International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-3, February 2020

Quality Healthcare Prediction using K-Means And Clara Partition Based Clustering Algorithm For Big Data Analytics

Madhura Chinchmalatpure, M.P. Dhore

Abstract*: Big Data is a collection of large or vast amount of information that grows at ever increasing rates. Big data is ordered, unstructured, semi structured or mixed data in natural world. Researchers are designing, implementing, analyzing different application. In medicinal industry large or vast amount of data is available but people are not able to extract the significant information. Healthcare big data analytics (HBDA) becomes "Healthier analytics" by fusion of techniques. In this paper, we discuss and implement algorithms of clustering using R-Studio tool. Clustering is defined as the method of partitioning set of patterns into similar groups called as clusters. We can extract the data from vast datasets in the form of clustering rules. These clustering techniques are scalable. Also, we compare the accuracies of two partition based clustering techniques k-means and Clara on healthcare datasets for giving good quality of healthcare services. Implemented results demonstrate the k-means method gives better accuracy values than Clara algorithm.

Keywords* : Healthcare Big Data analytics(HBDA), Partition based clustering techniques, Electronic Health Record(EHR), K-Means, Clara, Internet-Of-Things(IOT)

I. INTRODUCTION

Big Data is a collection of large or vast amount of data bring benefits to healthcare and gives good quality of healthcare. Big Data is ordered,unstructured,semistructured or mixed data in natural world[1]. In medicinal industry large or vast amount of data is available but people are not able to extract the important information. It makes possible to discover the healthcare trends, prevention of diseases and gives good quality of healthcare services.

Data mining Techniques will help doctors and patients to get improved quality of healthcare. A challenging job in medical industry is to provide effective treatment to patient and to examine the disease correctly. In Digitized world, vast amount of data is generated, to properly analyze that data big data concept is generated. The major reason of this epidemic are changes in lifestyle, such as inactive jobs, due to unhealthy diet, enlarge in job stress, dependence on smoking and tobacco.

Big data is characterized by multi "V" representation.

volume(amount of data) : Continuous monitor of healthcare sources
variety(range of data types and sources) : demonstration of data through variety of sources like EHR
velocity(speed of data in and out) : Fast speed processing for clinical decision support
veracity (e.g. Medicinal reports, electronic Health Record (EHR), biometrics information etc.) : Data mining from special environment

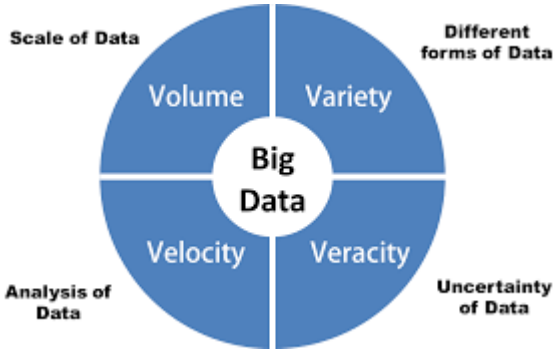

Fig 1: 4 V's of BDA

In medicinal industry large or vast amount of data has generated so it uses EHR for patients information, clinical report, doctors prescription, investigative reports, pharmacy information, health insurance data, data from social media and medical journals records. All these information collectively forms big data in medicinal datasets.

Machine Learning is a purpose of Artificial Intelligence, it focuses on progress of computer programs and analyze the big data. Data Mining is defined as pull out of indefinite information about data. Here we take three datasets as Antibiotics, medicare and unplanned hospital visits-Hospital database is designed for clinical purposes. In this we have to analyze,compile, summarize and organize machine learning challenges with Big Data. Big data is challenging research area, it refers to tool such as R-Studio that are applied to healthcare datsets to obtain the data from database in which to collect present data, preprocess data and evaluate data.[2] Analytics focus on statistical and mathematical analytics of data.

Revised Manuscript Received on February 10, 2020.
* Correspondence Author
Madhura Chinchmalatpure, Department of Inter Institutional Computer Cenre, RTM Nagpur University Campus, Nagpur, (MS)-India, madhura.naralkar@gmail.com
Dr. M. P. Dhore, Department of Computer Science, Shri Shivaji Science College, Congress Nagar, Nagpur (MS)-India, mpdhore@rediffmail.com

The analysis helps to identify the problem from the collected data set. Later it uses various algorithms for better output of data.

The purpose of mapping function is when we have the input data(X) we can predict the output variables(Y) for that information. It is called supervised learning because the method of an algorithm learnt as of the training dataset can be thought of as a supervised the learning process. [3]

COLLISION OF BIG DATA IN HEALTHCARE :

Big Data has a combination of different type of data and information and it's main source is Internet of Things (IOT), computer and machine generated data, Human generated data. Big data analytics have created dissimilar ways to assist them in considerable value and identify tools that are suitable for the new era. They focus on the following concept:

• **Right living.** Patients have to play an active role in their own health by making the right choice about diet, exercise, preventive care.

• **Right care**. Patients must be given appropriate treatment available.

• **Right provider**. Professionals who take care for patients must have strong performance records and be able to achieving the best outcomes about their patients.

• **Right value.** Providers should look for ways to get better value while maintain or improve the quality of Electronic health-care information.

• **Right innovation.** Patients focus on identify new therapies and approach to health-care delivery. They should also try to get better the innovation engines themselves, by advancing medicine and boost R&D. [6]

II. BIG DATA ANALYTICS ALGORITHMS :

Algorithms in ML are the important parts of BDA. The process of ML is used to make technique by analyzing large data as follows:

1) **Clustering Algorithms** – All the objects are consists of single clusters. The objects are divided into number of partitions by locating the points between the partitions. Investigation part of clustering is an unsupervised way[3]. The various clustering algorithms are mostly classified in to as follows:

a) Partitioned-based Techniques : FCM, K-means, K-mediods, Kmodes, Progeny, Canopy, PAM, CLARA.

b) Hierarchical-based Techniques : BIRCH, CURE, ROCK

c) Echidna Density-based: DBSCAN, OPTICS, DBCLASD

c) Grid-based Techniques : Wave-Cluster, STING, CLIQUE, OptiGrid.

d) Model-based Techniques : EM, COBWEB, CLASSIT, SOMs

2)**Machine Learning (ML) Algorithms** – : Algorithms in ML are the essential parts of BDA. The process of ML is used to build method by analyzing huge datasets are as follows:

f) Association Techniques: Association Rule Mining-Apriori algorithm

g) Kernel Methods: SVM, RBF

h) Decision Tree Techniques : CART, ID3, C4.5, C5.0

i) Instance based Methods: k-NN, SOM

j) Bayesian: Naïve-bayes, Averaged One Dependence Estimators

k) Ensemble Techniques : Boosting, AdaBoost, Random Forest

l)Regression Techniques : Linear Regression, Logistic Regression, Stepwise Regression,Ridge Regression,Lasso Regression, Elastic Net Regression, Discrete Choice Models,Probit Regression, Logit versus probit, Time Series Model ,Survival and duration Analysis, Multivariate adaptive regression splines. [4]

CLUSTERING CHALLENGES IN BIG DATA :

Big Data is generated throughout with smart devices, wearable devices, healthcare files are placed into datasets. HBDA represents new approaches to analytics and help in following areas like medical treatment, administration, health policy, clinical research and development, public health, patient profile analytics, output of clinical results, fraud detection. Clustering in Big Data is to recognize patterns. The properties of clustering methods are :

Type of Dataset : The composed data in the today's world contains both numeric and categorical attributes as unstructured data. Clustering algorithms work fruitfully either on purely numeric data or on categorical amorphous data.

Size of Dataset : The volume of the dataset has depend on both the time-efficiency of clustering and the clustering value.

Time Complexity : Most of the clustering methods must be repeated frequent times to improve the clustering importance and efficiency.

Stability: Stability corresponds to the capability of an algorithm to produce the partition of the data irrespective of the order in which the data are accessible to the algorithm.

High dimensionality: The number of extent of data increases, the data become gradually more sparse, so the distance measurement between pairs of points becomes meaningless and the average density of points anywhere in the dataset is likely to be low.

Cluster shape: A good clustering algorithm should be able to grip unstructured data and their wide variety of data types, which will create clusters of arbitrary shape. Many algorithms are able to identify only convex shaped clusters. [8]

III. DATASETS

To evaluate the efficiency of our method, experiments on Antibiotics, Medicare and Unplanned Hospital Visits-Hospital datasets are conducted. These Datasets are taken from Data.Medicare.Gov site. Antibiotics dataset has 17 instances and 5 attributes, Medicare dataset has 10001 instances and 17 attributes and Unplanned Hospital Visits-Hospital has 57925 instances and 18 attributes. They are publically available on the internet. Table shows the description of database. This dataset contains information of patients, medicines and gives the quality treatment to patient.

Table I: Sample Dataset

Sr. No.	Database	No.Of Attributes	Size
1	Antibiotics	5	17
2	Medicare	17	10001
3	Unplanned Hospital Visits-Hospital	18	57925

IV. METHODOLOGY

Clustering or information group is an unsupervised learning task in which fixed set of categories are identified as clusters based on the intra class comparison present in the information. Clustering means grouping of connected objects into a set known as cluster. Objects in one cluster are likely to be dissimilar when partially compared to objects grouped under another cluster. Clustering is one of the main tasks in exploratory data mining and is also a technique used in statistical data investigation.

Between each group there is maximum intra class similarity and minimum inter class similarity. Thus clustering technique is applied to identify the intrinsic grouping between a set of data which are not labeled or grouped. This technique can be used when the classes are not known and it does not analyze class labeled instance as used in classification. The attribute which provide the similarity must be identified to increase the similarity metric between clusters. Cluster properties can be evaluated to identify the profiles which distinguish one cluster from the other. The performance of good clustering method is measured by its ability to identify the patterns that are hidden and make maximum intra class similarity and decrease interclass similarity between other objects among clusters.[6]

All objects are measured as a only cluster. The objects are divided into number of partitions by iteratively locate the points between the partitions. Partition based clustering algorithm used to decompose the item into k-clusters so that each partition can be optimized using predefined criterion. The partitioning base algorithm like K-means, K-medoids PAM, CLARA and K-modes. Partition based algorithms can found clusters of Non convex shapes.[8] Some algorithm which can use this concept are planned as follows :

K-MEANS:

The K-Means method is a partitioning based for clustering. K-Means clustering method, groups the data is based on their nearness to each other according to the Euclidean distance. This algorithm is used in all metric spaces. The selection of K cluster centroids is done at random. It reassigns the points to its nearby centroids and recomputes centroids for newly assembled group. It takes ky as input consideration and partition as set of n objects from ky clusters. The mean values of objects is taken as correspondence parameter to form clusters. The cluster mean or center is formed by the random choice of kY objects. By comparing most similarity other object is assigning to the cluster. For each data vector the algorithm calculates the distance between data vector and each cluster centroid.This algorithm works as follows [8] :

Algorithm: Generate K-Means
Input: Training Data
Output: Decision Value

- Load dataset, Take the experiment of gathering a sample of observed values
- select number of k clusters
- Randomly generates k-cluster and evaluate cluster center and generate k random points as cluster centres using k-means function.
- Allocate each and every point to the nearest cluster center
- Use table() and Take Matrices by using Confusion Matrix(), and accuracy is being calculated

CLARA:

CLARA(Clustering large application based on random search) introduced by Kaufman and Rousseau to handle large datasets.. It makes sampling procedure along with PAM. It finds the medoids of the model. The sample is drawn in random way. The medoid of the sample would approximate the medoid of the entire dataset. Here, accuracy is the quality of clustering is measured based on the average dissimilarity of all objects in the entire datasets. In this we can select a node and compare it to user distinct number of neighbors searching for local minimum. This process is repeated if there is no perfect neighbor found.[9,10]

Algorithm: Generate CLARA algorithm
Input: Training Data
Output: Decision Value

- Load dataset, Take out the experiment of assembly a sample of observed values
- For i is equal to 1 to 5, repeat the subsequent steps:
- Draw a sample of objects at casual from the Complete dataset, and call method PAM to find k medoids of the sample.
- For each object in the complete data set, determine which of the k medoids is the a vast amount similar to that object.
- Calculate the average distinction of the clustering obtained in the previous step. If this value is less than the current smallest amount, use this value as the Current smallest amount, and retain the k medoids found in Step 2 as the best set of medoids obtained so far
- Back to the Step number 1 to start the next iteration.
- Use table() and get Matrices by using Confusion Matrix(), and accuracy is being calculated

V. RESULTS AND DISCUSSION

To authenticate the performance of algorithm in our research, we used healthcare datasets such as antibiotics, medicare, unplanned hospital visits. Two algorithms are discussed in RStudio tools to measure the performance of using several matrices and parameters using analysis of datasets. The above mentioned algorithm are compared in terms of evaluation matrices like accuracies calculated using the formula.

Accuracy : Accuracy is used for arrangement events and it is calculated as a ratio of amount of accurate predicted observations to total observations. Accuracy refers to the nearness of a measured values to a standard values or known values.

Table No. 2 represent that we can evaluate Accuracy on three databases using K-Means and Clara partition based techniques and they are used as testing and training purposes.

Table II : Results of Partition Based Algorithm

Sr. No.	Datasets	No.Of Attribute	Size	Field	K-means Accuracy	Clara Accuracy
1	Antibiotics	5	17	Gram	0.4375	0.375
2	Medicare	17	100001	Classification	0.2792	0.3634
3	Unplanned Hospital Visits-Hospital	18	57925	ComparedtoNational	0.1461	0.1285

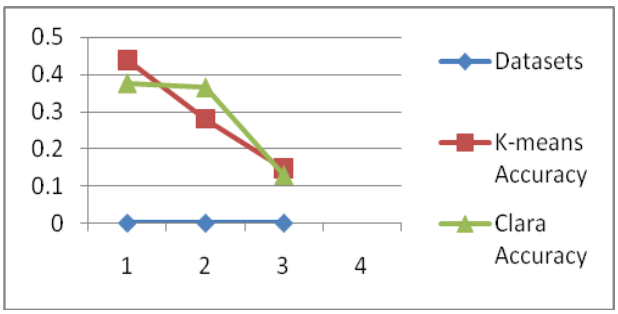

Fig 2: Comparison between K-Means and Clara

The partition based clustering methods k-means and Clara having Accuracies 0.4375, 0.2792, 0.1461 and 0.375,0.3634, 0.1285 on Antibiotics , Medicare and Unplanned Hospital Visits-Hospital datasets.

On y-axis figure shows the accuracies and on x-axis it shows databases of SVM and Naïve Baye's algorithm using data.

On these datasets, above graph shows that accuracy of K-Means Technique has higher performance than Clara, K-means partition based clustering algorithm's Performance is better than Clara Clustering algorithm.

Therefore, the accuracy of K-means partition based clustering algorithm is better than Clara Clustering algorithm.

VI. CONCLUSION

Prediction of quality treatment in healthcare and to build efficient classifier for medical Application is a major challenge in healthcare systems. The objective of projected work is to provide a study of dissimilar data mining classification techniques with their pros and cons, here we studied the two partition based classification techniques are used in Big Data Analytics on three databases Antibiotics,Medicare,Unplanned Hospital Visits-Hospital.

The algorithm is tested in terms of accuracy, the concert of k-means algorithm is better than CLARA. The comparison is on the basis of accuracy and numeral of clusters formed over healthcare datasets has been performed. K-means algorithm gives more accuracy when compared with CLARA. It shows higher performance in terms of all the other algorithms. Hence it is concluded that K-Means outperformed all other algorithms with CLARA for healthcare performance analysis datasets.

REFERENCES

1. Madhura Chinchmalatpure,Dr. M.P.Dhore," Big Data Analytics of SVM and Naïve Bayes Algorithm for Multiple Datasets", IOSR Journal of Engineering (IOSR JEN), www.iosrjen.org ISSN (e): 2250-3021, ISSN (p): 2278-8719, PP 40-43
2. Gemson Andrew Ebenezer J.1 and Durga S.2," BIG DATA ANALYTICS IN HEALTHCARE: A SURVEY", ARPN Journal of Engineering and Applied Sciences ©2006-2015 Asian Research Publishing Network (ARPN). All rights reserved. VOL. 10, NO. 8, MAY 2015, ISSN 1819-6608
3. ALEXANDRA L'HEUREUX, KATARINA GROLINGER, HANY F. ELYAMANY, MIRIAM A. M. CAPRETZ,"MACHINE LEARNING WITH BIG DATA: CHALLENGES AND APPROACHES", DOI 10.1109/ACCESS.2017.2696365, IEEE ACCESS
4. Venkatesh Saravanakumar M, Sabibullah Mohamed Hanifa, " BIGDATA: Harnessing Insights To Healthier Analytics – A Survey "
5. Madhura Chinchmalatpure,Dr. M.P.Dhore, "Big Data Analysis Using Regression Techniques " , 61st IETE Annual Convention 2018 on "Smart Engneering for Sustainable Development" Special Issue of IJECSCSE, ISSN: 2277-9477
6. Madhura A. Chinchmalatpure, Dr. Mahendra P. Dhore, "Review of Big data Challenges in Healthcare Application" , IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727 PP 06-09 www.iosrjournals.org
7. N.Valarmathy,S.Krishnaveni , "Performance Evaluation and Comparison of Clustering Algorithms used in Educational Data Mining", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7, Issue- 6S5, April 2019
8. T. Sajana, C. M. Sheela Rani and K. V. Narayana ,"A Survey on Clustering Techniques for Big Data Mining" , Indian Journal of Science and Technology, Vol 9(3), DOI: 10.17485/ijst/2016/v9i3/75971, January 2016
9. Mugdha Jain,Chakradhar Verma,"Adapting K-means for clustering in Big Data",International Journal of Computer Applications(0975-8887) Volume 101-No.1,September 2014
10. Raymond T. Ng and Jiawei Han," CLARANS: A Method for Clustering Objects for Spatial Data Mining", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 14, NO. 5,SEPTEMBER/OCTOBER 2002
11. B. Kranthi Kiran,Dr. A. Vinaya Babu,"A Comparitive Study of Issues in Big Data Clustering Algorithm with Constraint based Genetic Algorithm for Associative Clustering",International Journal of Innovative Research in computer and Communication Engineering,ISSN(Online):2320-9801

AUTHORS PROFILE

Madhura Chinchmalatpure is currently pursuing Ph.D. from Department of IICC, RTM Nagpur University Campus, Nagpur. She have completed MCA from GHRIIT,Nagpur.

Dr. M. P. Dhore is a Principal in Shivaji Science College, Nagpur. Educational qualification includes M. Sc. (Computer Science), M. Phil. (Computer Science) ,Ph. D. (Computer Science), Five students have been awarded Ph.D. under his supervision. He has published more than 40 research papers in International Journals.