Data Cluster Algorithms in Product Purchase SaleAnalysis

[1]Manjushree Nayak, [2]Dr. Bhavana Narain
[1]Research Scholar ,[2]Associate Professor
[1]MATS School of IT,MATS University,Raipur(CG),India,[2]MATS School of IT,MATS University,Raipur(CG),India

***Abstract*—** Big data has brought dramatic change in Online marketing. To analyze online big data there are various tools and methods. Clustering algorithm helps to partition online data and calculate various parameters. Main four algorithms plays an important role for data partition. In our paper we have reviewed four algorithm in second section. In third section we have discussed the methodology used in performance analysis of four clustering algorithm. In fourth section result and analysis of output. And in fifth section conclusion of the work we have done.

Keywords—Cluster;Algorithm;k-mean

I. INTRODUCTION

Purchase and sale of product is not any greater in targeted region, it could be visible within the shape of huge information transfer. R and Python are majorly used in evaluation of those varieties of statistics [1]. Bigdata performs crucial role in online world. product charge within the on-line international is extraordinarily obvious & can be a referred as an crucial component for on line buy. alternate in product pricing which is also taken into consideration as dynamic pricing isn't new & used by many to growth sales, its advantage to on line outlets. based totally on one of a kind strategies various models such as agent primarily based modeling, inventory based model, information pushed model, game concept model, machine getting to know version, simulation model, auction primarily based model has been proposed via distinct researchers [2]. All Models are based totally on one-of-a-kind clustering algorithm. on this paper we've reviewed 4 clustering algorithm and analyzed the overall performance of all four clustering algorithm with appreciate to parameter.

II. CLUSTERING ALGORITHMS

As we realize that clustering is a task for which many algorithms have been proposed. no clustering method is universally applicable, and distinctive techniques are in desire for extraordinary clustering purposes. clustering is beneficial in various fields like industries, education, banking, enterprise and agriculture and exceptionally in enterprise decision making. the diverse clustering algorithms discovered in the literature into awesome categories . The methods of various clustering algorithms can be extensively labeled underneath.A number of the essential clustering methods are as follows.

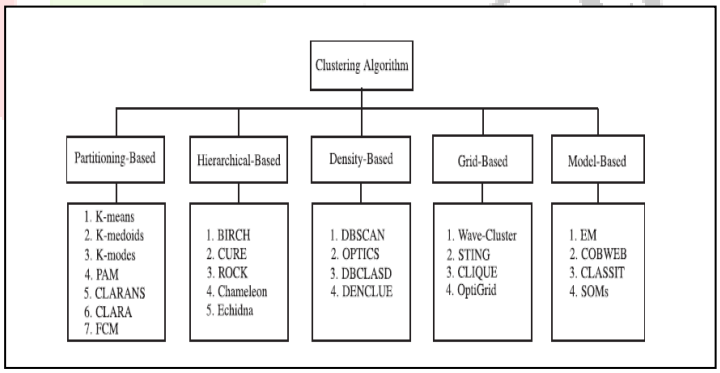

Figure1 : An Overview of clustering Algorithm [3]

Some of the important clustering methods are as follows.

II.1 Hierarchical Clustering Techniques wreak a hierarchical structure at the statistics objects and their step-sensible clusters, i.e. one excessive of the clustering structure is most effective one cluster containing all the items, the other severe is a number of clusters which equals the wide variety of objects. records are organized in a hierarchical way depending at the medium of proximity,proximities are received by using the intermediate nodes. a dendrogram represents the datasets, where individual data is provided by leaf nodes. The preliminary cluster step by step divides into numerous clusters because the hierarchy keeps. hierarchical clustering methods may be agglomerative (bottom-up) or divisive (pinnacle-down). An agglomerative clustering begins with one object for every cluster and recursively merges two or greater of the maximum suitable clusters. a divisive clustering begins with the dataset as one cluster and recursively splits the maximum suitable cluster. The process continues till a

stopping criterion is reached (frequently, the asked variety of k clusters). This hierarchical technique has a first-rate drawback though, which pertains to the fact that once a step (merge or break up) is accomplished, this cannot be undone. BIRCH, CURE, ROCK and Chameleon are some of the well-known algorithms of this class[4]

In Density Based Methodin density based method, this algorithm is devolped through Martin Ester, Hans-Peter Kriegel, Jorge sander and Xiaowei Xu in 1996. Density based method totally belong to partitional clustering. density primarily based clusters are defined as clusters which are differentiated from other clusters by using various densities meaning a collection which have dense area of gadgets can be surrounded by way of low density regions. density based method are of types: density based totally connectivity and density primarily based capabilities. density primarily based connectivity is related to training statistics point and DBSCAN and DBCLASD comes beneath this while density features is associated with information factors to computing density features defined over the underlying attribute area and DENCLUE comes beneath this[5]

2.1 Partitioning clustering methodsthis method partition the data object set into clusters where every pair of object clusters is either distinct (hard clustering) or has some members in common (soft clustering).Partitioning clustering begins with a starting cluster partition which is iteratively improved until a locally optimal partition is reached. [6] [7]. K-mean is the most popular method of working.

k-means procedure is easily programmed and is computationally economical, so that it is feasible to process very large samples on a digital computer.Partitioning an N-dimensional Population into k sets on the basis of a sample. The process, which is called 'k-means,'

The k-means process was originally devised in an attempt to find a feasible method of computing such an optimal partition.

The k-means procedure consists of simply starting with k groups each of which consists of a single random point, and thereafter adding each new point to the group whose mean the new point is nearest. After A point is added to a group, the mean of that group is adjusted in order to take account of the new point. Thus At each stage the k-means are, in fact, the means of the groups they represent [8][9].

This is the Pseudo code of the k-mean algorithm.

1. *Set the initial seed/Centre points*
2. *Do*
3. *For Each data point in the data set do*
4. *Find the nearest seed point using the distance metric*
5. assign the facts factor to the cluster with the closest seed point.
6. *End For*
7. *Calculate the mean of the data point in each cluster*
8. *Assign these mean values as the new center point*
9. *While at least one data point changes cluster*

The main function of k-means is to minimize the distance between the objects in a cluster & maximize the distance between the clusters[9]

K-means is the process of partitioning an N-dimensional data into k sets on the basis of a sample.

$W^2(S) = \sum k\, i = 1 . \left| z-\mu_i \right|^2 dp(Z)$

is low for the partitions S.

Where ,

S=Data set

P=Probability mass function for the data set S.

S={S1, S2, ……., Sk}

µi=i=1,2,……..,k is the conditional mean.

- K-means procedure is easily programmed and is computationally economical, so that it is feasible to process very large samples on a digita computer.
- K-means represents a generalization of the ordinary sample mean.
- The K-mean procedure consists of simply starting with k groups each of which consists of a single random point, and thereafter adding each new point to the group whose mean the new point is nearest. After a point is added to a group, the mean of that group is adjusted in order to take account of the new point.

  Thus at each stage the k-means are, in fact, the means of the groups they represent.

  Let,

$E_{N=}$z1,z2……=Random sequence of points (vectors) .
P=fixed probability measure, used to select each point, independently of the preceding ones.
Thus,
$P[z_1 \in A] = p(A)$
$P[z_{n+1} \in A \mid z_1, Z_2, \ldots - , Zn] = p(A)$, n = 1, 2, ….,
A=Measurable set in E_N.
Relative to a given k-tuple
X = (X1, X2, …… , Xk), xi E E_N, i = 1, 2, ….., k,
$S(x) = \{S_l(X), S_2(X), \ldots\ldots , S_k(X)\}$ of E_N,
Where S(x)=minimum partition.
minimum distance partition S(x) is defined.

(1) $S_l(x) = T_l(x)$, $S_2(x) = T_2(x)S_1(x)$,…… -,
$S_k(X) = T_k(X)S_1(X)S_2(X)$ ….. Sk-l(^{X}),
where

(2) $Ti(x) = \{\varepsilon:\varepsilon \epsilon E_N, I \varepsilon - x_i I \leq I \varepsilon - x_j I, j = 1, 2, \ldots ,k\}$. Si(x) set contains the points in EN nearest to x_i, [11].

2.2 Expectation–maximization (EM) algorithm

The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step[10].

III. METHODOLOGY

We have developed a program in R . In R program we have taken four clustering algorithm and tested in four parameters
Step 1. Choose the clustering algorithms: To perform the comparative analysis, four clustering algorithms are chosen namely Hierarchical algorithm, EM algorithm, Density based algorithm and partition based algorithm.
Step 2. **Choose the dataset:** The "customer" data set has been chosen and downloaded from "http:/http://archive.ics.uci.edu/ml/datasets/online+retail" in excel sheet we have converted.xlsx sheet to .csv file format.
Step 3. Used dataset in R: We haveloaded data file for analysis.
Step 4. Preprocess data: After loading of the dataset the next step is to preprocess the dataset using Caret package in R. This is our final dataset in file Online _Retail.csv file.
Step 5. Apply clustering algorithms: We have used this preprocessed file for four clustering algorithms.
Step 6. Store the result: We have run four algorithms and results are stored into the tabular form. Our parameter are number of cluster formed, number of iteration, time taken to build clusters, accuracy of clustered data.
Step 7. Plot the graph: we have represented results in graphical format
We have done programming in R language to get given table

Algorithm	Number of Clusters	Cluster instance	Number of Iteration	Time	Accuracy
Hierarchical Algorithm	4	61%	7	0.20s	58.34%
EM Algorithm	4	39%	5	0.78s	57.56%
Density based Algorithm	4	74%	5	0.35s	55.33%
Partition Algorithm	4	75%	7	0.25s	59.60%

Table 1 : Result of performance of four Clustering Algorithm

IV. RESULT AND ANALYSIS

After performing the practical work of clustering algorithms, now we have discussion of these four algorithms. This section presents the experimental result of each of the four clustering algorithms using Online_Retail data. The experimental results are presented in Table 1.The simulation result is obtained for all four clustering algorithms.

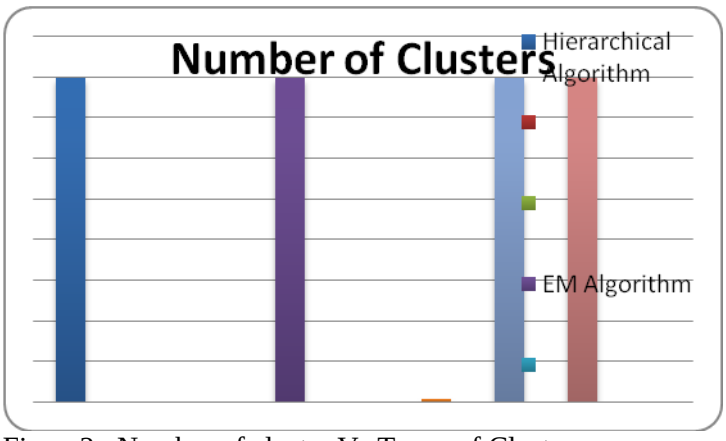

Figure2 : Number of cluster Vs Types of Clusters

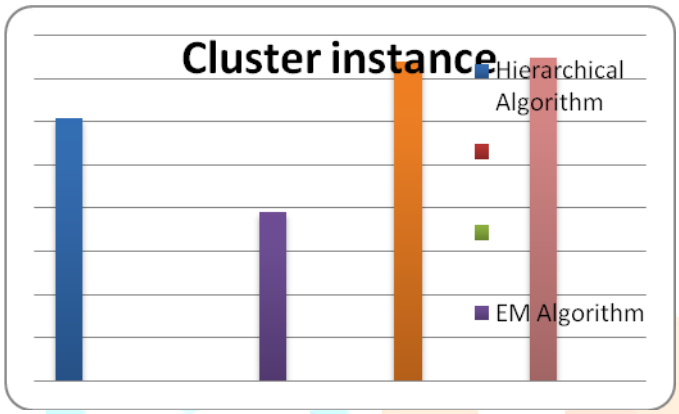

Figure3 : Clusteinstancer Vs Types of Clusters

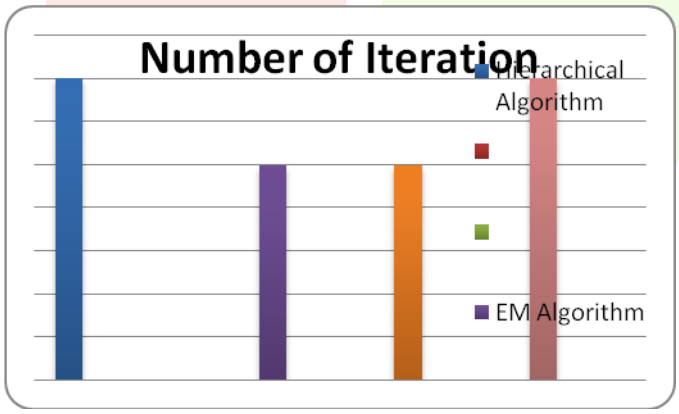

Figure4 : Number of Iteration Vs Types of Clusters

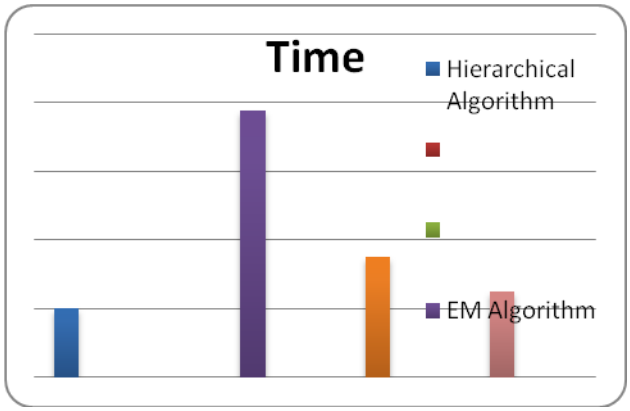

Figure5 : Time taken Vs Types of Clusters

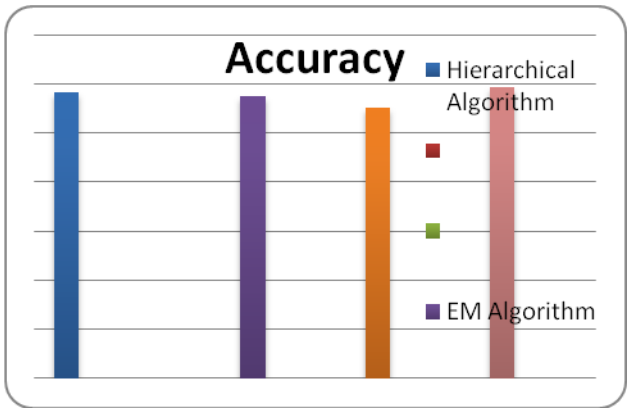

Figure6 : Accuracy of cluster Vs Types of Clusters

V. CONCLUSION

On this paper, comparative study has been performed on the k-means ,Hierarchical, EM , Density primarily based clustering algorithms and partition based algorithm. Comparision is done on on line _retail dataset the use of R language and the comparative effects are presented in the form of table and graph. the comparative study is achieved on the basis of accuracy and efficiency parameters. Hierarchical clustering takes greater time to shape clusters and much less accuracy with normalized dataset. Density based clustering form clusters with equal accuracy as partition based clustering. After applying normalization only simple partition clustering algorithms forms clusters with less time and more accuracy than other algorithms. In terms of time and accuracy partition cluster algorithm produces better results as compared to other algorithms.

References

[1]. Manjushree Nayak, Bhavana Narain, "Imapact of R and Python in Big Data Analysis",International conference on Innovative trends in Engineering Science and Management (ITESM-2017) 10-11 November 2017,MSEIT, Raipur.

[2]. Rajan Gupta and Chaitanya Pathak,"A Machine LearningFramework for Predicting Purchase by onlinecustomers based on Dynamic Pricing", ScienceDirect, Procedia Computer Science 36 2014, pp-599 – 605.

[3]Fahad A,Alshatri N,Tari Z,Alamari ".A survey of clustering algorithms for Bigdata:Taxonomy ad Empirical analysis" .IEEE Transcations on Emerging Topis in computing.2014 Sep;2(3):267-79

[4]Pooja Batra Nagpal,Priyanka Ahlawat Mann"comparative study of Density based clustering algorithm"International Journal of Computer Applications (0975 – 8887) Volume 27– No.11, August 2011.

[5].Pooja Bhandari,"Study of Various Clustering Algorithms Used by WEKA Tool",International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 ,Volume-4, Issue-8,pp-37-40.

[6].Raj bala,Sunil Sikka, Juhi Singh,"A Comparative Analysis of Clustering Algorithms", International Journal of Computer Applications (0975 – 8887) Volume 100 – No.15, August 2014,pp-36-39.

[7]. Rui Xu, "Survey of Clustering Algorithms", IEEE transactions on neural networks, vol. 16, no. 3, may 2005, pp 645-678.

[8]. Kardi Teknomo,"K-Means Clustering Tutorial",http:\\people.revoledu.com\kardi\ tutorial\kMean, Last Update: July 2007.

[9] Ashish kumar singh& grace rumantir,"clustering Experiments on Big transaction Data for market segmentation"Proceedings of the international conference on bigdata science and computing,Article No-16,2014.

[10]MacQueen, J. (1967) Some Methods for Classification and Analysis of Multivariate Observations. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, 1, 281-297.
[11]. Garima Sehgal, Dr. Kanwal Garg,"Comparison of Various Clustering Algorithms",(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, pp-3074-3076.

Author

Mrs.Manjushree Nayak

.

Mrs.Manjushree Nayak,pursuing Ph.D at MATS Universities,Raipur,has been working in the field of computer science last 15 years.specialization in Machine Learning, Bigdata, R &Python.

Dr. Bhavna Narain

Dr. Bhavna Narain is working as Associate Professor in MSIT Department of MATS University Raipur, C.G. Ph.D. (Computer Application), M.Phil (Computer Science), Her work experience is fifteen years. Her area of interest is computer networks (adhoc, mesh) and Digital Image Processing. She has published 41 papers in national and international journals and conferences, two books and worked as co-pi in two minor projects. She has been working as state student coordinator in Computer Society of India and awarded as best teacher and researcher by national organization.