DYNAMIC AUTO K-MEANS ALGORITHM

ABSTRACT:

Today most of the software projects/applications involved AI, ML and DS. In the IT industry and academic Institutions these three subject streams are becoming popular. On the other side due to automation and digitalization the volume of data is increasing. Generally we find a good number of applications on Clustering and its various implementation methodologies and so on Research area. In specific, we have many a number of K-Means Clustering methodology approaches. In most research papers K-means is treated as UN-supervisory, but over a period time the results of Un-supervisory clustering may provide some input for latter repetition of clustering the same data. Hence I thought a "Dynamic k-means Algorithm" is more useful as data groups may change at short Intervals.

The Dynamism in clustering can be:

- **Number of centroids(k)**
- **Actual value of centroids**
- **The cluster group of values**

From these analyses we are attempting an Algorithm that should be able to dynamically change above dimensions (Parameters) .i.e. the Algorithm initially accepts' k' and centroids then cluster the data, store these parameters later, It dynamically predict probable 'k' and centorids and keep grouping.

Thus, making this ("Dynamic auto k-means Algorithm") Un-supervisory initially and supervisory later by taking 'k' from or predicting 'k' from previous values.

LITERATURE SURVEY (Review)

A literature review is an overview of the previously published works on a specific topic. The term can refer to a full scholarly paper or a section of a scholarly work of an article.

A literature review consists of an overview, a summary, and an evaluation ("critique") of the current state of knowledge about a specific area of research. lt may also include a discussion of methodological issues and suggestions for future research.

A literature review helps understanding the subject content in different dimensions and provides inputs for new ideas. Following are some of the papers We have studied, clarified our doubts and formed the subject title of our paper.

Youguo Li, Haiyan Wu[1] in his paper titled **"A Clustering Method Based on K-Means Algorithm"** combine the largest minimum distance algorithm and the traditional K-Means algorithm to propose an improved K-Means clustering algorithm. This improved K-Means algorithm effectively solved two disadvantages of the traditional algorithm, the first one is greater dependence to choice the initial focal point, and another one is easy to be trapped in local minimum.

Jun Wu, Li Shi, Wen-Pin Lin , Sang-Bing Tsai , Yuanyuan Li, Liping Yang, and Guangshu [2] in their paper "An Empirical Study on Customer Segmentation by Purchase Behaviours Using a RFM Model and K-Means Algorithm" They dealt with a real-world problem in an enterprise. A RFM (recency, frequency, and monetary) model and K-means clustering algorithm are utilized to conduct customer segmentation and value analysis by using online sales data. Customers are classified into four groups based on their purchase behaviors. On this basis, different CRM (customer relationship management) strategies are brought forward to gain a high level of customer satisfaction. The effectiveness of this approach is supported by improvement results of

some key performance indices such as the growth of active customers, total purchase volume, and the total consumption amount.

Jaswanth Reddy Vulchi[3] has focused on effective decisions and decision making. With the help of machine learning technique one can sort out the data and can find the target group by applying several algorithms to the dataset. Without this, It will be very difficult and no better techniques are available to find the group of people with similar character and interests in a large dataset. Here, The customer segmentation using K-Means clustering helps to group the data with same attributes which exactly helps to business the best. We are going to use elbow method to find the number of clusters and at last we visualize the data.

Prashant Sharma[4] Emphasized and gave importance to volume of data and its availability on internet Even though the nature of individual data is straightforward, the sheer amount of data to be analyzed makes processing difficult for even computers. To manage such procedures, we need large data analysis tools. Data mining methods and techniques, in conjunction with machine learning, enable us to analyze large amounts of data in an intelligible manner. k means is a technique for data clustering that may be used for unsupervised machine learning. It is capable of classifying unlabeled data into a predetermined number of clusters based on similarities (k).

Kristina P. Sinaga and Miin-Shen Yang [5] Focused on various versions of k-means and how un-supervisory k-means can be tuned to supervisory k-means. There are various extensions of k-means to be proposed in the literature. Although it is an unsupervised learning to clustering in pattern recognition and machine learning, the k-means algorithm and its extensions are always influenced by initializations with a necessary number of clusters a priori. That is, the k-means algorithm is not exactly an unsupervised clustering method. In their paper, they construct an unsupervised learning schema for the k-means algorithm so that it is free of initializations without parameter selection and can also simultaneously find an optimal number of clusters. That is, we propose a novel unsupervised k-means (U-k means) clustering algorithm with automatically finding an optimal number of clusters without giving any initialization and parameter selection. The computational complexity of the proposed U-k-means

clustering algorithm is also analyzed. Comparisons between the proposed U-k-means and other existing methods are made. Experimental results and comparisons actually demonstrate these good aspects of the proposed U-k means clustering algorithm.

The Review of literature provided us good knowledge and various thoughts on the subject matter of k-means clustering algorithm and the result is our paper titled "DYNAMIC AUTO K-MEANS ALGORITHM"

INTRODUCTION:

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters (groups). K-means clustering is a method of vector quantization that aims to partition "n" observations into "k" clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. To achieve this objective, K-means looks for a fixed number (k) of clusters in a dataset." A cluster refers to a collection of data points aggregated together because of certain similarities.

Clustering is a key technique in pattern recognition, data mining, and knowledge discovery. The aim is to uncover the (hidden) structure underlying a given collection of objects. If, the data is fixed or relatively permanent then traditional k-Means clustering methods can used.

But today, as large volumes of data are created almost every day the group dynamics of data changes. This change in data in turn leads to the change in number of groups and the values or elements in the group. Many a number of

times we may need to compare and analyze cluster's before and after changes in data. This leads to requirement of Dynamic clustering techniques.

This type of clustering embraces different scenarios: dynamic features, dynamic data objects and dynamic clusters. Dynamic clustering is also relevant to many other clustering situations involving very large data, data streams, incomplete data, noisy data, unbalanced data, and structured data. This makes the grouping more challenging and quite interesting. Moreover, because clustering can also be just one step in a multi-step complex system, its importance grows. In applications like control, artificial vision, surveillance, etc., clustering quality is vital for the well-functioning of the whole system.

Dynamic clustering as a form of unsupervised online/incremental machine learning that considers two concepts:

(1) Learning methods that are incremental to devise the clustering model and
(2) Self-adaptation of learning model (parameters and structure).

The incremental learning methods solve the problem of time-intensive re-training and memory constraints. Dynamic (online/incremental) clustering has been attracting the attention of the many research communities (like data mining, pattern recognition, machine learning, etc.)

This encouraged us to attempt the implementation of dynamic clustering using machine learning techniques.

METHODOLOGY:

A research methodology gives research legitimacy and provides scientifically sound findings. It also provides a detailed plan that helps to keep track on work, making the process smooth, effective and manageable.

Research methodology is a way to systematically solve the research problem. It may be understood as a science of studying how research is done scientifically. In it we study the various steps that are generally adopted by a researcher in studying his research problem along with the logic behind them.

In our paper we are assuming and recommend following the "simulation" methodology for our problem of "Dynamic Auto K-Means Algorithm". As we already stated K-Means Clustering deals with grouping of data as per some predefined parameters or arguments. We choose to derive these parameters from the given data rather than human providing random values as they may be biased, for example, as a part of clustering we have to choose "K" and their initial centroids. Our algorithm dynamically derives these parameters and recommended to validate this value of parameters and also whether they are appropriate to give data or not. As a part of choosing the best parameters we also recommend comparing these value's with previously values and previous groups, the change in the volume of data. We would like to use machine learning techniques for better supervisory algorithms.

Our algorithm initially calculates a threshold value as centroids of k-means and based on this value the number of clusters are formed. We recommend that If, the solution of centroids is wrong then clustering result is volatile and the number of iterations will be increased hence optimization of iterations is a subject to concern.

Finally we would like to test the derived clusters or group whether the they satisfy the principle of " If, the Euclidian distance between two points is less than or

equal to the threshold value then these two data points must be in the same group".

This methodology needs implementation with large data sets and with different types of data such as text, image. Also we recommend rigorous testing and fine tuning of various methods suggested.

CONCLUSION

References

- **Youguo Li, Haiyan Wu "A Clustering Method Based on K-Means Algorithm**

  LINK: https://www.researchgate.net/publication/271616608

- **Jun Wu,1,2 Li Shi,1 Wen-Pin Lin , 3 Sang-Bing Tsai , 4 Yuanyuan Li,2 Liping Yang,2 and Guangshu Xu5 , "An Empirical Study on Customer Segmentation by Purchase Behaviors Using a RFM Model and K-Means Algorithm"**

LINK: https://www.hindawi.com/journals/mpe/2020/8884227/

- **Jaswanth Reddy Vulchi "Customer Segmentation Using K- Means Clustering Algorithm**

  LINK: https://www.researchgate.net/publication/355587534

- **Prashant Sharma "Understanding K-means Clustering in Machine Learning(With Examples)"**

  LINK : https://www.analyticsvidhya.com/blog/2021/11/

- **Kristina P. Sinaga and Miin-Shen Yang "Unsupervised K-Means Clustering Algorithm"**

  LINK : https://www.researchgate.net/publication/340813602