between each data point is calculated.
- The minimum distance is found between A and B, D and E.
- Both clusters form a single cluster say, AB and DE
- Again, distance between AB, C, DE has been calculated.
- The minimum distance found between AB, C they form a single cluster.
- Similarly, Distance calculation is carried out for all the data points and finally they form a single clusteras ABCDE.

*C. DBSCAN Clustering*

There is a drawback in k-means clustering and agglomerative hierarchical clustering as its fail to create arbitrary shapes. Thus, DBSCAN helps to overcome this issue. DBSCAN groups data based on high density. The most interesting part in DBSCAN is it is robust to outliers, and it can easily detect the noise from the group of data points. In k means clustering, there is a need to determine k value to form number of clusters, but in DBSCAN there is no need to specify k value or number of clusters. It requires two parameters,

- **Epsilon**: Radius of circle to be formed around data points.
- **Min-points**: Minimum number of points to be inside radius of circle.

- DBSCAN classifies data points into three types, (a) Core Point (b) Border point (c) Noise. Consider Min point: 3, then the number of data points with at least 3 inside the circle with including itself is represented as "core point". If the data point within 3 but greater than 1 can be represented as "border point". If it contains only one data point inside the circle it is noise. DBSCAN uses Euclidean distance-based measure to calculate distance between points. To use DBSCAN technique it is much more important to select epsilon and Minimum point value. Selecting epsilon value and Min point follows some criteria, the minimum point

should be one value greater than number of dimensions. Min Point = Dimension+1. Epsilon value can be decided using elbow graph or k distance graph. The maximum curvature value can be selected as epsilon to get more accurate value. Hence, DBSCAN approach can be more useful for clustering related problem as when compared to k- means clustering or Hierarchical clustering. But K-means and hierarchical clustering can be useful for some applications also. Density based problems can be solved using DBSCAN algorithm.

## IV. Association Rule Mining

### *A. Apriori Algorithm*

Apriori algorithm is used in Market basket analysis, which is used to find the relationship between two products, items, or objects and find the frequent item set from candidate k items. If a person buys a product 'x', then he/she might buy product 'y' in combination with x. The main goal of Apriori algorithm is to find the association rules between items. It is also called as frequent pattern mining, which finds frequent items purchased by the customer. Hence, this is helpful in marketing companies or industries to find the customer behavior and cross sell items to the customers. Apriori algorithm operates on the databases with large number of transactions. Three main components to calculate Apriori algorithm is, support, confidence, and lift.

1. *Working of Apriori algorithm*
    - Consider for example if support count is 2 and confidence is 60%.
    - The dataset contains transaction id and frequent item set purchased by a customer.
    - If there are 5 items say, A, B, C, D, E. and k: 1 initially then find the number of transactions for each item.
    - If the support count for each item is less than the minimum support count 2, then prune that item.
    - Next, again check with k=2, combine 2 items and check the combination of transactions between two items. Again, if the support count for two items less than minimum support count prune it. Similarly, theprocess will be repeated until there is no itemset leftover with minimum support count.
    - Finally, generate rules for result set of items and check with confidence value of item set with given confidence value i.e., 60%. If the confidence value less than given value, reject that pair and item set with maximum confidence value will be selected as frequent item pairs purchased by customers.
    - This will get to know about the customer and enhance their purchasing experience and marketing sectorswill gain more profit.

### *B. FP - Growth*

There is a drawback in Apriori algorithm, since it must scan the database again and again to find the candidate k items it makes algorithm slower. To overcome this issue, FP Growth has been introduced. It is also calledas Frequent pattern growth algorithm.

- *Working of FP-Growth*
    - Consider the support value to be 3. The items with equal or maximum support value are sorted indescending order.
    - Creating ordered item set, comparing the actual item set with sorted item (item with more than 2 supportcount) and considering only the item in ordered item creating a separate set.
    - Now all the ordered item set will be inserted in trie data structure.
    - Next step is to find the conditional pattern base by using the trie data structure.
    - After that identifying conditional frequent pattern is important to identify frequent item pair. Thus,finding frequent pattern by using conditional frequent pattern.
    - Finally, calculating confidence value for frequent item and comparing with given confidence value.Identifying, if then rules for frequent item sets by using confidence value.

## V. Applications of Unsupervised Learning

There are many real-world applications that make use of unsupervised machine learning and enhancing userexperience by using any of the algorithms.

- Recommender system
- Customer segmentation
- Targeted Marketing
- Identity management, Item categorization

- In Genetics and Anomaly Detection
- Image, speech and pattern recognition

## VI. Challenges and Research Directions

Unsupervised machine learning has lots of benefits and it is helpful in real world problems, and it is useful for users to enhance their experience and thus helpful in making end to end applications and software. Even though, it has some benefits there is a challenge in unsupervised learning technique since it automatically does everything without human intervention. Some of the challenges of unsupervised techniques are,

- Complexity in computation due to large volume of data.
- It may give inaccurate results too. Inaccurate results should be solved because it is important parameter to consider when it is applied in real world problems.
- Even though it automatically predicts, or groups data based on some criteria, there is a need of human intervention in validating the result.
- It may take longer time to train the data points.
- Collecting data in specific domain like IOT, Cyber security and agriculture, network traffic is not straight forward.
- In depth investigation is required while collecting real world data.

But when all these challenges are solved successfully, then unsupervised give more advantages in business environment like gaining profit to marketing entities, enhance user experience by providing more accurate results for example, recommender system, which recommends similar things what user likes the most, makes companies to understand about their customers to track their behavior, activities etc. The success of machine learning based solution in specific domain or application needs good quality of data and choosing appropriate algorithms. Hence, effectively handling features, processing, and good level of maintaining the dataset leads to best performance by the machine learning model with high accuracy and leads to build effective and intelligent application.

## VII. Conclusion

Unsupervised learning is a powerful tool which can be used for large databases. There are various kinds of applications developed using unsupervised learning technique. Variety of algorithms are there to solve problems and give more accurate results. Both advantages and disadvantages are there in unsupervised learning technique, when successfully solved these challenges faced by unsupervised learning, it gives more profit to companies, strengthen relationship between companies and their customers, improve user performance and there is a lot of advantages when using unsupervised machine learning technique. It automatically solves problems by grouping the data points based on similarity without any human intervention. Less human intervention, more automatic process with fairly good results is unsupervised machine learning. Hence, this paper describes the complete survey on unsupervised machine learning with its applications, algorithms and its challenges faced in real world scenarios.

## References


[1] A. Toshniwal, K. Mahesh and R. Jayashree, "Overview of Anomaly Detection techniques in Machine Learning," 4th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), pp.808-815, 2020, doi: 10.1109/I-SMAC49090.2020.9243329.

[2] Li, N., Shepperd, M., & Guo, Y., "A systematic review of unsupervised learning techniques for software defect prediction," Information and Software Technology, Vol.122, pp.106287, 2020.

[3] Rodrigues, J., Belo, D., & Gamboa, H., "Noise detection on ECG based on agglomerative clustering of morphological features," Computers in biology and medicine, Vol.87, pp.322-334, 2017.

[4] Shakeel, P. M., Baskar, S., Dhulipala, V. S., & Jaber, M. M., "Cloud based framework for diagnosis of diabetes mellitus using K-means clustering," Health information science and systems, Vol.6, No.1, pp.1-7, 2018.

[5] [5] M. Sujithra, P. Velvadivu, J. Rathika, R. Priyadharshini and P. Preethi, "A Study on Psychological Stress of Working Women In Educational Institution Using Machine Learning," 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2022, pp. 1-7, doi: 10.1109/ICCCNT54827.2022.9984460.

[6] Liang, Z., Wang, C., Duan, Z., Liu, H., Liu, X., & Ullah Jan Khan, K., "A Hybrid Model Consisting of Supervised and Unsupervised Learning for Landslide Susceptibility Mapping" Remote Sensing, Vol.13, No.8, pp.1464, 2021.

[7] Kuang, H., Qiu, Y., Li, R., & Liu, X., "A hierarchical K-means algorithm for controller placement in SDN- based WAN

architecture," 2018 10th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), IEEE, pp.263-267, February 2018.
[8] Addagarla, S. K., & Amalanathan, A., "Probabilistic Unsupervised Machine Learning Approach for a Similar Image Recommender System for E-Commerce," Symmetry, Vol.12, No.11, pp.1783, 2020.
[9] S. Siddharth, R. Darsini and M. Sujithra, "Sentiment analysis on twitter data using machine learning algorithms in python", Int. J. Eng. Res. Comput. Sci. Eng., vol. 5, no. 2, pp. 285-290, 2018.
[10] Acar, E., & Yener, B., "Unsupervised multiway data analysis: A literature survey," IEEE transactions on knowledge and data engineering, Vol.21, No.1, pp.6-20, 2008.