Clustering: Review on Partitioned Clustering Algorithms

Swayanshu Shanti Pragnya
School of Computer Science and Engineering
Centurion University
Jatni, Khurdha Road, 752050

Abstract: **- In present generation Big data is the most impenetrable, enliven and most desirable research subject which is leading in all aspects. Unresolved nature of this analysis is the biggest question to all, going to be solved soon. Literally Cluster is the smallest part of the Analysis purpose which somehow fulfil the side by collaboration part of big data. Big data is a term which identifies datasets which are large in terms of size. Big data introduces unequally identified statistical and computational challenges. This paper is to study on the problems which are related to big data, clustering and comparison in between FCM and K-means algorithm.**

Keyword: ***Big data, clusters, partitioned clustering, clustering algorithms, FCM, K-means.***

INTRODUCTION:

In the software field data analysis is the most vital and significant tool as to process voluminous data. To precise and easy way of keeping, extracting, inserting and modifying data, we need some technique. Big data is not only big by size but also starting from variety to veracity, it is vastly enlarged. So for analyzing these big data, any technique is required must. Cluster is the technique which segregate data in their perspective reason. Data is a collaboration of small lexemes with some valid sense and different from other sense. In a few year of time use of data in databases are increased. For making any suitable decision data are collected

And analyzed is the purpose of keeping data. Database is an organized collection of data so that Data can be easily accessed, managed, updated and inserted.

Big Data:

Data is some facts which may be organized or raw, gathered for use. Any block of information which is meaningful and will be used in further can be define as data. Big data is collection of full volume data sets. Now-a-days big data is becoming the main focus at every place. Big data is related to everywhere in all aspects.

II. CHARACTERISTICS OF BIG DATA:

Previously big data was characterized as 3Vs but in present 2 more characters are included i.e. value and veracity. Variety: It defines as data in the form of structured, UN structured, semi structured, probabilistic, multi factor or statistical manner. Basically collection of different type of data.

1. Value: Value represent a vital role in case of big data. As it means data must have some statistical or hypothetical importance with proper meaning.
2. Velocity: Most often big data are generated at a high speed by including arrays, sensors or multiple events and need to process in near real time or batch.
3. Veracity: This dimension of big data includes 2 aspects i.e. Consistency or certainty and trust worthiness with authentication.
4. Volume: In big data volume indicates to the size, scale, dimension and amount of any data or data sets. Ex more no. of records, tables and file transaction.

III . LIMITATIONS IN BIG DATA:

Heterogeneity and Incompleteness: Though many types of data sets are combined together which leads to heterogeneity. Here working with heterogonous data leads to more challenging

SECURITY AND PRIVACY CHALLENGES:

This is the measure problem seen in case of big data. When personal information combined with such a large scale of data leads to complexity in their privacy of data. Quality of data decreases due to improper utilization.

Fault Tolerance: Two methods which increase this tolerance in big data are

1. Divide whole computation being done by tasks and assign these tasks into different nodes for computation
2. One node is assigned the work of observing that these nodes are working properly. If something happens that particular task restarted.[1]

But sometimes whole computation can’t be divided into independent tasks which lead to fault tolerance.

Endogeneity: It refers to a feedback loop in a system. We can model and predict ants or algae far better than we can model and predict markets.Clustering is one of the solution in big data issues.

IV. CLUSTERING:

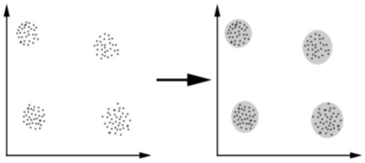


Fig 2 Showing Clustering example [2]

Cluster: Cluster can be defined as group of similar or very similar objects. In a computer system a group of servers and other resources that act alike single system is known as cluster.

Clustering is an ongoing process or a under supervision technique of data mining. Clustering means grouping of similar objects together and separating dissimilar ones.

In this figure we find that 4 clusters where data can be easily divided. But here we make clusters according to the nearest distance from each other i.e. known as Distance based clustering. Here objects are grouped together according to their descriptive fit concept not by simple similarity measures.[2]

V. TYPES OF CLUSTERING:

1. Partitioned Clustering:
   Here classification is done equally. Partitioned algorithms divide or classify the data set in mutually disjoint partitions

Advantages:

- Simple and Scalable
- Suitable for almost all data sets

Advantages:

- Problems involving point linkages are well suited. Ex taxonomy trees.
- Flexibility in embedded regarding the level of granularity.

Disadvantages:

- Making correction after splitting is not in able.
- Interpretability in cluster descriptors are poor

It basically uses 2 parameters Epsilon and minimum points of each cluster and one least point.

Advantage: no need to specify the number of clusters in advance and easily handle cluster with arbitrary shape.
Disadvantage: Not handling the data points with varying densities and results depend on the distance measures.[2]

Major Benefits of cluster are High performance, scalability, system availability, and major issues are scalable performance(scaling of resources), availability support, Fault tolerance and recovery(cluster machines are designed to eliminate all single point failures).

Basically a process of segmenting similar and dissimilar one.

- Spherical clusters which are well in shape.

Disadvantage:

- Clusters descriptors are very poor
- Phase initialization is very sensitive.
- Dealing with non –convex clusters are varying in density and size
- Degradation in high dimensional spaces.
- Entrapment into local optima is very frequent

2. Hierarchical Clustering:
3. Hierarchical clustering algorithms divide datasets in hierarchy clusters.
   It can be done by 2 approaches i.e.
   Top-down: It is complex due to absence of subroutine. It is more efficient and accurate

And Bottom-up.: It is not complex and less efficient with accurate.

- Massive datasets
- Expensive
- Effectiveness degradation in high dimension spaces due to dimensionality phenomenon.

4. DENSITY BASED CLUSTERING:

It is discovering the cluster which are arbitrary in shapes and the noise in a spatial database.

Clustering can be done by following certain algorithms. The algorithms are completely depended upon the following properties are:-

1. Type of attribute handled by algorithm
2. Complexity
3. Size of data and database
4. Detection of Outlier : Finding very similar object
5. Dependency in ordering tuples in database
6. Choosing initial cluster
7. Calculation of center.s

Clustering Benefits:

In case of Linux clustering server:

1. Completely a Scalable solution. Addition of resources are allowed.
2. If a cluster server needs any maintenance, it can be done by stopping it while handling the load over to other servers.
3. It is reliable and easy to configure.

In case of Oracle:

1. The value of clustered key is stored once which reduces memory.
2. Query retrieval is very fast.

Server Clustering:

1. If one server is having some problem and did not solved then it will take solution from cluster and don't affect to another server.
2. Resource addition is easy.

SQL server:

1. Only one instance run over the server.
2. Usage of hardware is full

VI. Limitations:

The potential problems with clustering are:-

1. Distance measure identification: For numerical attributes distance measurement is difficult. In case of Murkowski distance where maximum distance calculation is found by numerical attribute.
2. Number of clusters: If number of class labels are unknown then number of cluster finding is not easy.
3. Class label lacking
4. Database structure: In real life databases are not having sufficient structured tuples which lead to missing of data. If data cannot be structured properly then clarification and completeness will not occur.
5. Different attributes in Database: Database may not contain distinctively numerical or categorized attributes.

1. High cost: Though cluster needs proper hardware and a design which is costly compared to non – clustered server management.
2. Clustering needs more h/w and severs to establish one, monitoring and maintenance is hard which increases the infrastructure.

In case of oracle:

1. It takes longer to update records when the fields in the clustering index are changed.
2. Avoiding clustering index constructions, there is a risk that many insertions will happen on almost the same clustering index value

3. Possibility that cluster s/w will fail in another sub system that would not occurred in standalone operation.
4. Recovering from unanticipated s/w or h/w state is difficult.
5. While performing management tasks showing of operator error,

SQL server:

1. Only one instance run which make other servers idle. Full utilization of money will not done.
2. Any more Configuration needs more license.
3. If failover occurs and one server is running then the performance is adversely affected.

VII. K-MEANS ALGORITHM:

By referring [2] we knew that in case of k-means algorithm first we have to choose any no. of desired cluster, choose an initial point. Then we have to calculate classification, centroid classification, and convergence classification. This algorithm is basically used in case of huge number of variables.

Algorithm:

Step 1: Set K- Choose a number desired clusters,K

Step 2: Initialization – To choose k starting points which are used as initial estimates of the cluster centroids. They are taken as the initial starting values.
$\mu i=$ some value $,i=1,...,k$
Step 3: Classification – To examine each point in the dataset and assign it to the cluster whose centroid is nearest to it.
$ci=\{j:d(xj,\mu i)\leq d(xj,\mu l),l\neq i,j=1,...,n\}$
K-Means:

Strength

-Computationally faster execution occur in case of huge number of variables.

-K-means produce tighter clusters, especially when clusters are globular

Step 4: Centroid calculation – When each point in the data set is assigned to a cluster, it is needed to recalculate the new k centroids.
$\mu i=1|ci|\sum j\in cixj,\forall i$

Step 5: Convergence criteria – The steps of (iii) and (iv) require to be repeated until no point changes its cluster assignment or until the centroids no longer move [7]

-Time complexity is less. Faster computation

Weakness

-For predicting K-value is difficult.

-It does not work well in case of global cluster

-Different initial partition can result in different final clusters.

-In case of different size and density of clusters, it does not work well.

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{\frac{2}{m-1}}}$$

VIII. FCM (FUZZY-C-MEANS) ALGORITHM:

It is one kind of method of clustering which allows a single piece of data to belonging more clusters using recognition of patterns. On the basis of minimizing following objective function are :

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m}\|x_i - c_j\|^2 \quad , \quad 1 \le m < \infty$$

where m can be considered as any real number that must be greater than 1, u_{ij} is membership degree of x_i in cluster j, x_i is the ith of d-dimensional measured data, c_j is the d-dimension center of the cluster, and ||*|| is any kind of norming which is expressing the equity in between any data measured and center. Partition of Fuzzy is carried out by an iteration that

1. Initialize $U=[u_{ij}]$ matrix, $U^{(0)}$

2. At k-step: calculate the centers vectors $C^{(k)}=[c_j]$ with $U^{(k)}$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

3.Update $U^{(k)}$, $U^{(k+1)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{\frac{2}{m-1}}}$$

4. If $\| U^{(k+1)} - U^{(k)}\| < \varepsilon$ then STOP; otherwise return back to step 2.

by updating the cluster centers iteratively and also the membership grades for data point [8].

Minimize $$E(U,V) = \sum_{i=1}^{k}\sum_{j=1}^{n}(u_{ij})^m \|\bar{x}_j - \bar{v}_i\|^2$$

too optimized of the objective function shown above, with the update of membership u_{ij} and the cluster centers c_j by:

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}} ,$$

This iteration will stop when $\max_{ij}\left\{\left|u_{ij}^{(k+1)} - u_{ij}^{(k)}\right|\right\} < \varepsilon$, where ε is a termination criterion between 0 and 1, whereas k are the iteration steps. At local minimized point of Jm, the procedure converges.

Algorithm:

Subject to $$\sum_{i=1}^{k} u_{ij} = 1 \qquad \forall j = 1,\dots,n$$

Solving constrained optimization

- By Lagrangian multipliers

$$L_j(U,V) = \sum_{i=1}^{k}\sum_{j=1}^{n}(u_{ij})^m \|\bar{x}_j - \bar{v}_i\|^2 + \alpha_j\left(\sum_{i=1}^{k} u_{ij} - 1\right)$$

FCM

Strength

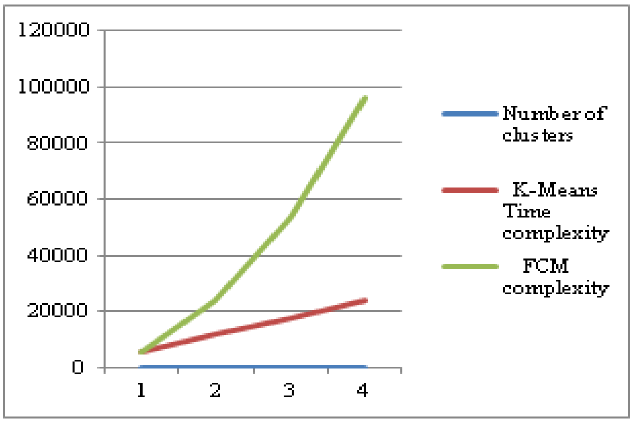
1 Gives best result for overlapped data set and comparatively better then k-means algorithm.
2 k-means at data point where exclusively belong to one center at one cluster, here data point is assigned membership to each cluster center as a result of which data point may belong to more than one cluster center.

Weakness

1 Priority specification of the number of clusters.
2 With lower value of β we get the better result but at the expense of more number of iteration.
3 Euclidean distance measures can unequally weight underlyingfactors.

IX. COMPARISON IN BETWEEN K-MEANS AND FCM ALGORITHM:

The time complexity of K-means is O(ncdi) and time complexity of FCM [3] is O(ndc2i)



Number of Clusters	K-Means Time Complexity	FCM Time Complexity
1	6000	6000
2	12000	24000
3	18000	54000
4	24000	96000

[4]

Fig 5 TC of K-Means and FCM by varying number of clusters [4]

After studying paper [4] we came to know that among both of the algorithms K-means is comparatively good in terms of different clusters. Time complexity of K-means algorithm is less than of FCM so more preferable is k-means.

Analysis after referring number of Papers:

K-Means Algorithm
-It was extremely faster than FCM in all datasets containing the clusters scattering in regular or irregular patterns[5]

- In case of image clustering it produces fairly higher accuracy and requires less computation. [6]

-By considering computation time it is best.[7]

- produces fairly higher accuracy and requires less computation[8]

- After apply normalization this clustering algorithm forms clusters with less time and more accuracy than other algorithms in bank document using WEKS tool.[9]

Fuzzy-C-Means Algorithm

-It is an algorithm based on more iterative fuzzy calculations, so its execution was found comparatively higher as it is expected[5]

-It produces nearly similar result but require more computations.[6]

-It should be improved in terms of its computation time.[7]

- more computation time is required than K-means because of the fuzzy measures calculations involved in the algorithm[8] -not accurate [9]

TIME COMPLEXITY OF K-MEANS AND FCM WHEN NUMBER OF CLUSTERS VARYING

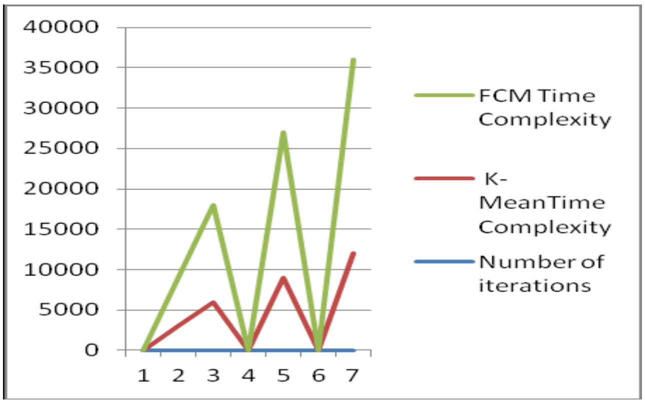


Fig 6 TC of K-Means and FCM by varying number of iterations [4]

CONCLUSION

In this paper we have got an overview about big data concepts, limitations, clustering, types of clustering and algorithms related to clustering. Finally we have studied partitioned clustering and its 2 main algorithms, at the same time comparison among FCM and K-means algorithm on the basis of several point out of papers. By bringing up several papers we came to know that both algorithms are having some demerits and merits mean while among two, FCM is having more demerits than K-means due to its computational algorithm and complexity in solving.

REFERENCES:

[1] A Survey on Clustering Techniques in Medical Diagnosis

[2] A Review: Clustering Techniques for Medical Image Segmentation. Juilee Anil Katkar, Trupti Baraskar Reseach Scholar, Department of Information Technology Assistant Professor, Department of Information Technology Maharashtra Institute of Technology, PuneUniversity Pune 411048

[3] A. Rui and J. M. C. Sousa, "Comparison of fuzzy clustering algorithms for Classification", International Symposium on Evolving Fuzzy Systems, 2006 , pp. 112-117.

[4] Comparative Analysis of K-Means and Fuzzy C-Means Algorithms Soumi Ghosh Department of Computer Science and Engineering

[5] Comparison of K-Means and Fuzzy C-Means Algorithms on Different Cluster Structures Zeynel Cebeci1, Figen Yildiz2

[6] Comparitive Analysis Of K Means And Fuzzy C Means Algorithm by Poonam fauzdar and Sujata Kindri
International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Vol. 2 Issue 6, June – 2013

[7] Performance Comparison of Fuzzy C Means with Respect to Other Clustering Algorithm Tejwant Singh1, Mr. Manish Mahajan2 1 Research Scholar, 2 Associate Professor, Dept of IT, Chandigarh Engineering College, Mohali, India

[8] " A Comparative Analysis of Fuzzy C-Means Clustering and K Means Clustering Algorithms" Mrs. Bharati R.Jipkate and Dr. Mrs. V. V. Gohokar

[9] A Comparative Analysis of Clustering Algorithms Raj bala Research Scholar (M.Tech) Amity University Haryana, India
Sunil Sikka, PhD.

Author Profile

Swayanshu Shanti Pragnya is currently in 4th year of computer science engineering at Centurion University of technology and management 2014-18. During 2016-17 researched about big data, big data issues, clustering algorithms, partitioned clustering algorithm and FCM. Published a paper on digital image processing in IJSRD journal 2016.