Implementation of DBSCAN Clustering to Relate Various Parameters to Predict Primary Education Growth Based on Previous Data

Mayank Mittal, Nitin Goyal, Mohit Kumar

ABSTRACT--- Primary Education can be defined as first step of compulsory education which contributes in the development of future of country. Our aimis to develop a prototype based on DBSCAN Clustering Algorithm to monitor the primary education and corresponding utilization of previous data to estimate and predict the future growth of primary Education in India.The major goal of primary education is achieving basic literacy and numeracy amongst all people as well as establishing foundations in science, mathematics, geography, history & social sciences. DBSCAN Clustering Algorithm can be utilized to establish relationship between human resources, infrastructure, government expenditure, actual utilization of these resources and the outcome which is socio-economic makeover of Society of India. It would help to predict and formulate correct path to transform the process of growth of Indian Primary Education System.

I. INTRODUCTION

This project is basically a research-based project. In this project we tried to establish the relationship between Indian Primary Education and previously available data. This project is based on DBSCAN[4] clustering algorithm for which we have developed an algorithm based on DBSCAN [1]Clustering[7] algorithm which is a density based spatial clustering of applications with noise. It can be extended to develop a higher level system which can predict the required resources for Indian Primary Education plans and can monitor the utilization of these allocated resources. In this project we consider some of the factors that affect Indian Primary Education and we also specify the most dominant factors that affect Indian Economy.

The current health of primary education in India is not a very good fact to cherish upon. Despite of major efforts and projects by the Indian government to ameliorate the current scenario of primary education in India the situation is still worse. Thousands of crores of rupees and expert committees havecome and gone but their efforts could still not be seen anywhere.

We have back tracked available previous data on these primary factors and applied DBSCAN Clustering Algorithm to develop some interesting relationships that have direct contribution to the Indian Economy.

We have made an effort to develop a prototype using DBSCAN Clustering Algorithm to develop clusters of states and develop a method to map the effect of various factors that are part of primary education with the economy of various states of India and our prototype is an effort that can help us to find better ways of fund utilization and programs in developing primary education so as to grow the Indian Economy.

Primary Education:

Primary Education can be defined as first step of compulsory education which contributes in the development of future of country. The major goal of primary education is achieving basic literacy and numeracy amongst all people as well as establishing foundations in science, mathematics, geography, history & social sciences.

Factors Affecting Primary Education:

- States
- Girls Enrolment Percentage
- Teacher Student Ratio
- Number Of Govt. Schools
- Number Of Private Schools
- Enrolment In Govt. Schools
- Enrolment In Private Schools
- Number Of Female Teachers

GDP:

- Gross Domestic Products refers to the market value of all final goods and services produced within a country in a given period.
- It is often considered as indicator of a country's standard of living.
- Gross Domestic Product is related to National accounts, a subject in macroeconomics .
- GDP= private consumption + gross investment + govt. spending + (exports - imports)

Factors Affecting GDP:

- Indian agriculture contribution
- Industry contribution
- Service contribution
- Total GDP

Revised Manuscript Received on August 14, 2019.

Mayank Mittal, Assistant Professor, Dept. of CSE, DVSIET, Meerut, Uttar Pradesh, India.

Nitin Goyal, Assistant Professor, Dept. of CSE, DVSIET, Meerut, Uttar Pradesh, India.

Mohit Kumar, Assistant Professor, Dept. of CSE, DVSIET, Meerut, Uttar Pradesh, India.

II. DBSCAN ALGORITHM

This algorithm tries to classify the data points based on the density of the data points in vicinity. As the density or the number of data points available in defined unit of space is monitored to classify these data points it can be problematic sometimes when the density is variable. However the DBSCAN Algorithm works great for normal densities. There are two basic parameters those control the DBSCAN Algorithm. **Epsilon**, which is the minimum distance between two points those must be considered to be classify on the basis of distance, this is like radius of a virtual sphere which can be drawn considering current data point as centre of this sphere. Second thing is **minimum number of data points,** permissible for consideration for the classification. If a data point having epsilon radius and having n number as minimum number of data points is considered as Core point.

There are three types of data points.

Core Points,

data points having more than or equal to minimum number of data points around the data point having epsilonradius is known core point. Data Points within the epsilon can be further explored.

Border Points,

data points having less than minimum number of data points around the data points having epsilon radius is known as border point.

Outliers,

the data points not core point nor border point is known as Outliers.

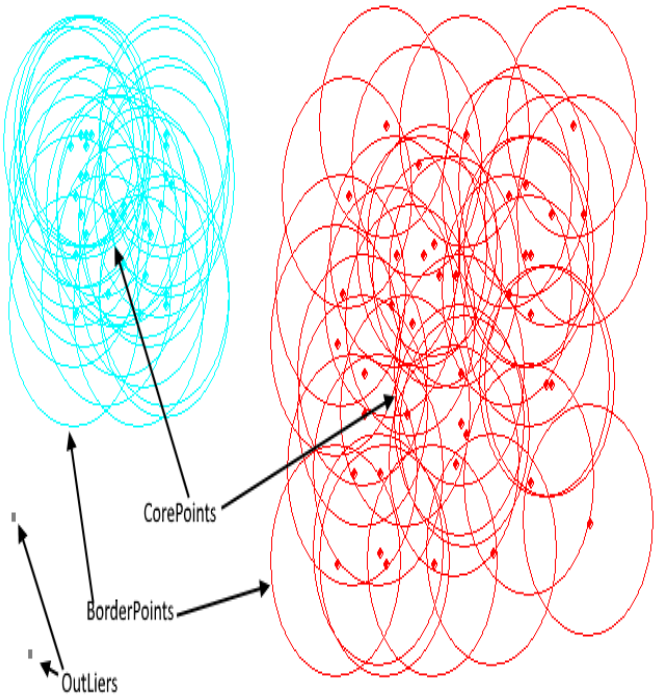

Fig 1. DBSCAN (Core points, Border Points and Outliers)

The results may change or vary drastically based on these parameters, so Epsilon and minimum number of data points must be chosen mathematically. on the initialization of algorithm , when it finds a core point it reaches all the neighbours of that point one by one. Reaching itsneighbour it finds if that Neighbour is a core point or not. If that point is a core point all its neighbour are also visited and this process is carried on till it reaches an end. However, if it encounters a neighbour as border point, only that point is considered within the cluster. The neighbour of such a point is not visited. Once all the points which are reachable from the initial point of start are done, a cluster formation is completed.

The Algorithm is implemented in JAVA.

Following is the steps involved to implementation of the algorithm

- There are two arrays (A_v and A_c of the size of the data set received by the program. A_v stores if the data point is visited and A_c stores the cluster id corresponding to a gene Id. All indices of arrays are mapped to a gene Id.
- In second step a Matrix is created which is identified as distance matrix having number of rows and columns equal to number of genes. Matrix can be identified as distanceMatrix.
- Algorithm can start from first data point. We traverse rows and search the gene Ids which are within Epsilon radius from that gene Id, than we check if the number of data pointsnear to this point having greater than or equal to the specified minimum number of data points.
- If above conditions are satisfied, than current point is Core Data Point and we have to further expand the cluster by calling a separate method (explore()).
- We mark each visited data point in Array A_v as the points are traversed one by one.
- For each gene Id if the gene Id has more neighbours, within Epsilon, than specified minimum points, then all it neighbours are also visited in the same call of the explore() method.
- If there is a neighbour whose number of neighbour given from search() method is less than minimum number of points, then that point is a border point. For such points, their neighbours are not included to visit in the same call of the explore() method.
- Once, one call to the explore() method is over, all the points visited are marked or come under one cluster Id.
- Program terminates when all data points are visited.

III. EVALUATION SETUP AND ENVIRONMENT

We have taken relevant data set related to primary education in India which belongs to different states and union territories of India. These data sets have huge data regarding several dimensions like number and type of school, number and type of enrolments in these schools, the faculty and the infrastructure available, amount of funds provided by the government. This data give us wide contrast of generated demand each and every respective year or state and supplied infrastructure to this demand. We can predict the pattern of demand on the basis of this setup and the corrective measures can be taken on the basis of these predictions.

We have used JAVA and Python as programming language to implement the algorithm. We can vary and adjust optimise the results on the basis of epsilon and number of minimum data points within the epsilon.

Jaccard [2] index and Rand [3] index are used for experimental evaluation and measuring the similarity. Jaccard index is also known as Intersection over Union or Jaccard similarity index.

Jaccard'sindex can be defined as

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

$$J(A,B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Where $0 <= J(A,B) <= 1$, if A and B both are empty than $J(A,B) == 1$.

Rand Measure or Rand Index is used to measure the accuracy or similarity between data clusters in Data Mining. Rand index can be used even when the labels are not used.

If we want to compare two X and Yare partition of Set $A = \{\alpha 1, \alpha 2, \alpha 3 \ldots\}$, where partition $X = \{x1, x2, x3 \ldots\}$ having r subsets and partition $Y = \{y1, y2, y3 \ldots\}$ has s subsets.

Than let k be the number of pairs having same number of subsets in X and Y, l be the number of pairs with different subsets in X and y, m be the number of pairs in A that are same in X and different in Y and n be the number of pairs in A that are different in X and same in Y.

Than Rand index R, can be defined as

$$R = \frac{k + l}{k + l + m + n}$$

In other words, k + l can be considered as the number of agreements in X and Y and m + n can be considered as number of disagreements between X an Y.

We can predict the accuracy and similarity among data using Jaccard and Rand Indices.

We have used Windows laptop with i3 processor and four GB of RAM for the experiments.

IV. EXPERIMENTS AND RESULTS

Several results are received and analysed with given data. Number of tests, amount of data and fields of data can be varied according to the requirements. We have considered only few test cases. Data can be analysed and inferences can be drawn on the basis of different test.

India as a country has several states and union territories. For example the distribution of male and female literacy in India is variable state wise or even district wise. If we categorise the states having same distribution of male and female literacy than using DBSCAN we can take district and state wise data and cluster the data to identify states and union territories having similar literacy distribution.

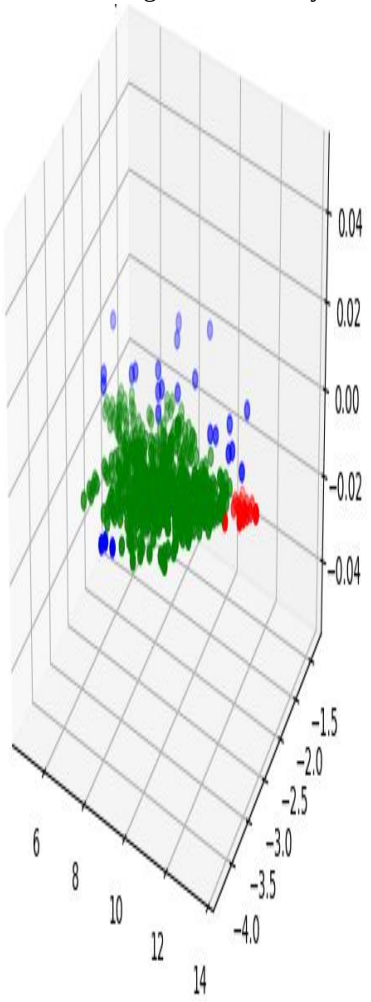

Fig 2.3-D Graph for Sex Ratio district wise

The above graph displays the distribution of sex ratio in different districts. Graph is able to divide the districts into three categories. First category where sex ratio where is less than one, second category where sex ratio is greater than one and the outliers shown in blue where it unexpected. There can be several reasons for this this malicious or wrong da or exceptional sex ratio in these districts.

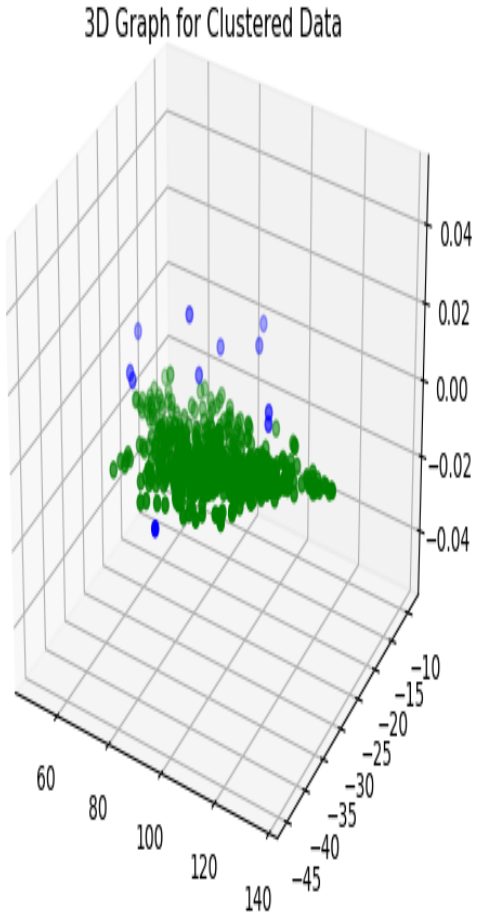

Fig 3.3-D Graph for male female literacy ratio district wise

Above graph shows district wise male female literacy. Blue colour dots shows the outliers, these are those districts either having inappropriate data or large difference in literacy.

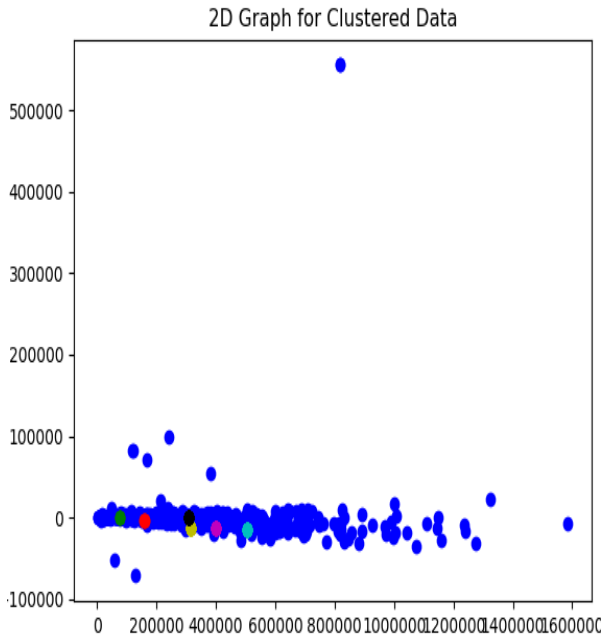

Fig 4.2-D Graph for population and number of schools district wise

Above graph displays the number of schools allocated for given district population. Dots in the graph having colour other than blue displays the districts with exceptional or inappropriate data.

Our research is based on non-linear problem that has a lot of uncertainty associated with it.

Indian Educationalsystem[8][9] is prone to lot of uncertainties with it. The exact behaviour and pattern of growth of Indian Economy cannot be assessed with surety.

The factors that contribute to Indian Economy are subjected to constant change and uncertainties. These are time-variant and area-variant domains and no sure relationship can be established with GDP.

V. CONCLUSION

Education is the chief defence of any country. In order to compete and stand together with the major developed economies of the world, INDIA needs to have a very well planned and properly implemented quality education system. Currently many new mega projects have been launched by the government to ameliorate the lacking education scenario of INDIA. Free Education scheme for poor children under 14yrs has been recently launched by government.This streamline the government efforts in order to channelize proper fund utilization so as to ensure that it benefits the common masses. This proposed technique helps us to estimate the relationship between Indian Primary education and Indian Economy in order to specify the area where we lack of. Once we will know the area of improvement, we will be able to take necessary steps to improve it. By analysing the key parameters of Indian GDP with the major factors of Primary Education, we have tried to develop clusters that depict areas of country where the growth is not synonymous with education. These areas are our concerned regions of country where we need to properly develop resources and channelize funds so that a proper growth in education system can be implemented throughout the country in order to ensure quality education system.

REFERENCES

1. Ester, Martin, Kriegel, Hans-Peter, Sander, Jrog, Xu Xiaowei(1996), Simoudis, Evangelos, Han, Jiawei, Fayyad, Usama M. A density-based algorithm for discovering clusters in large sp.
2. Kosub, Sven, " A note on the triangle inequality for Jaccard distance".
3. W. M. Rand(1971), "Objective criteria for the evaluation of clustering methods", Journal of the American Statistical Association.
4. https://en.wikipedia.org/wiki/DBSCAN
5. Ling, R. F. (1972-01-01). "On the theory and construction of k-clusters". The Computer Journal.
6. Lloyd, S. "Least squares quantization in PCM",IEEE Transactions on information Theory"

7. https://en.wikipedia.org/wiki/Cluster_analysis
8. National educational policy, 1986,https://mhrd.gov.in/sites/upload_files/mhrd/files/document-reports/NPE86-mod92.pdf
9. S. Chandra, A. Joshi, M. Guar, "Report of the Review Committee on the Delhi School Education Act".
10. Everitt Brian, Cluster analysis Chichester, West Sussex, U.K.
11. Dunn, J. "Well separated clusters and optimal fuzzy partitions" Journal of Cybernetics.