



Survey on mining clusters using new k-mean algorithm from structured and unstructured data

¹T.Nelson Gnanaraj ²Dr.K.Ramesh Kumar ³N.Monica

¹ M.Tech Scholar, Department of IT, Hindustan University, Padur, Chennai.

Nelsonraj.27@gmail.com

² Associate Professor, Department of IT, Hindustan University, Padur, Chennai.

rameshkumark.dr@gmail.com

³ M.Tech Scholar, Department of IT, Hindustan University, Padur, Chennai.

monitech26@gmail.com

ABSTRACT

Big data is the popular term used in the current era for extracting knowledge from large datasets. Bigdata is the collection of large and complex dataset. The challenge in big data is volume, variety and velocity (3V's). variety can be classified into structured, unstructured data and semi structured. Structured data are the identifiable data, which is organized in some structure. Data stored in the relational database are example of structured data. Unstructured data are the data without identifiable structure, audio, video and images are few examples. All web and bioinformatics data comes under semi structure data which does not have any regular structure, it is neither structured nor semi structured. Clustering the one of the best technique in knowledge extraction process. It is nothing but grouping of similar data to form a clusters. The distance between the data in one clusters and other should not be less. Many algorithms are practiced for clustering, in that k-mean clustering is the one of the popular term for cluster analysis. The main aim of the algorithm is to partition the dataset into k clusters based on some computational value. The limitation of k-mean clustering is that it can be applied to either structured or unstructured, not in combination of both. This project overcomes that limitation by proposing new k-mean algorithm for extracting hidden knowledge by forming clusters from the combination of both structure and unstructured dataset.

Keywords: Big Data, structured, unstructured data, clustering, k-mean

1. INTRODUCTION

The Objective of this literature survey is to provide brief review on clustering techniques and different methods associated with it. It also discuss briefly about the concepts of Big data and different variety of data involved in the clustering technique.

1.1 DATA MINING

Data Mining is the process of knowledge extraction. The ultimate goal of data mining is to mine knowledge or information from the dataset and makes it into understandable form for the future use. Data mining is the process of finding correlations or patterns among dozens of field in large relational databases^[1]. Many Companies follows data mining in order get the consumer focus. Data Mining is used in many fields such as retail, medical, organization, etc. There are many methods associated in knowledge extraction. They are outlier detection, clustering, classification, regression, summarization, sequential pattern mining and Association rule learning^[2]. This Paper discuss clearly about the clustering techniques.

1.2 CLUSTERING

Clustering is the one of the popular method in data mining; it is the process of grouping the information in the dataset based on some similarities. In definition it can be described as "It is the task of grouping a set of objects in the same group (clusters) are more similar to each other than to those in other groups (clusters)^[3]". Clustering the unsupervised learning process, i.e. there is no predefined classes. Data modeling puts clusters in a historical perspective rooted in mathematics, statistics, and numerical analysis^[4]. From a machine learning perspective clusters correspond to hidden pattern, the search for clusters is unsupervised learning and the resulting system represents a data concepts. Data mining ads to clustering the complication of very large datasets, so this imposes unique computational requirements on relevant clustering algorithms^[4]. Each clusters consists of objects in the other clusters. Formally the clusters structure is represented as a set of subset $C = C_1, \dots, C_k$ of S , such that $S = \bigcup_{i=1}^k C_i$ and $C_i \cap C_j = \emptyset$ for $i \neq j$. Consequently, any instances in S belongs to exactly one and only one subset.

2. CLUSTERING METHODS

The clustering technique in data mining involves different methods in extracting the hidden knowledge. Many clustering methods have been developed but each uses different principle. Failey and Raftery(1998) suggest dividing the clustering methods into two groups :Hierarchical and partitioning methods. Han and Kamber (2001) suggest categorizing the methods into additional categories: Density based clustering, distribution based clustering, model based clustering, and grid based clustering^[5].

2.1 HIERARCHICAL CLUSTERING

It is also called as connectivity based clustering. In this method, the clusters are formed based on some hierarchy (top-down or bottom-up).Hierarchical clustering is based on the idea of objects being more related to nearby objects farther away. It is divided into two types.

2.1.1 AGGLOMERATIVE

It is the bottom up approach that each objects in the dataset represents a cluster of its own and then clusters are successively added until the desired cluster is obtained^{[5][6]}.

2.1.2 DIVISE

It follows the top-down approach. All the instance of dataset belongs to single cluster initially, and then they are subdivided into many clusters .this process continues until desired cluster is obtained.

The clusters are represented in dendrogram. It is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering^[7].

2.2 PARTITIONED CLUSTERING

In partitioned clustering, the objects in dataset are relocated by moving from one cluster to other based on some computational value. It is also known as the centroid based clustering method. In this method, the number of cluster should be predefined by the user. Namely a relocating method iteratively relocates points between the k-clusters. K-mean algorithm is the most popular algorithm in partitioned clustering.

2.3 DISTRIBUTION BASED CLUSTERING

Distribution based clustering provides fast and natural clustering of very large databases. It

automatically determines the number of clusters to be generated^[10].It is an iterative process.

The similarity of each object with each of the currently existing clusters is calculated .Initially no clusters exist^[10];if the similarity calculated is greater than the given threshold value. The object is placed in the relevant cluster otherwise new cluster is created.

2.4 DENSITY BASED CLUSTERING

Density based algorithm apply a local cluster criterion; clusters are regarded as regions in the data space in which the objects are dense, end which are separated by regions of low object density^[12].

DBSCAN is the method used in density based clustering, In contrast to many methods, it features a well –defined clusters model called “density reach ability”^[3].It is similar like linkage clustering method only difference is ,it is based on the connection points within the certain threshold^[3].

2.5 MODEL BASED CLUSTERING

Model based clustering helps to optimize the fit between the given data and some mathematical model. It also describes the characteristics of each cluster. Where each group represents the class. Model based clustering is of two types^[5], they are

- Decision trees
- Neural networks

2.5.1 DECISION TREES

In decision trees the leaf denotes the concepts and contains a probalistic description of that concept. Many algorithms are developed for unlabelled data. The well known algorithm is COBWEB.

2.5.2 NEURAL NETWORKS

This type of algorithm represents each cluster by a neuron or “prototype”. The input data is also represented by neurons, which are connected to the prototype neurons.SOM algorithm is used for neural network clustering.

2.6 GRID BASED CLUSTERING

Clustering operation is performed on the grid structure (space partitioned into finite number of cells).The main advantages of this clustering are its fast processing time^[5].

3. K-MEAN CLUSTERING

The fast simple and most popular partitioned clustering method is k-mean, developed by mac Queen in 1967. This method considers the mean value of the objects in the dataset, to form a cluster. It aim to partition n Observation into k-cluster in which each observation belongs to the cluster with the nearest mean^[13]. k-mean algorithm steps are as follows^[14]

3.1 STEPS IN K-MEAN ALGORITHM

- Place k points into the space represented by the objects that are being clustered these points represent initial group centroids
- Assign each objects to the group that has the closest centroid
- When all objects have been assigned recalculate the position of the k-centroid
- Repeat step 2 and 3 until the centroid no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

```

Input: D=Dataset
          K= The Number of centers
Output: Set of k centroid  $c \in C$ 
           representing a good partitioning of D
           into k clusters
1  Select the initial cluster
   centroids c
2  repeat
3    Changed=0
   //Find the closest centroid to
   every data point d...
4    for all data point  $d_i \in D$  do
5      assignedCenter =  $d_i$ .center
6      for all center  $c_j \in C$  do
7        Compute the squared
   Euclidean distance  $dist = dist$ 
   ( $d_i, c_j$ )
8      if  $dist < d_i$ .center Distance
   then
9         $d_i$ .center Distance =  $dist$ 
10        $d_i$ .center = j
11      end if
12     end for
13     if  $d_i$ .center  $\triangleleft$ 
   assignedCenter then
14       changed ++
15       Recompute  $c_j$ .new for
   next iteration
16     end if
17   end for
18 Until changed == 0.
    
```

Figure 1: k-mean Algorithm

The error sum of square is calculated by

$$E = \sum_{i=1}^k \sum_{p \in C_i} \text{dist}(p, C_i)^2 \quad (3)$$

Where,

- E = sum of square error
- K = number of clusters
- P = An object
- $C_i = i^{\text{th}}$ cluster
- C_i = the centroid of cluster i

The goal of k-mean algorithm is to produce the solution such that there are no other solutions with lower SSE. The advantages of k-mean is it works with a large number of variables faster than the hierarchical clustering .then it produces tighter clusters than the hierarchical clusters especially if the clusters are globular^[16]. k-mean algorithm also contain some disadvantages .that is they doesn't work well with non globular clusters ,different initial partition can result in different final clusters^[16].

4. BIG DATA

Big data is the word describing the large volume of both structured and unstructured data, which cannot be analyzed using traditional techniques and algorithm. It requires the unique algorithm, technique and analysis process.

As of 2012, limits on the size of data sets that are feasible to process in a reasonable amount of time were on the order of Exabyte's of data.^{[17][18]} Scientists regularly encounter limitations due to large data sets in many areas, including meteorology, genomics^[19] connectomics, complex physics simulations^[20] and biological and environmental research.^[21] The limitations also affect Internet search, finance and business informatics. Data sets grow in size in part because they are increasingly being gathered by ubiquitous information-sensing mobile devices, aerial sensory technologies (remote sensing), software logs, cameras, microphones, radio-frequency identification readers, and wireless sensor networks.^{[22][23]} The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s;^[15] as of 2012, every day 2.5 quintillion (2.5×10¹⁸) bytes of data were created. The challenge for large enterprises is determining who should own big data initiatives that straddle the entire organization. 90 percent of data are created in last two years alone.

Big data mainly characterised by three V's namely: volume variety and velocity. There are other 2V's available such that veracity and value. But scientists are mostly concentrating on only 3V's. fig:2 shows the three characteristics of Bigdata.

4.1 VOLUME

Volume denotes the size .every data more than 50 million terabytes of data are produced from different field such as social network, email, sensor networks etc. those collected large dataset should be managed properly.

4.2 VARIETY

Variety is nothing but different types of data .every day different varieties of data are produced in every field .they have to be managed in the sense that they produce some knowledge.

4.3 VELOCITY

Velocity denotes speed of the data. Processing the data stored in different dataset are time consuming, they have to be managed properly in way that they process the request in short span of time. In fraud detection velocity is the main aspect

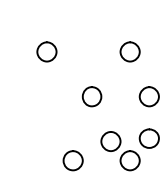
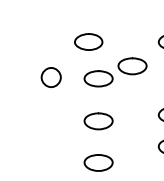
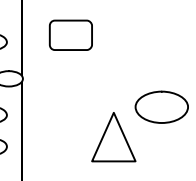
Volume	Velocity	Variety
 <p>Data at rest</p>	 <p>Data in Motion</p>	 <p>Data in many forms</p>

Figure 2: Big data 3v's

5. VARIETY

Big data consist of three different types of data. These data are combined to form large dataset. The types are classified as

- Structured
- Unstructured
- Semi structured

5.1 STRUCTURED DATA

Structured data are identifiable data, that they have some predefined structured .numerical data, program code are comes under structured dataset. Since they are stored in relational databases in form of rows and columns. Structured dataset are available in many fields such as organization data,

5.2 UNSTRUCTURED DATA

Unstructured data are data without identifiable structure. Audio, images and video files are unstructured data 90% of big data comprises of unstructured dataset

5.3 SEMI STRUCTURED DATA

Semi structured data are the combination of both structured and unstructured data.web and bio informatics data are semi structured data. For example in web front end contains the unstructured data such as images, tweets, etc and at back end it contains the structured data that stored in databases.

6. RESEARCH DIRECTION

Clustering is the key for big data problems, because large dataset are difficult to label. It evolves for more time and also it provides efficient browsing and search of the data in the dataset. Each and every application existing nowadays are using clustering .For example facebook uses clustering technique in order group the users and Google news used to group the similar data on search. Likewise youtube, Google search engines are also using clustering. It can also used in efficient Image retrieval. Many algorithm are proposed for mining clusters in large databases, some of them are CURE, BIRCH, RAD, NPUST.

Traditional clustering algorithms extract knowledge from spherical shapes and similar sizes. So in order to cluster the non -spherical clusters and varies sizes new algorithm called CURE is proposed .CURE algorithm achieve this by representing each clusters by a certain fixed number of points that are generated by selecting well scattered points from the clusters by a certain fixed number of points that are generated. In this algorithm in order to handle the large databases two steps are followed ,first the sample of data is partitioned and partially clustered, and then they are again clustered in the second step in order to get desired result.

There are generally two types of attributes involved in data clustering, metric and non metric. The BIRCH algorithm uses metric attributes, so that optimization problem exist in statistics literature can be avoided. In Additional, database-oriented Constraint: The amount of memory available is limited and we want to minimize the time required for I/O. So Birch algorithm is proposed in order to reduce the I/O cost its architecture also provides opportunity for parallelism and for interactive or dynamic performance tuning based on knowledge about the dataset, gained over the course of the execution. It is the first algorithm proposed to address outliers in databases and noise handling.

But the CURE algorithm technique differs from BIRCH in two ways. First, instead of pre clustering with all the data points, CURE begins by drawing a random sample from the database. Second in order to further speed up clustering process, CURE partially clusters the random sample.

CLARANS is the algorithm proposed for clustering the spatial databases, which contributed three main things that, first to identify the spatial structure in the data. Second is to handle the polygon objects efficiently and at last, two algorithms are built to discover the relationship between spatial and non-spatial attributes.

Another Algorithm for clustering large databases is the RAD, it is proposed in order to overcome the time consumption. When data sizes increases time to process is also increases simultaneously. So in order to perform clustering fastly and to attain the quality result RAD algorithm is developed.

NPUST, which is the advance version of KIDSCAN. It is a hybrid density-based approach, which partition the dataset using k-mean and then clusters the resulting partition with IDSCAN, finally the closest pairs of clusters are merged until the natural number of clusters of dataset is obtained.

The traditional clustering algorithm need multiple data scan to achieve the desired result, so it will not suitable for large databases since it will be more expensive. Scalable Clustering framework applicable to a wide class of iterative clustering, this method is based on identifying regions of the data that are compressible, regions that must be maintained memory, and regions of the data are discardable.

DENCLUE (DENSity based CLUstEring) is the algorithm proposed for clustering large multimedia databases. The basic idea of this algorithm is to model the overall point density analytically as the sum of influence functions of the data points. Clusters can then be identified by determining density -attractors. Then main advantage of this approach is it has good clustering properties in dataset with large amount of noise and it allows a compact mathematical description of arbitrarily shaped clusters in high -dimensional datasets.

7. CONCLUSION

This survey briefly review the clustering technique and its different methods .It also described the concept of k- mean algorithm which is used to find the clusters from structured and unstructured data set..the Research Direction clearly explain that the algorithm proposed are applied to either structured or unstructured data. So this survey will be helpful for performing the clustering technique in

combination of variety of data using k-mean algorithm.

Thus we have given an overall coverage on Clustering which provide shortly, a outline of the recent work which gives general view of the field. Clustering has achieved tremendous progress which gives the various set of applications. This survey is prepared to our new project titled Mining cluster using new k-new algorithm from variety of data.

REFERENCES

1. Xiaowei Xu, Martin Ester, Hans-Peter,Kriegel, Jörg Sander,University of Munich,Oettingenstr. 67,D-80538 München”A Distribution-Based Clustering Algorithm for Mining in Large Spatial Databases”.
2. ,Hans-Peter Kriegel Martin Pfeifle,University of Munich, Germany University of Munich,” Density-Based Clustering of Uncertain Data”
3. Francis, Matthew (2012-04-02). "Future telescope array drives development of exabyte processing". Retrieved 2012-10-24.
4. "Community cleverness required". *Nature* **455** (7209): 1. 4 September 2008. doi:10.1038/455001a.
5. "Sandia sees data management challenges spiral".HPC Projects. 4 August 2009.
6. Reichman, O.J.; Jones, M.B.; Schildhauer, M.P. (2011). "Challenges and Opportunities of Open Data in Ecology".*Science* **331** (6018): 703–5.doi:10.1126/science.1197962.
7. Hellerstein, Joe (9 November 2008). "Parallel Programming in the Age of Big Data". Gigaom Blog.
8. Segaran, Toby; Hammerbacher, Jeff (2009). *Beautiful Data: The Stories Behind Elegant Data Solutions*. O'Reilly Media. p. 257. ISBN 978-0-596-15711-1.
9. Hilbert & López 2011
10. "IBM What is big data? — Bringing big data to the enterprise". www.ibm.com. Retrieved 2013-08-26.
11. Oracle and FSN, "Mastering Big Data: CFO Strategies to Transform Insight into Opportunity", December 2012
12. Shan Suthaharan,Department of Computer Science,University of North Carolina at Greensboro,Greensboro, NC 27402, USA,s_suthah@uncg.edu”Big Data Classification: Problems and Challenges in Network Intrusion Prediction with Machine Learning,”
13. Navneet Gopal,Poonam Goyal,K Venkatramaiah,Deepak P C,and sandoop P S,Department of computer science and information systems,BITS,pilani,”An Efficient Density Based Increment clustering Algorithm in

- Data Warehousing Environment “International conference on computer Engineering and Application 2009,Singapore.
14. Zengyou He,Xiaofei Xu,ShengChun Deng,Department of computer Science and Engineering ,Harbin Institute of Technology,china”A link clustering based approach for clustering categorical Data”this work supported by High technology Research and development of china.
 15. Zengyou He,Xiaofei Xu,ShengChun Deng,Department of computer Science and Engineering ,Harbin Institute of Technology,china”Clustering Mixed Numeric and categorical Data:A Cluster Ensemble Approach”
 16. Zhexue Huang,CSIRO Mathematical and Information sciences,Australia”clustering Large datasets with mixed Numeric And Categorical Values”
 17. Elham Karoussi,Associate professor Nouredine Bouhmala”Data Mining K-clustering Problem”
 18. David Loshin ,”Integrating structured and unstructured Data”TDWI Research.
 19. Qiuyue Wang,School of Information ,Renmin University of china, ,Jinglin Kang,Key lab of Data Engineering and Knowledge Engineering,MOE,China”Integrated Retrieval Over Structured and Unstructured Data”.
 20. Imran R.Mansuri,Sunita Sarawagi,IIT BomBay”Integrating unstructured data into relational databases”.
 21. Tapas Kanungo,Senior Member,IEEE,David M.Mount,Member,IEEE, Nathan S.Netanyahu,Member,IEEE,Christine D.Piatko,Ruth Silverman,and Angela Y.Wu,Senior Member,IEEE “ An Efficient k-means clustering Algorithm : Analysis And Implementation”
 22. Chen et al,CLUE : Cluster based retrieval of images by unsupervised learning,"IEEE Tans. On Image Processing ,2005.
 23. Sudipto Guha¹,Rajeev Rastogi², and Kyuseok Shim³.CURE:An Efficient Clustering Algorithm For Large Databases, published in Information Systems,vol 26,No.1,35- 58,2001.