World Academics Journal of Engineering Sciences **Research Paper**
Vol.5, Issue.1, pp.01-06, December (2018) **E-ISSN:** 2348-635X

An Efficient Cluster Formation Approach for Handling Large Scale Data

Anil[1], Rajendra Gupta[2*]

[1,2]Rabindranath Tagore University, Raisen, India

Available online at: www.isroset.org

Received: 29/Jul/2018, Accepted: 26/Aug/2018, Online: 31/Dec/2018

Abstract- The basic K-means algorithm is the most widely used clustering algorithm and the best-known of the partitioning-based clustering methods. Choosing the initial centroids randomly results in poor clustering. Instead of optimal clustering (or global optimum), sub-optimal clustering (or local minimum) is obtained. One of the problems associated with the basic K-means algorithm is that empty clusters can be obtained if no points are allocated to a cluster during the assignment step. So implementation in basic k-mean algorithm gives better solution for large scale data analysis. In this paper, the k-mean algorithm and its implementation have done and the obtained results are also discussed. After the analysis, it is found that the implementation in k-mean algorithm provide better outcome.

Keywords- Basic k-mean, implemented k-mean, Large Scale Data, Fuzzy C4.5

I. INTRODUCTION

The data may have any facts, any figures, numbers or text which can be processed by the computer system. Now-a-days, organizations are accumulating enormous and growing amounts of data in different formats and different databases. The data can be treated as given below:

- Transactional or operational data such as, cost, payroll, sales, inventory, payroll or accounting
- Non-operational data which include forecast data, industry sales and macro-economic data
- Meta Data – the data about the data itself, which can include logical database design or data dictionary definitions

1.1 Challenges with large scale data

There are so many challenges observed in developing Business Intelligence with unstructured or large scale data data.

- **Terminology** – the terminology is not clear in various disciplines. Among researchers and analysts, there is a need to develop a standardized terminology.
- **Accessing unstructured textual data** – the unstructured data is stored and saved in a enormous variety of formats.
- **Volume of data** – As it is clear that up to 75 percentage of all the data exists as unstructured data. The need for word-to-word and semantic analysis is required.
- **Searching capacity of unstructured textual data** – A simple search on some data, eg. “mango”, results in links where there is a reference to that precise search term. The author Inmon et.al. 2018, gives an example: “a search is made on the term felony. In a simple search, the term felony is used, and everywhere there is a reference to felony, a hit to an unstructured document is made. But a simple search is crude. It doesn't find references to crime, arson, embezzlement, murder, and vehicular homicide even though these crimes are types of felonies.”

1.2. Detailed Survey and Comparative Analysis

Following is the depiction of earlier proposed algorithms, its scalability, efficiency, the shape of cluster and the input data to study and anlysing the large scale data :

Table 1 : Comparison chart for algorithms for large scale data

Algorithm	Scalability and Efficiency	Noise	Shape of cluster	Input data
K-Means	Scalable in processing large datasets.	Sensitive to noise and outliers.	Works well only with clusters of convex shapes.	Works only on numerical data.
PAM	Works well for small datasets but not for large datasets.	Not very sensitive to noise and outliers.	-	Works on data of all attributes.
CLARA	Can deal with larger datasets in comparison to PAM	Not very sensitive to noise and outliers.	-	Works on data of all attributes.

	efficiency depends on sample.			
CHAMELEON	-	-	Good at finding clusters of arbitrary shape.	Works on data of all attributes.
DBSCAN	Does not work well for high dimensional data.	Handles noise effectively.	Good at finding clusters of arbitrary shape.	-

1.3. Issues related to Large Scale Data

The following issues have been observed during the study of data mining algorithms :

- Scalability
- Data Quality
- Data Ownership and Distribution
- Dimensionality
- Privacy Preservation
- Streaming Data

1.4. Limitations and Objectives

The limitations of K-means algorithm while large data is as given below :

- Choosing the initial centroids randomly results in poor clustering.
- Instead of optimal clustering (global optimum), suboptimal clustering (local minimum) is obtained.
- Empty clusters are formed if no points are allocated to a cluster during the assignment step (because of which squared error will be larger than necessary, wastage of memory).

The objective of the research study is to

- Analysing the mining algorithms for clustering
- Clustering the Large scale datasets
- Compare the performance of proposed algorithm with earlier algorithm

II. LITERATURE REVIEW

The frameworks have been designed to Extract Context Vectors from Unstructured Data using Large scale data Analytics for analyzing large scale datasets. The paper proposes a framework for computing context vectors of large dimensions over large scale data, trying to overcome the bottleneck of traditional systems. The aim of the paper is to examine and propose a framework for computing context vectors of large dimensions over large scale data, trying to overcome the bottleneck of traditional systems. This paper explores the Hadoop cluster on Amazon Elastic Cloud, perform the benchmark of data load time with traditional data processing application and Hadoop. Secondly, they analyzed the unstructured data in Zeppelin with Spark [1].

The Extensible Query Framework for Unstructured Medical Data having a large scale data Approach shows the extensible query based framework contains built-in modules but is flexible in allowing the user to import their own, making it extensible. The framework runs the modules in a Hadoop cluster making it efficient by utilizing the distributed computing capability of large scale data approach. The framework is tested through simulation[2,3]. The framework allowed the user to run a different module using the previous output to further analyze the unstructured data it also enabled the user to import a new module.

The Agglomerative Algorithm to discover Semantics from Unstructured large scale data paper presents a graph model and an agglomerative algorithm for text document clustering. Author had tested. The algorithm in three different data sets and presented working scenario and also compared with traditional clustering algorithms, such as k-means, principal direction division partitioning, Auto-Class and hierarchical clustering.

The review of Large scale Architecture for Voice and Data Services in Mobile Communication paper gives a detail review of hybrid routing protocols for mobile communication[4,5]. With the considering Proactive routing protocol and Reactive routing protocol discoveries the routing tables only for the destination that has traffic going through. The common problem associated with network is mobility management. To overcome this problem author proposed a model for higher degree of coverage with less traffic. Future work of this paper focuses on the reduced delay with increased packet delivery ratio and better control over paths.

The large scale data clustering validity paper describes a new fuzzy validity index able to interpret the best partition of Large scale data clustering. Called Fuzzy Validity Index with Noise-Overlap Separation (FVINOS), this new technique provides sufficient interpretation of the properties of the large scale data by detecting the overall geometric structure within and between clusters[6].

The clustering time-stamped data using multiple nonnegative matrices factorization[7-9]. In this paper, an approach for clustering time-stamped data and discovering the evolutionary trends of the clusters by using Multiple Nonnegative Matrices Factorization (MNMF) with smooth constraint over time. To utilize time-stamped information in the clustering process, an extra object-time matrix is constructed in our proposed method. Experimental results on real data sets demonstrate that the proposed approach outperforms the comparative algorithms with respect to Fscore, NMI or Entropy[10].

III. BASIC K-MEANS ALGORITHM

Basically K-Means is an iterative process that divides a given data set into *K* disjoint groups. The K-means is the most widely used clustering principle, and the best-known of the partitioning-based clustering methods.
The basic K-means Algorithm is formally described as follows -

- Select K points as initial centroids.
- Repeat it
- form K clusters by assigning each point to its closest centroid
- Re-compute the centroid of each cluster
- until Centroids do not change

The major operations of K-means is very well illustrated in the algorithm, which shows how, starting from three centroids, the final clusters are found in four assignment-update steps. In this concept, each area shows :
(1) the centroids at the start of the iteration and
(2) the assignment of the points to those centroids. The centroids are indicated by the "+" symbol. All points belonging to the same cluster have the same marker shape

In the first step, the points are assigned to the initial centroids, that are all in the larger group of points. In this example, the mean as the centroid is treated. After points are assigned to a centroid, the centroid is then updated. Again, the figure for each step shows the centroid at the beginning of the step and the assignment of points to those centroids. In the second step, points are assigned to the updated centroids, and the centroids are updated again.

IV. PROPOSED METHOD

4.1 Implementation in K-means

According to the survey, every hour more than 35,000 status or data are updating in server and a huge number of data is being collected, that is called large scale data. The revolution in scientific and technological facet has affected the size of data which increases on a daily basis with an aim to improve profitable activities. In present days, there is a growth in Technology, Business that enlarges the data at a faster date.

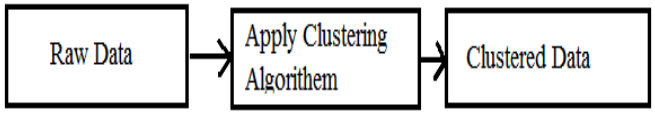

Figure 1 : General representation of large data

All the data from different sources are collected at one point these large set of data are stored in data center, receiving the huge data at an instance is difficult task. For an example, a system may include data of weather conditions, population of particular country, traffic information, organization information, when a person need to access any one above mentioned data which is difficult to get at instance. Therefore to make search easier data is divided into clusters and stored into data center. The clustering helps in reducing redundancy and fault tolerance by making sure that there is no single point of failure. The clustering deals with searching a *pattern* in a collection of un-labeled data.

A *cluster* is a collection of objects which are "similar" between them and are 'dissimilar' to the objects belonging to other clusters". The cluster analysis groups data objects based only on information found in the data that describes the objects and their relationships. The goal is that the objects within a group be similar (or related) to one another and different from (or unrelated to) the objects in other groups. Better clusters are formed with greater the similarity (homogeneity) within a group and the greater the difference between groups. Basically clustering systems yield a data description in terms of clusters that possess strong internal similarities. Therefore the cluster is considered as a collection of objects, which are similar to one another within the same cluster and dissimilar to the objects in other clusters [11,13].

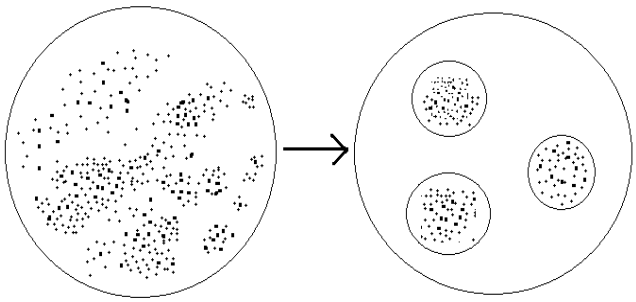

Figure 2 : Cluster formation example

To simplify the goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. In the proposed system, we are presenting an implemented K-means algorithm, that overcomes the above mentioned problems occurring because of choosing initial centroids and formation of empty clusters. This is achieved by intelligently allocating sufficient number of initial centroids, thus leading to better formation of clusters [12].

The proposed K-mean scheme also eliminates the empty clusters that are formed in the process of clustering.

Step 1: Given a huge unstructured data set

Step 2: Assign K random points as centroids in unstructured data.

Step 3: For all C_i to $C_k$whereC_i: i^{th} Centroid

If distance $\{C_i, C_{i+1}\}$<y …… (1)
//Compare the distance between two centroids[2]
// where y is a constantsuch that y >= D
Remove C_i
//If two centroids are likely to fall in one cluster then remove one centroid

Step 4: If $\sum C_i < K$
Assign X centroids such that $X + C_i = K$.

Step 5: Repeat Step 3 and 4 until K centroids are formed and no two centroids are within y.

Step 6: Assignment step: Allocate the data points to their closest centroids (closeness is determined based on Euclidean distance)

Step 7: K clusters are formed
Step 8: If any of these clusters is empty (the empty cluster are those for which no data points are assigned)

For the each value of C_i to C_k

$$D=\sqrt{(C_e - C_i)^2 + (C_e - C_j)^2}$$

By using distance measure technique determine the nearest cluster centroid to this empty cluster centroid.

Step 9: Merge the empty cluster with closest centroid.

Step 10: Repeat Steps 6 and 7 until all empty clusters are eliminated.

4.2 Implementation process of K-mean algorithm

4.2.1 Initialization

The First phase of algorithm is initialization phase were the given pair of words are searched into huge unsupervised data set. We have taken websites which includes journal papers. The journal papers are accessed by different streams like research papers on Medical, Engineering, Agriculture papers etc.

A researcher enters a pair of words “Skin Diseases”, the word skin is classified under medical stream although in Image processing area skin color analysis, detection terms are also used. When a person enters these two key words the website will display publication papers of both Medical and Engineering. In initialization phase the second step is classification it searches each document with respect to given pair of words. To find the given key words in huge unsupervised dataset fuzzy C4.5 helps us to preprocess the attribute [13-15].

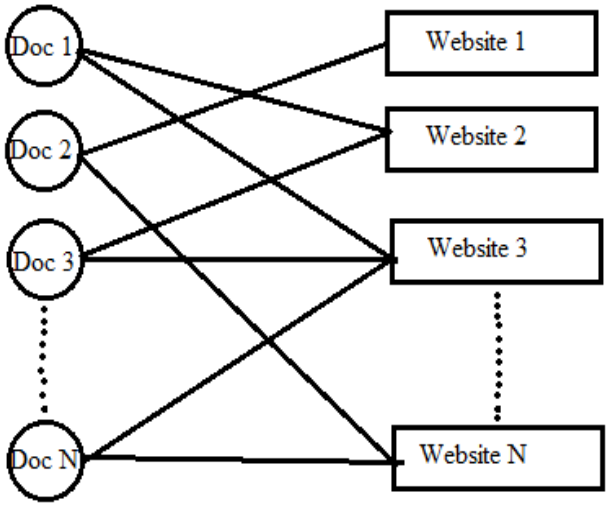

Figure 3 : Large scale database searching from various websites

4.2.2 Representation of Fuzzy C4.5

The each document is pre-processed is to reduce the number of attributes using Fuzzy C4.5 classification algorithm and converted into vectors of numerical values. If the vectors of numerical values are already available in the searching folder corresponding to the available documents, then preprocessing of the documents and converting into vectors need not be carried out using Text Pattern Mining using Radial Fuzzy C4.5 algorithm[16].

The algorithm steps are as given below :

- ✓ C4.5 is Ross Quinlan’s technique based on decision trees
- ✓ It is a refinement of the 1986 technique called ID3
- ✓ C4.5 is used in products such as Clementine
- ✓ Uses post pruning approach, but uses a specific technique to estimate predicted error rate
  - o Pessimistic pruning
- ✓ The algorithms generates decision rules
  - o E.g. If Attribute 1 = A and Attribute 2 = B, then Classification = Class C

4.2.3 Vector Space model

Widely used document representation is vector space (or "bag of words"). Each document d is represented by a vector (w/, .. , WMi), where w where w/ is the importance (weight) of term tj in document d . Weights can be determined by one of well established method of document classification - TFIDF, as follows:

$$d\ i = TF(ti)\ \&log(-v/.tJ),$$

where TF(tJ is term frequency - number of occurences of term ti in document d; D is number of documents and DF(til is document frequency - number of documents, where term ti is occured. Second argument of the product is also called

inverse document frequency. TF IDF method gives greater weight to the terms, which more frequently appear in more documents.

Many issues specific to documents are discussed more fully in information retrieval texts to represent our proposed algorithm the documents are represented using the vector-space model. In this model, each document, ***d***, is considered to be a vector, *d*, in the term-space (set of document "words"). Here a vector forms by determining the term weight by using Radical Basis Function.

$$di1 = (tf1, tf2, \ldots, tfn)$$

where oij is the occurrence of word i shown in class j, C is the total number of classes and M is the total number of different words. An element of a significance word vector for a word i in class j is represented as wij. Documents and queries are represented as vectors.

Each dimension corresponds to a separate term. If a term occurs in the document, its value in the vector is non-zero. Several different ways of computing these values, also known as (term) weights, have been developed. One of the best known schemes is tf-idf weighting.

The definition of *term* depends on the application. Typically terms are single words, keywords, or longer phrases. If words are chosen to be the terms, the dimensionality of the vector is the number of words in the vocabulary (the number of distinct words occurring in the corpus).
Vector operations can be used to compare documents with queries.

V. RESULT AND DISCUSSION

In this research work, we have conducted experiments on four large-scale datasets, which are depicted in Table 1.

- **RCV11**: This is a subset of an archive of 805414 categorized newswire stories from the source Reuters Ltd. This dataset have 233854 documents in 108 categories. Now remove the features appearing less than 100 times in the corpus that results in 1875 out of 35236 keywords.

- **CovType2**: It consists of 521052 instances for predicting forest data from cartographic variables.

- **ILSVRC20123**: This is a subset of ImageNet which contains 1100 object categories and around 1.4 million images. We have used the 5092-dimensional features extracted by the convolution neural networks (CNN) model to represent the images.

- **MNIST8M4**: This consists of around 5.4 million images of handwritten digits from 0-9.

Table 2 : Large scale datasets and its sample size

Datasets	Sample size	Features	Classes
RCV1	805414	12	108
CovType	521052	16	152
ILSVRC2012	1100	20	23
MNIST8M	540000	20	112

Table 3 : Comparison of k-mean and implemented k-mean for large datasets

Dataset	k-mean		Implemented k-mean	
	Time (in second)	Speed	Time (in second)	Speed
RCV1	154	1x	141	1x
CovType	18	1x	17	1x
ILSVRC2012	286	1x	188	1x
MNIST8M	488	1x	320	1x

The following figure compares the execution times of the k-mean algorithm and implemented k-mean algorithm for the four datasets. As the number of datasets to be clustered increases, the time dedicated to basic K-means exceeds. The proposed implemented algorithm takes less time to employ than the basic K-mean algorithm because of its processing of large datasets through several different systems. The implemented algorithm uses the Euclidean distance as a similarity measure and the proposed work also use inter and intra-clustering measures to obtain better and high-quality clusters with high levels of intra-cluster similarity and thus low levels of inter-cluster from large datasets.

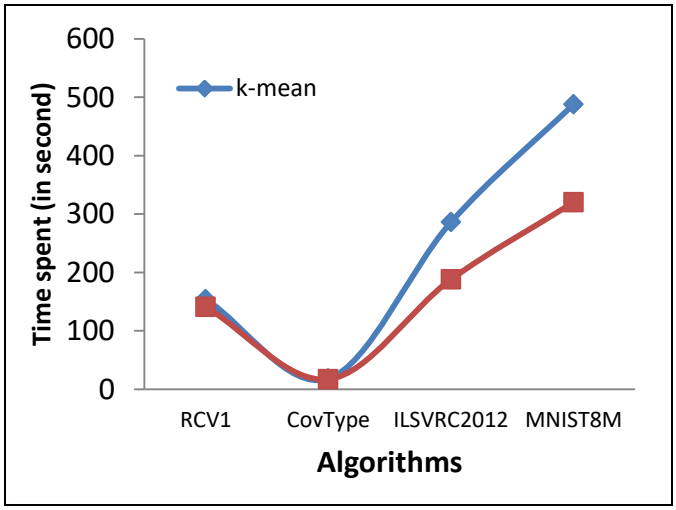

Figure 4 : Comparison of k-mean and implemented k-mean for large datasets

VI. CONCLUSION

Generally handling huge data from different source is a very big task. In every second, more than 100 types of data is collected and the managing and analyzing it is another

typical task. Since, there are different techniques to handle datasets so many researchers are showing keen interest to large scale data to real environment. In this research work, an implemented K-means algorithm is presented that can handle big problems associated with choosing initial centroids and also formation of empty clusters. The algorithm performs better with respect to time utilization and efficient cluster formation.

The basic K-means method is a very popular clustering approach due to its clarity, simplicity and reasonable execution time efficiency when applied on small datasets. A clustering method however, maintain cluster efficiency when large datasets are also involved by considering inter and intra-clustering distances between data objects in a dataset.

Since, the clustering is a challenging issue that is shaped by data used and the problems associated with it. The proposed algorithms show improvements in terms of its execution time. The main objective of this research work is to accelerate and scale-up large datasets to obtain prominent high-quality clusters.

REFERENCES

[1] Bhagyashri S. Gandhi ,Leena A. Deshpande , "The Survey on Approaches to Efficient Clustering and Classification Analysis of Large scale data", *International Journal of Engineering Trends and Technology (IJETT)* – Volume 36 Number 1- June 2016

[2] V. Estivill-Castro, Why so many clustering algorithms: A position paper, SIGKDD Explorations Newsletter, 4 (2002), pp. 65–75

[3] Sami Ayr ¨ am ¨ o Tommi K ¨ arkk ¨ ainen ¨," Introduction to partitioning-based clustering methods with a robust example" *University of Jyvaskyl ¨ a¨ ISBN 951-39-2467-X*, ISSN 1456-4378, 2006

[4] R. Duda and P. Hart, Pattern Classification And Scene Analysis, *John Wiley & Sons*, Inc., Ny, 1973.

[5] V. Estivill-Castro, Why So Many Clustering Algorithms: A Position Paper, *Sigkdd Explorations Newsletter*, 4 (2002), Pp. 65–75.

[6] A. K. Jain and R. C. Dubes, Algorithms For Clustering Data, Prentice-Hall, Inc., Upper Saddle River, Nj, Usa, 1988.

[7] M. S. Aldenderfer and R. K. Blashfield, *Cluster Analysis, Sage Publications, London, England*, 2016.

[8] Tanvir Ahmad, Rafeeq Ahmad, Sarah Masud, FarheenNilofer, "Framework to Extract Context Vectors from Unstructured Data using Large scale data Analytics", Pages: 1 -6, *Ninth International Conference on Contemporary Computing (IC3)* ,2016 IEEE

[9] Harleen,Naveen Garg, "Analysis of Hadoop Performance And Unstructured Data Using Zeppelin", Year: 2016, Pages: 1 - 6, *International Conference on Research Advances in Integrated Navigation Systems*, April 06-07, 2016 IEEE

[10] Radhika K R, Pushpa C N, Thriveni J, Venugopal K R, "EDSC: Efficient Document Subspace Clustering Technique for High-Dimensional Data", *International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT),2016 IEEE.*2016, 222

[11] Joe Tekli, "An Overview on XML Semantic Disambiguation from Unstructured Text to Semi-Structured Data: Background, Applications, and Ongoing Challenges", *IEEE Transactions on Knowledge and Data Engineering* Year: 2016, Volume: 28, Issue: 6Pages: 1383 -1407,

[12] SarmadIstephan, Mohammad-Reza Siadat, "Extensible Query Framework for Unstructured Medical Data – A Large scale data Approach", *IEEE International Conference on Data Mining Workshop (ICDMW)*Year: 2015, Pages: 455 - 462, DOI: 10.1109/ICDMW.2015.67, 2015

[13] I-Jen Chiang, "Agglomerative Algorithm to Discover Semantics From Unstructured Large scale data" , *IEEE International Conference on Large scale data (Large scale data)*Year: 2015 Pages: 1556 -1563, 2015

[14] Aarti Rahul Salunke, ArunNathaGaikwad, "Review of Unstructured Architecture for Voice and Data Services in Mobile Communication", Year: 2015,Pages 1-7, DOI: 10.1109/ICECCT.2015.7226190,2015 *IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT).*

[15] Mania Tlili, Tarek M. Hamdani, "Large scale data clustering validity", *6th International Conference of Soft Computing and Pattern Recognition* 348-352, 2014,.

[16] Ganeshayya Shidaganti, S. Prakash, "Feedback analysis of unstructured data from collaborative networking a Large scale data analytics approach", *CIMCA IEEE, International conference on circuits, communication, control and computing*.2014, 343-347,

