IEC
Institute of Electronics and Computer

Journal of Artificial Intelligence and Systems, 2024, 6, 85-95
https://iecscience.org/journals/AIS
ISSN Online: 2642-2859

A Review on Analysis of K-Means Clustering Machine Learning Algorithm based on Unsupervised Learning

Manish Suyal[a,*], Sanjay Sharma[b]

SGRR University, Department of CA & IT, Dehradun, India
[a]suyal.manish922@gmail.com;[b]dean.cait@sgrru.ac.in
*Corresponding author

How to cite this paper: Manish Suyal, Sanjay Sharma (2024). A Review on Analysis of K-Means Clustering Machine Learning Algorithm based on Unsupervised Learning. Journal of Artificial Intelligence and Systems, 6, 85–95. https://doi.org/10.33969/AIS.2024060106.

Received: December 29, 2023
Accepted: March 29, 2024
Published: April 15, 2024

Abstract

The process of machine learning is understood within Artificial Intelligence. Machine learning process gives the tools the ability to learn from their experiences and improve themselves without any coding. In machine learning, we program a computer or machine in such a way that the user wants the work done by the machine. It can give such work and in this process the computer does its work on the basis of the data already with it and gives its performance. The objective of writing the paper is how K- Means clustering algorithm is applied on the model dataset based on unsupervised learning. We used to pass feature data and label data to machine learning model in the supervised learning. But the method of unsupervised learning algorithms is different. In this we do not give the feature data and target data to the model. The dataset model uses only the input data for processing and the output data has no meaning in the model. Accordingly on the basis of the similarities found in the data and model predict the desired output. K-means is clustering algorithm based on unsupervised learning in which data and objects are separated into different clusters in such a way that objects that have similar properties are put in one cluster and objects that have different properties are put in separate cluster.

Keywords

Predictive Machine Learning Model, Unsupervised Strategy Learning Technique, K-means Algorithm, Cluster Forming Algorithm, Artificial Intelligence, Cluster, Input Data

1. Introduction

In today's environment machine learning is a futuristic technology which is in high demand these days. If you do online job search, then you will also get to see the most jobs in these areas, artificial intelligence, machine learning and data science and the most money will also be available in these jobs. In the coming times, the use of these technologies will increase even more rapidly because the upcoming device is of the smart machine. That's why all the big companies of the world are investing in technology like AI, AR, VR and machine learning strategy and competing to keep themselves ahead of others. Overall system learning is one of the top latest technologies at the moment and as many big companies from Google to face book, everything is using machine learning. In computer science, system learning has divided in two parts, supervised and unsupervised technique. The unsupervised strategy is the exact opposite of the supervised technique. In this the output data is not given to the model only input data is given. That is, by analyzing the input data that the model receives, it makes its own predictions. In this paper K-means cluster strategy is easily explained in the paper, which is based on unsupervised learning technique. The Cluster technique is very prominent method in unsupervised learning. Clustering is also called cluster analysis. K-means is a system learning technique learning in which data and objects are separated into different clusters in such a way that objects that have similar properties are put in one cluster and objects that have different properties are put in separate cluster.

- K-Means algorithm on very large data which are related to each other that above can be implemented easily.
- The large datasets can be easily scaled with the K-means clustering system learning technique which works in an unsupervised manner.
- The K-means strategy gives a strong guarantee of the convergence.
- K-means strategy algorithm easily generalizes the clusters for the various shapes and structure, for example elliptical clusters.
- K-means algorithm is very easily adapts to new examples.

2. Related Work

If we want to work on the unsupervised learning in future, then we have to read the review papers of previous years which is based on unsupervised learning.

Khanum M, Mahboob T, [1] have noticed the paper provides a comprehensive survey of unsupervised learning and methodologies under the machine learning. These

cluster techniques are fed to learn highly complex non-linear models with large amount of data where many parameters are in unlabeled form. In computer Science, K-means is considered as the most sophisticated and prominent cluster method. The year of publication of K-means cluster technique was 1955.

The study [2] proposes Dey. A, In today's time, most of the discussion is being done on various machine learning algorithms, with the help of which we are able to fulfill various purpose like data mining, image processing and pre- dictive analytics etc. The main purpose of using machine learning algorithms is that the systems learn from the algorithm what to do with the data at a time. After that the system automatically does its job smoothly.

The Self-Organizing neural networks learns by identifying hidden patterns in unlabeled input data with the help of unsupervised learning algorithms. The unsupervised learning takes a lot of advantage of the lack of direction for the educational algorithm because this algorithm allows seeing patterns that were never considered before [3].

In the paper, we take into account all the basic elements required to perform the clustering process, such as distance, similarities measures and evaluation indicators. We have tried to experience K-means cluster technique in old and new way. In the Paper, there are two main algorithms K-means and K- mediods. The K-means technique tries to update the center of the cluster to complete its task which is represented from data point of the center. The repe- tition process continues until certain criteria for concurrence are met. The K- mediods is in a way better than K-means to agreement with the discrete data [4].

Bindra, Mishra [5] have analyzed, the paper attempts to explain how to address the expansion of the partial issue of cluster formation. The study stud- ies, reviews and analyzes some cluster strategy falling under various categories instance to fulfill its purpose and currently gives us a lot of profitable comparison efficiency. The K-means cluster technique is depend on distance and divides the data into prearranged clusters or groups. The Euclidean and cosine formulas can prove useful in finding distance. The K-means algorithm locates the means to show the exact center of the cluster and displays the re- sult in extreme values but the K-mediods strategy calculates the cluster center using the actual point. Primarily, the strategy tries to reduce the genreal dis- parity of items to their nearest item.

The author [6] is currently explaining the scenario; nowadays it is very important to have cluster technique for big data. The paper provides an example of how cluster strategy can prove to be a solution for big data and gives suggestions on the source of

cluster techniques based on the big data. The cluster techniques are one of the major techniques of system learning which is prom- inently focused for mining internal data and segmentation of data set into multiple subjects.

The paper describes, unsupervised learning has been studied to solve partial differential equation easily based on the numerical approach [7].

The study [8] describes, a comprehensive approach is presented on system learning strategy that can be easily implemented to enhance the intelligence and ability of a model. The major contribution of the study is the principal of various system learning techniques and various real world applications domains such as security systems, smart cities, healthcare, e-commerce and agriculture.

Govindasamy, Velmurugan [9], explain in the paper the nowadays K-means technique is being used to solve the problem of clustering. In the first phase the data is divided into groups by looking the similarities; if we make a mistake in choosing the number of clusters will mess up the complete development so that the results will be wrong.

Maheswari [10] analyzed the customer is a major asset to the business. In today's time companies do not invest much for customer relationship management. The online shopping is also increasing day by day. The people are showing more interest to visit popular websites and they spend very less contacting people to choose their products. The online shops are paying more money to analyze customer preference, needs and buying behavior through machine learning techniques.

Saroj, Kavita [11] has noticed in the sentences of the paper that, comprehensively review the various forms of K-means and updated K-mean cluster strategy. The cluster technique is used in lot of fields such as statistical, patterns identification and machine learning. The cluster analysis is a data mining tool that is increasingly used nowadays in large and multidimensional databases. The clustering is one of the most important techniques of machine learning in which data is divided into groups of similar objects and different objects into another group. The clustering is a suitable example of unsupervised learning.

In other words, it can be said that no training data is provided to the model and the model automatically does its analysis and start its mechanism automatically [12].

In today's time, the purpose of text clustering is to analyze and fine-tune the data of text on a large scale which can be grouped into several categories in which similar text is kept in one group and different text in another group. The paper tries to explain, about hierarchical cluster has been stated. In hierarchical cluster technique, we can

create different clusters of text by measure the cosine similarity, dice coefficient and jaccard similarity coefficient [13].

The study [14], help us to understand the process is start from the root node of the tree structure and progresses by applying conditions at each non-leaf node resulting homogenous subset.

The machine learning is primarily a part of artificial intelligence which is a major component of digitization solution, which has attracted major attention in the digital sectors. The author has paid much attention in the paper to express a brief problem of different system learning technique that has been used most frequently. The intent of the author is to help in making a proper decision towards selecting the suitable system learning technique to meet all the requirements of an application [15].

3. Clustering Methodology: Unsupervised Machine Learning

The cluster technique is most followed in unsupervised learning; clustering is the process in which similar things are grouped together. The goal of clustering is in unsupervised learning, finding the similarities in a data set. In unsupervised learn- ing, we understand the clustering algorithm through a simple example. Suppose you have a fruit basket containing different types of fruits. Now your task is to keep one type of fruit separately, so here we use clustering algorithm for this. For clustering algorithms all these fruits are same, it doesn't know which fruit is of what kind then it finds similarity between these fruits. So it just takes their first attribute like color and first separates these fruits according to their colors. Then it separates them according to other properties or attributes like their size and shape, so by finding similarities between these fruits. It makes a different group of these fruits which is called as cluster and this process is called clustering. We can understand the supervised learning as example. When the system or model is trained with supervised technique, a system is given certain datasets (for example, we can consider the color of apple is red, weight is taken as 20 gram, the shape is round and the height is taken as 5 cm). Using these dataset, the proposed system model predicts the outputs. The two types of data fields are given to the system learning model in these fields, one field will be of the feature data and other field will be of label data. Using these data, algorithm is used to train the ML model, what kind of output to predict. Suppose we have to train a system model, which recognizes mango. For this we have to use supervised learning algorithm. For this you have to specify the system model, what does the mango look like, to do this we have to give the feature data to the system model (such as mango color is yellow, shape is round and taste is sweet).

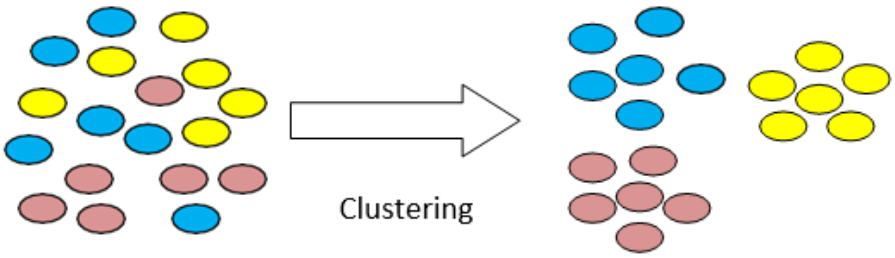

Figure 1. Forming different types of clusters using clustering algorithms

The most popular algorithm of unsupervised learning is known as K-means clustering. The machine is taught to learn by the K-means, in which by working on the data and object, after that separated into different clusters, in such a way that objects with identical properties are put in one cluster and objects with different properties are put in another cluster. The objects in each group are different from the objects of the other cluster.

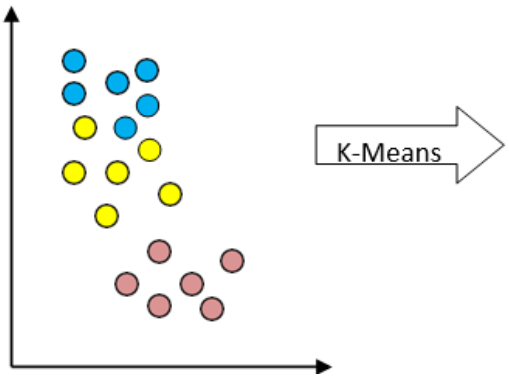

Figure 2. Before applying K-means clustering

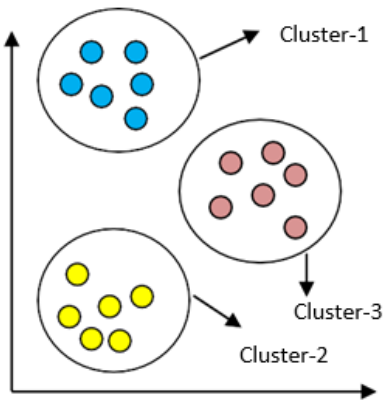

Figure 3. After applying K-means clustering

To understand the working of K-means cluster algorithm, we have to understand step by step.

Step-1: In the first step, choose the number K; take judgment about the numbers of the clusters.

Step-2: Any K data points can be taken randomly.

Step-3: Each data points must be assigned to a nearest barycenter which will be from the preconcert K- clusters.

Step-4: To figure out the change value and keep a novel barycenter for each cluster.

Step-5: Let's iterate the third step, which mean reallocate each data points to the novel nearest barycenter of every cluster.

Step-6: If any reallocating repeats again and again then go to step-4 otherwise go to finish.

Step-7: In this way, the K-means clustering system will be ready

The formula for finding the distance measure shows the similarities between two elements and tries to influences the shape of various clusters things.

$$D = \sqrt{(x_1 - c_1)^2 + (x_2 - c_2)^2}$$

Where x_1 and x_2 =data point and c_1 and c_2 = closest point to centroid.

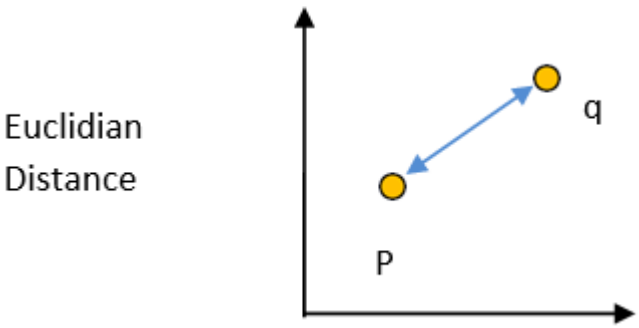

Figure 4. Euclidian Distance formula between two points

The k-means cluster technique works by dividing objects into different types of clusters mentioned by the number 'K.', so if we can say K = 2, that the object is divided into cluster c_1 and c_2.

4. Result and Evaluation

The K-means algorithm can be understood with a simple example as follows. Apply K-mean clustering for the following datasets. Apply K-mean clustering for the following dataset.

$$K = [2, 3, 4, 10, 11, 12, 20, 25, 30] \quad (1)$$

First of all, we have to take a mean value is often give to you. We have to make clusters on the basis of mean value from the dataset. The value of K may have been given to you. Suppose K=2, this means you have to create two clusters but let's see how to make it. Now we take two mean values from the data set according to us.

$$M1 = 4, M2 = 12 \quad (2)$$

Now we have to create two clusters K1 and K2, only the value that is closest to 4 will come in the K1 cluster and only the value that is closest to 12 will come in the K2 Clusters.

$$K1 = [2,3,4], K2 = [10,11,12,20,25,30] \quad (3)$$

Now we have to get M1 and M2 again, earlier we had taken out random. M1 will be the average of the K1 cluster.

$$M1 = (2+3+4)/3 = 9/3 = 3 \quad (4)$$

M2 will be the average of the K2 cluster.

$$M2 = (10+11+12+20+25+30)/6 = 108/6 \quad (5)$$

We are not getting same mean value over yet so we have to make cluster again.

$$K1 = [2,\ 3,\ 4,\ 10], K2 = [11,\ 12,\ 20,\ 25,\ 30] \quad (6)$$

M1 will be the average of the K1 cluster.

$$M1 = (2+3+3+10)/4 = 4.75 = 5 \quad (7)$$

M2 will be the average of the K2 cluster.

$$M2 = (11+12+20+25+30)/5 = 19.6 = 20 \quad (8)$$

$$K1 = [2,\ 3,\ 4,\ 10,\ 11,\ 12] \ and \ K2 = [20,\ 25,\ 30] \quad (9)$$

M1 will be the average of the K1 cluster.

$$M1 = (2+3+4+10+11+12)/6 = 7 \quad (10)$$

M2 will be the average of the K1 cluster.

$$M2 = (20+25+30)/3 = 25 \quad (11)$$

$$K1 = [2,\ 3,\ 4,\ 10,\ 11,\ 12] \ and \ K2 = [20,\ 25,\ 30] \quad (12)$$

M1 will be the average of the K1 cluster.

$$M1 = (2+3+4+10+11+12)/6 = 7 \quad (13)$$

M2 will be the average of the K1 cluster.

$$M2 = (20+25+30)/3 = 25 \quad (14)$$

When we start getting the same mean, then we have to stop the process.

$$K1 = \{2,\ 3,\ 4,\ 10,\ 11,\ 12\}\ and\ K2 = \{20,\ 25,\ 30\} \quad (15)$$

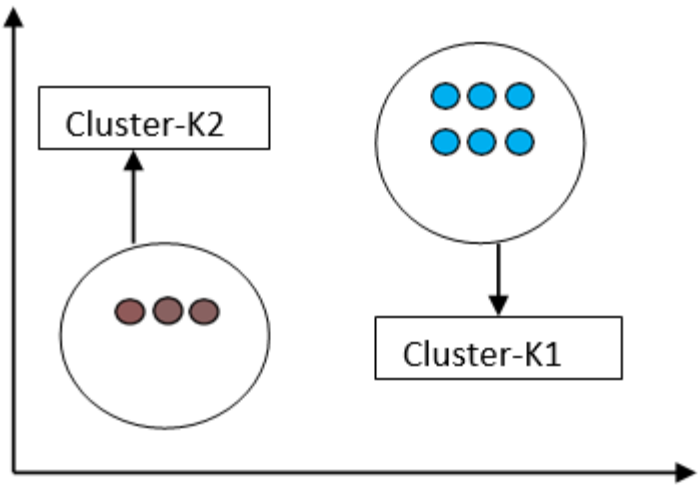

Figure 5. K1 consist 6 (2, 3, 4, 10, 11, 12) data points and K2 consist 3 (20, 25, 30) data points

5. Conclusion and Suggestion for Future Work

The message in the paper has been noticed that, we have presented the unsuper- vised learning and K-means clustering algorithm in a very easy way. The trouble that comes in understanding the unsupervised system strategy and K-means tech- nique has been explained in very simple words and example. The unsupervised technique comes under the part of machine learning that deals with unlabeled data. The unsupervised learning can be understood as the opposite of the supervised learning is neither an example is given to the model nor an output is given; only input data is given and based on that the model makes predictions. After working the paper, we understand that data has to be grouped and its clusters have to be created in unsupervised learning. In such a situation, working on a data set be- comes complex because you do not even have labeled data in it. Whereas in su- pervised learning there is a supervisor in which labeled data is given and processed and there is no need to form groups in unsupervised learning. In this way, it can be said that unsupervised machine learning is more complex that supervisory machine learning. In unsupervised learning, K-means cluster technique is used to operate on the data set. K-means clustering

algorithm is useful in collecting data sets. K-Means algorithm on very large data which are related to each other that above can be implemented easily. The large datasets can be easily scaled with the K-means cluster strategy which is focused on unsupervised technique. The K- means technique gives a strong guarantee of the convergence. K-means technique easily simplifies the clusters for the various shapes, colors and sizes, take elliptical clusters for example. K-means strategy is very easily adapts to new examples. In the paper we have taken a small data set for applying the K-means learning algorithm. We find the similarities in all the data elements of the dataset and form clusters of them using K-means algorithm. By the K-means algorithm we know how to find similarities between each other elements in the elements of the dataset and how to from their clusters. After working in the paper, we know an important thing in the coming time all the machines made with unsupervised learning will be one of the very smart and trending technologies. The K-means cluster strategy very easily finds the similarities among data elements of the dataset and very easily to form clusters on the basis of the similarities. The clusters can be easily created by taking the distance of each element from the other element in the dataset us- ing Euclidian distance formula. While writing the paper, we know that unsupervised learning is one such study. The Paper concludes that unsupervised learning will be the future of technology in the times to come. The unsupervised learning is being used nowadays in product recommendation, you want to buy a product on Amazon, but you did not buy it and the next day you were watching YouTube and an advertisement for the same product appeared on it. That you had gone to buy on Amazon, but you did not buy it. Then you visited a face book and there was also an advertisement show of the same product your online behavior is track and machine learning algorithms are applied on it and then you are shown the same advertisement in which you are interested. Nowadays unsupervised learning is being used for Google translate; you can translate many languages of the world among themselves. In future we can work on semi supervised machine learning algorithm. This algorithm falls between both supervised learning uses label data and unsupervised learing uses unlabeled data.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Khanam, M., Mahboob, T., 2015, “A Survey on Unsupervised Machine Learning Algorithms for Automation, Classification and Maintenance,” International Journal of Computer Application, 119(13), pp. 0975-8887.

[2] Dey, A., 2015, "Machine Learning Algorithms: A Review," International Journal of Computer Science and Information Technologies, 7(3), pp. 1174-1179.
[3] Sathya, R., Abraham, A., 2013, "Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification," International Journal of Advanced Research in Artificial Intelligence, 2(2), pp. 34-38.
[4] Xu, D., Tian, Y., 2015, "A Comprehensive Survey of Clustering Algorithms," Ann. Data. Sci., 2(2), pp. 165-193.
[5] Bindra, K., Mishra, A., 2017, "A Detailed Study of Clustering Algorithms," International Conference on Reliability, Infocom Technologies and Optimization, 5(17), pp. 371-376.
[6] Wu, C., 2019, "Research on Clustering Algorithm Based on Big Data Background," International Journal of Physics, 12(37), pp. 1-6.
[7] Cai, Z., Chen, J., 2019, "Deep Least-Squares Methods: An Unsupervised Learning-Based Numerical Method For Solving Elliptic Pdes," Journal of Computational Physics, 10(16), pp. 1-20.
[8] Sarker, I., 2021, "Machine Learning: Algorithms, Real World Applications and Research Directions," SN Computer Science, 2(160), pp. 2-21.
[9] Govindasamy, K., Velmurugan, T., 2017, "A Study on Classification and Clustering Data Mining Algorithms based on Students Academic Performance Prediction"