Student Segmentation using Clustering Algorithms

A Synopsis Submitted
in Partial Fulfillment of the Requirements
for the Course of
Minor Project - I
In
Third year – Fifth Semester of
Bachelor of Technology
specialization
In

Artificial Intelligence and Machine Learning

Under

Sujoy Chatterjee

By

500075940	**R177219148**	**Rohan Nyati**
500076347	**R177219143**	**Rajneesh**
500075224	**R177219170**	**Shantanu Jaswal**

DEPARTMENT OF INFORMATICS
SCHOOL OF COMPUTER SCIENCE
UNIVERSITY OF PETROLEUM AND ENERGY STUDIES,
BIDHOLI, DEHRADUN, UTTRAKHAND, INDIA
Sep,2021

Synopsis

1.Introduction

There has been a lot of deal about the coronavirus that started in 2019 and due to that the students have not been able to join the college for studies and colleges have been running in online mode for more than a year .

The current trend of research on machine learning has been great due to many students studying it in different streams .Machine learning (ML) is the learning process developed for machines using various mathematical computational algorithms that can improve automatically through experience or by the use of data .

As the colleges have been planning to open again after one harsh year we are planning to create a project that helps colleges to get an idea about how many students are fully vaccinated and will be joining the college after this pandemic. We will be creating some random data on which we will be implementing clustering algorithms that will help us get predictions about the current mental state of the students in terms of academic knowledge and how much they were able to learn and understand during this pandemic on the basis of which colleges will be able to predict which student needs how much help and how many extra classes will be required for that student and will be able to group them on the basis of that , thus help them get back on track .

2.Motivation

The Motivation behind opting for this is that for the past one year students have been studying in online mode and because of which their knowledge about the topics have reduced by a great deal . So our motivation here is to help the college in finding such students and help them get back on their feet and become industrial ready .

As a ML enthusiast, we wanted to explore the core of these algorithms using a structured programming approach to gain the pure insights and working of this algorithm.

3.Related work

DBSCAN is a density based algorithm [1] used for clustering that requires only one input parameter and supports the user in determining an appropriate value for it. It discovers clusters of arbitrary shape. Finally, DBSCAN is efficient even for large spatial databases.
The Eps-neighborhood of a point p,
denoted by:-

$N_{Eps}(P)$, is defined $N_{Eps}(P) = \{q \in D \mid dist(p,q)$ _< Eps}.

A naive approach could require for each point in a cluster that there are at least a minimum number (MinPts) of points in an Eps-neighborhood of that point.

We require that for every point p in a cluster C there is a point q in C so that p is inside of the Eps-neighborhood of q and NEps(q) contains at least MinPts points.

A point p is directly density-reachable from a point q wrt. Eps, MinPts if:-
1) p ∈ $N_{Eps}(P)$ and
2)| $N_{Eps}(P)$ | > MinPts (core point condition)

To find a cluster, DBSCAN starts with an arbitrary point p and retrieves all points density-reachable from p wrt. Eps and MinPts. If p is a core point, this procedure yields a cluster wrt. Eps and MinPts. If p is a border point, no points are density-reachable from p and DBSCAN visits the next point of the database

INCREMENTAL DBSCAN CLUSTERING [2] insertion of some new data items into the already existing clusters .The new data which are not inserted into any clusters, they are treated as noise or outliers. Sometimes outliers which fulfil the Minpts & eps criteria , combinly can form clusters using DBSCAN.

$$\% \ \delta \text{ Change in database} = \frac{(\text{ new data - old data}) \text{ X } 100}{old\ data}$$

Incremental DBSCAN clustering algorithm [3] is used to handle dynamic databases. It has the ability to change the radius threshold value dynamically.The actual DBSCAN approach is not suitable for a large multidimensional database which is frequently updated. In that case, the incremental clustering approach is much better.

The K-Means algorithm based on dividing [4] is a kind of cluster algorithm, and it is proposed by J.B.MacQueen. This algorithm which is unsupervised is usually used in data mining and pattern recognition. Aiming at minimizing cluster performance index, square-error and error criterion are foundations of this algorithm. The K-Means algorithm based on dividing is a kind of cluster algorithm, and has advantages of briefness, efficiency and certainty.

$$D = ||X\text{-}Z|| = [\sum_{i=1}^{n} (x_i\text{-}z_i)^2]^{0.5}$$

D is the distance of X and Z in n-dimensional space, where X and Z are two samples.

The sample pattern congregation is $\{x\} = \{X_1, X_2, \ldots\ldots X_n\}$, and we classify it to C classes, they are $S_1, S_2, S_3, S_4 \ldots\ldots S_c$. M_j and S_j are mean vectors. So:

$$M_j = \frac{1}{Nj} \Sigma_{X\in Sj} X, \ N_j = |S_j|$$

where N_j and S_j are number of samples, so we define cluster criterion function as:-

$$J = \sum_{j=1}^{c} \sum_{X\in Sj} ||X\text{-}M_j||^2$$

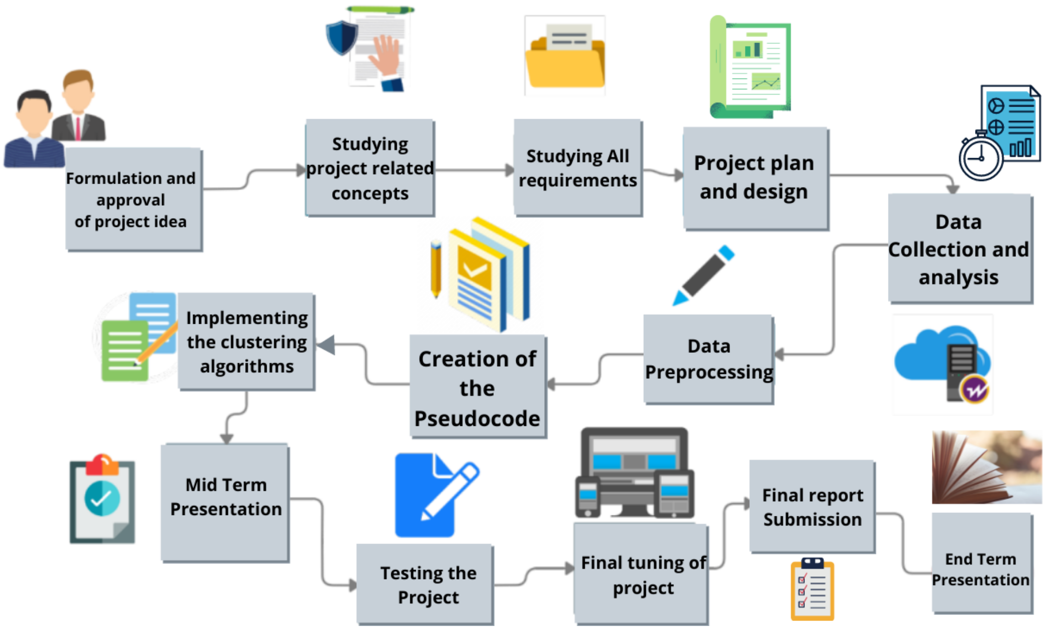
The incremental K-means algorithm[5] presented in this paper is similar to the block sequential algorithm with the exception that each block is accessed only one time.

Each block is going through a set of l epochs of K-means where the final centers of block ïare used as the initial centers for block i+1.

4. Methodology

Predictive analytics is the use of data, statistical algorithms and machine learning techniques to identify the likelihood of future outcomes based on historical data. It involves certain manipulations on data from existing data sets with the goal of identifying some new trends and patterns. These trends and patterns are then used to predict future outcomes and trends.
The steps given below are the one we go back and forth to achieve the best prediction.
Steps -
1. Problem understanding and definition
2. Data collection and preparation
3. Dataset understanding using clustering algorithm
4. Data analysis
5. Data validation
6. Deployment

5.Plan of work


Formulation and approval of project idea
Studying project related concepts
Studying All requirements
Project plan and design
Data Collection and analysis
Data Preprocessing
Creation of the Pseudocode
Implementing the clustering algorithms
Mid Term Presentation
Testing the Project
Final tuning of project
Final report Submission
End Term Presentation

References

[1] A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise Martin Ester, Hans-Peter Kriegel, Jiirg Sander, Xiaowei Xu Institute for Computer Science, University of Munich Oettingenstr. 67, D-80538 Miinchen, German https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf

[2] International Journal of Enterprise Computing and Business Systems Analysis and Study of Incremental DBSCAN Clustering Algorithm SANJAY CHAKRABORTY Prof. N.K.NAGWANI National Institute of Technology National Institute of Technology (NIT) Raipur, CG, India. https://arxiv.org/ftp/arxiv/papers/1406/1406.4754.pdf

[3] A Technical Survey on DBSCAN Clustering Algorithm Nidhi Suthar1 , Prof. Indr jeet Rajput2 , Prof. Vinit Kumar Gupta 3 1 Department of Computer Engineering , Hashmukh Goswami college of Engineering, Vahelal, Gujarat. https://www.ijser.org/researchpaper/A-Technical-Survey-on-DBSCAN-Clustering-Algorithm.pdf

[4]A Clustering Method Based on K-Means Algorithm Youguo Li, Haiyan Wu Department of Computer Science Xinyang Agriculture College Xinyang, Henan 464000, China https://www.researchgate.net/publication/271616608_A_Clustering_Method_Based_on_K-Means_Algorithm/link/57da70fc08aeea1959316130/download

[5]Dynamic Incremental K-means Clustering Bryant Aaron, Dan E. Tamir Department of Computer Science, Texas State University, San Marcos, Texas, USA, Naphtali D. Rishe, and Abraham Kandel School of Computing and Information Sciences Florida International UniversityMiami,Florida,USA http://cake.fiu.edu/Publications/Aaron+al-14-DK.Dynamic_Incremental_K-means_Clustering_IEEE-downloaded.pdf