ANALYSING STUDENT PERFORMANCE BY COMBINING DBSCAN AND GRID BASED CLUSTERING METHODS

[1]Vijay Rai, [2]Pooja Patre,
[1] Research Scholar, [2]Assistant Professor,
[1]Computer Science and Engineering, Vishwavidyalaya Engineering College, Lakhanpur, India
[2] Computer Science and Engineering, Vishwavidyalaya Engineering College, Lakhanpur, India

***Abstract*—** Clustering is the way toward making a gathering of conceptual items into classes of comparable items. The primary favourable position of bunching over arrangement is that, it is versatile to changes and helps single out valuable highlights that recognize diverse gatherings. The real necessities of bunching calculations is Scalability, Ability to manage various types of traits, Discovery of groups with property shape, High dimensionality, Ability to manage uproarious information, Interpretability. The point of the present work is to analyse the performance of the student on the basis of collected sample data of semwise student results. For analysis generation the DBSCAN algorithm is combined with Grid based algorithm for producing accurate and fast output.

***Keywords*—**Information Mining**,**Clustering, Density based clustering,Grid based clustering,Grid boundary points.

I. INTRODUCTION

Information Mining is characterized as separating data from enormous arrangements of information. At the end of the day, we can state that information mining is the methodology of mining learning from information. The data or information separated so can be utilized for any of the accompanying applications.

- Market Analysis
- Fraud Detection
- Customer Retention
- Production Control
- Science Exploration

Information mining is exceptionally valuable in the accompanying spaces:

- Market Analysis and Management
- Corporate Analysis and Risk Management
- Fraud Detection

Aside from these, information mining can likewise be utilized in the zones of creation control, client maintenance, science investigation, games, crystal gazing, and Internet Web Surf-Aid.

One of the information mining procedures is clustering. Clustering is isolating the informational collection into gatherings to such an extent that information focuses with comparative properties are assembled together. There are different calculations that can perform clustering. These calculations are comprehensively grouped into the accompanying classes:

- Partitioning grouping
- Hierarchical grouping
- Density based grouping
- Grid based grouping

Cluster investigation or cluster analysis is a procedure that is utilized arrange objects into gatherings with the end goal that the objects having a place with one gathering are substantially more like each other and not quite the same as other question gatherings. It has wide applications, including business sector or client division, design acknowledgment, natural examinations, spatial information investigation, Web archive grouping, and numerous others. Cluster investigation can be utilized as a remain solitary information mining instrument to pick up understanding into the information appropriation or can fill in as a pre-handling venture for other information mining applications working on the distinguished groups. The nature of clustering can be evaluated in view of a difference of objects, which can registered for different sorts of information, including interim scaled, paired, clear cut, ordinal, and proportion scaled factors, or mixes of these variable kinds.

Thickness based calculations find the group as per the areas which develop with high thickness. It is the one-filter calculations. The main approach called the thickness based availability clustering pins thickness to a preparation information point. It can locate the discretionary melded bunches and handle commotion.

A.DBSCAN Clustering

Density Based Spatial Clustering of Applications with Noise It is of partitioned compose clustering where more thick areas are considered as group and low density area are called noise. Steps of calculation of DBSCAN are as per the following:

- Arbitrary select a point r.
- Retrieve all focuses thickness reachable from r w.r.t Eps and MinPts.
- If r is a center point, bunch is shaped.
- If r is an outskirt point, no focuses are thickness reachable from r and DBSCAN visits the following purpose of the database.
- Continue the procedure until the point that the greater parts of the focuses have been handled.

Major drawback of density based approach is if it fails in case of neck type of dataset and it does not work well in case of high dimensionality data.

The real hindrance is it relies upon just the quantity of cells in each measurement in the quantized space.

B. Grid Based Clustering

This kind of clustering is worried about the esteem space that encompasses the information focuses not with the information focuses. This calculation utilizes the multi goals network information structure and utilize thick matrices to shape bunches. Grid Density based calculations require the clients to determine a density size estimate or the thickness limit, the issue here emerge is that how to pick the size or thickness threshold. To beat this issue, a procedure of adaptive grid types are suggested that consequently decides the span of grids in view of the information appropriation and does not require the client to determine any parameter like framework estimate or the size estimate of grid [6].

C.CLIQUE

It is a grid based strategy that discovers thickness based clustering in subspaces. Faction performs grouping in two stages. In the initial step, CLIQUE parcels each measurement into non-covering rectangular units, along these lines dividing the whole space of information objects into cells. In the meantime it recognizes the thick cells in every one of the subspaces. The unit is thick when the division of aggregate information focuses surpasses the info show parameter. In the second step, CLIQUE utilizes these thick cells to frame groups, which can be subjective.

V. LITERATURE SURVEY

Bo.Wu [10] presents a grouping approach by quick find and find of thickness peaks and thickness based spatial clustering of utilizations with noise, thus numerous others are accounted for to be fit for finishing this task however restricted by its calculation time of shared distances between focuses or designs. Without the estimation of shared distances, this work shows an elective strategy to satisfy grouping of information with any shape and noise much speedier and more proficient.

Elankavi et.al [11] proposed Fast Clustering Algorithm is utilized for choosing the subset of highlights or features. A Fast grouping calculation renders proficiency and viability to find the subset of highlights. Quick grouping calculation work should be possible in two stages. The initial step is to moving out unessential highlights from the dataset, the immaterial highlights are evacuated by the highlights having the at threshold over the predefined limit. What's more, the second step is to wiping out the excess highlights from the dataset, the repetitive highlights is expelled by developing the Minimum Spanning Tree and separate the tree having the edge remove more than its neighbour to shape the different bunches, from the groups includes that are emphatically connected with the objective highlights are chosen to form the subset of highlights.

Wu et.al [12] proposes a novel grouping technique called Spatial Clustering with Multiple Density-Ordered Trees (SCMDOT). Roused by the possibility of the Density-Ordered Tree (DOT), first dataset is represented by the methods for building Multiple Density-Ordered Trees (MDOT). In the developing procedure, we force extra imperatives to control the development of every Density-Ordered Tree, guaranteeing that they all have high spatial comparability. Moreover, a progression of MDOT can be progressively created from locales of meager territories to the thick regions, where every Density-Ordered Tree, additionally regarded as a sub-tree, speaks to a group. In the consolidating procedure, the last groups are acquired by over and over combining a reasonable match of clusters until the point when they fulfill the normal clustering result..

Singh.at.el [13] the two most imperative process amid which information's are gathered and investigated are affirmation and arrangement. The positioning of the college relies upon scholarly execution and arrangement of the students. Aside from scholastic execution there are different components which help in understanding the general execution of the student. In this examination work, the information mining method is utilized to comprehend the execution of student and gathering the students under different classes as an student need to reliably enhance to contend in this day and age.

V.METHODOLOGY

This paper is focused on the implementation of a density and grid based data mining techniques for clusters of various shapes and sizes. Detailed descriptions come in the following parts:

- Normalization and Scaling: At first, the original data set is normalized to [0,1] in each dimension and then scaled into [1,N_grid] range grid, where N_grid is the size of grid in each dimension. This simplifies the calculations of nodes' local densities.
- Calculation of Node's Local Density: Instead of computing pattern's local density, the node's local density will be used. In this proposed method, nodes are only located in the grid with integer coordinates.

The soft decision using fuzzy type approximation is proposed to determine the node's local density.

- Sparse Matrix Operations: It is noted that most parts of the nodes in the standard grid have no densities because of blank margins between clusters, which are used to separate different clusters. That means there is no need to deal with these zero-density nodes that will surely cost time. Thus, instead of processing all the nodes, sparse matrix that keeps non-zero nodes will be used and processed in this proposed method to accelerate it.
- Finding Mountain Ridges: The outlook of nodes' local densities is like mountains with different heights, so the task of clustering is redefined as finding the mountain ridges. To fulfil it, mountain ridges are detected one by one starting with the peaks, which are the nodes with higher local densities. For the first mountain ridge, it starts with the grid node with the largest local density among all nodes. Then it labels the neighbour nodes, which have distances of 1, into this mountain if the neighbour nodes are not the edges. The mountain edges are the nodes with densities smaller than Edg% of this mountain peak density. From the merged node(s), keep merging its/their neighbour node(s) into this mountain until all edges of this mountain are reached. Then it continues to find the next mountain ridge, starting with the grid node that has the largest node's local density among the unlabelled nodes, and keep labelling nodes until all mountain ridges are detected. There are two strategies to terminate mountain ridges searching. One of them is by checking if the starting node has a larger local density than the node containing noise or not. Another way to terminate it is to see whether the number of nodes labelled into one mountain ridge is larger than some constant or not. The number of mountain ridges corresponds to the number of clusters.The algorithm is as given below:

1. Read multivariate understudy information from csv dataset.
2. Normalize and scale the understudy information into [1,N] framework;
3. Calculate hub's neighbourhood thickness;
4. Find mountain edges, If there is no clamor in the information, mark the individual l designs and the conceivable fringe designs into various mountain edges; generally, other than the naming procedure, the identification of commotion is additionally connected

VI.EXPERIMENTAL RESULTS

The experimental clustering approaches were carried out using php environment. PHP represents Hypertext Pre-processor. PHP is an incredible and broadly utilized open source server-side scripting language to compose powerfully produced site pages. PHP contents are executed on the server and the outcome is sent to the program as plain HTML.PHP can be coordinated with the quantity of mainstream databases, including MySQL, PostgreSQL, Oracle, Sybase, Informix, and Microsoft SQL Server. The current implementation consists of basic implementation of gid based and db scan approaches for analysis.

A. Space Consumption

The comparative analysis of space consumption among the algorithms is as given below.

Table: Space consumption

	Existing	Proposed
Space	2666700	2111450
	2450500	2222333
	2000700	2000522
	2004599	2000049

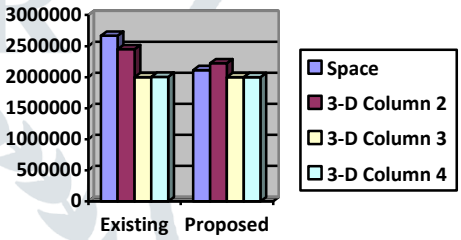

Figure: Space consumption graph

The above space consumption reading clearly shows that the proposed approach has less space consumption as compared to traditional existing algorithm. The graphical analysis is more explanatory as it shows that the reading is much higher in case of improved approach therefore the most efficient one. In the X-Axis the performance of algorithm is denoted and in Y-Axis the space consumed by the approach is denoted.

B. Time Analysis

The comparative analysis of running time among the algorithms is as given below.

Table: Execution time

	Existing	**Proposed**
Time (in millis)	100000	30000
	155000	45000
	145000	45300
	203000	55500
	230000	64000

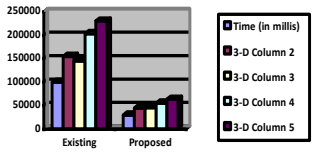

Figure: Analysis for execution time

The above execution time reading clearly indicates that approach proposed has less execution period in milliseconds as compared to the existing approach.The graphical analysis is more explanatory as it shows that the reading is much higher in case of improved approach, therefore the most efficient one. In the X Coordinate-Axis the performance of algorithm is denoted and in Y-Axis the execution period in milliseconds is denoted.

C. Dataset Used

The dataset used is the UCI dataset which is collected from following sources:

- https://archive.ics.uci.edu/ml/datasets/User+Knowledge+Modeling
- https://data.world/datasets/multivariate

VII. Conclusion and Future Scope

Currently there are many algorithms used for clustering data. With increasing high dimensionality of data more sophisticated algorithm are used like Density based and hierarchical clustering approaches. The proposed Fast Clustering Algorithm is utilized for choosing the subset of highlights. A Fast grouping calculation renders proficiency and viability to find the subset of highlights. Quick grouping calculation work should be possible in two stages. The initial step is to moving out immaterial highlights from the dataset, the superfluous highlights are evacuated by the highlights having the at incentive over the predefined limit.The implementation focuses on conducting and experimental study to evaluate our algorithm against SCDOT andDBSCAN on representative spatial datasets .Future extension expects to actualize the calculation for ongoing web of things based stream.

References

[1] Nelofar Rehman," Data Mining Techniques Methods Algorithms and Tools", IJCSMC, Vol. 6, Issue. 7, July 2017, pg.227 – 231

[2] I.A.Venkatkumar & S.J.K Shardaben ," Comparative study of Data Mining Clustering algorithms",IEEE,2016

[3] G.Thangaraju , J.Umarani & Dr.V.Poongodi," Comparative Study of Clustering Algorithms: Filtered Clustering and K-Means Cluttering Algorithm Using WEKA" , International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, Issue 9, September 2017 .

[4] T. Velmurugan & T. Santhanam," A Survey of Partition based Clustering Algorithms in Data Mining: An Experimental Approach". Information Technology Journal Volume 10 (3): 478-484, 2011.

[5] J.Yadav & M.Sharma," A Review of K-mean Algorithm", International Journal of Engineering Trends and Technology (IJETT) – Volume 4 Issue 7- July 2013.

[6] K. Chitra & Dr. D.Maheswari," A Comparative Study of Various Clustering Algorithms in Data Mining", IJCSMC, Vol. 6, Issue. 8, August 2017.

[7] Han, J. and Kamber, M. Data Mining- Concepts and Techniques, 3rd Edition, 2012, Morgan Kauffman Publishers..

[8] P.Nagpal & P.Mann," Comparative Study of Density based Clustering Algorithms", International Journal of Computer Applications (0975 – 8887) Volume 27– No.11, August 2011

[9] Han, J. and Kamber, M. Data Mining- Concepts and Techniques, 3rd Edition, 2012, Morgan Kauffman Publishers.

[10] Bo Wu." A Fast Density and Grid Based Clustering Method for Data with Arbitrary Shapes and Noise", IEEE,2010.

[11] R. Elankavi, R. Kalaiprasath & R. Udayakumar," Fast Clustering Algorithm For High-Dimensional Data". International Journal of Civil Engineering and Technology (IJCIET) Volume 8, Issue 5, May 2017.

[12] X.Wu, H.Jiang and C.Chen**,"** SCMDOT: Spatial Clustering with Multiple Density-Ordered Trees". International Journal of Geo-Information, May 2017.

[13] Ishwank Singh , A Sai Sabitha & Abhay Bansal ," Student Performance Analysis Using Clustering Algorithm",Ieee 2016