DOI: https://doi.org/10.24113/ijoscience.v6i5.290

A Review on Unstructured Data using k-Mean Algorithm

Akhilesh Sharma
M. Tech. Scholar
Department of Computer Science and Engineering
Oriental Group of Institutes Bhopal, India
akhilesh.sharma.mca@gmail.com

Abstract: **Unstructured data are the data without identifiable structure, audio, video and images are few examples. Clustering one of the best techniques in the knowledge extraction process. It is nothing but a grouping of similar data to form a cluster. The distance between the data in one cluster and the other should not be less. Many algorithms are practiced for clustering, in that k-mean clustering is one of the popular terms 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 with both. This paper overcomes that limitation by proposing a new k –mean algorithm for extracting hidden knowledge by forming clusters from the combination of unstructured datasets.**

Keywords: **structured data, clustering, unstructured data, k-mean**

I. INTRODUCTION

Clustering is an unsupervised learning Algorithm. It deals unlabeled data. It is used to find the group similarity. The important part of clustering is measuring the distance between data points. Clustering permits to find the hidden relationship between data points. Intra cluster minimization and inter cluster maximization will be used for creating good clusters. Clustering analysis is a process of Identifying groups. The clusters used to find the relationship between variables. The process involved is

1. Preparing a dataset.
2. Preprocessing of data.
3. Generating clusters of the data.
4. Interpreting results –validating clusters

If the data is good and clean then the models can be constructed easily, so that the performance will increase [1]. To make the data sensible and extract meaningful value from unstructured data, classification and clustering are used. Clustering is one of the techniques to sort and organize the data in to logical grouping before analysis process starts. Clustering is also performed based on the distances [2].

II. LITERATURE REVIEW

Duong Van Hieu et al. [3] proposed algorithm for reducing executing time of the k-means .They implemented this by cutting off a number of last iterations. In this experiment method 30% of iterations are reduced, so 30% of executing time is reduced, and accuracy is high. However, the choosing randomly the initial centroids produces the instable clusters. Clustering result may be affected by noise points, so it produces inaccurate result.

Li Ma et al. [4] developed a solution for improving the quality of traditional k-means clusters. They used the technique of selecting systematically the value of k i.e number of clusters as well as the initial centroids. Also they reduced the number of noise points so the outlier's problem solved. This algorithm produces good quality clusters but it takes more computation time.

Omar Kettani, Faical Ramdani, Benaissa Tadili [5] work covers an algorithm designed for automatic clustering. This method computes the correct number of clusters on tested data sets. This method was compared with G-means. The comparison of algorithm shows that the proposed approach much better than G-means in terms of clustering accuracy.

Sk Ahammad Fahad et al. [6] in this paper proposed method first finds the initial centroid and puts an interval between those data elements which will not change their cluster and those which may change their cluster in the

subsequence iterations. So that it will reduce the workload significantly in case of very large data sets. We evaluate our method with different sets of data and compare with others methods as well.

III. CLUSTERING

It makes an important role in data analysis and data mining applications. Data divides into similar object groups based on their features, each data group will consist of collection of similar objects in clusters. Clustering is a process of unsupervised learning. Highly superior clusters have high intra-class similarity and low inter-class similarity. Several algorithms have been designed to perform clustering, each one uses different principle. They are divided into hierarchical, partitioning, density-based, model based algorithms and grid-based.

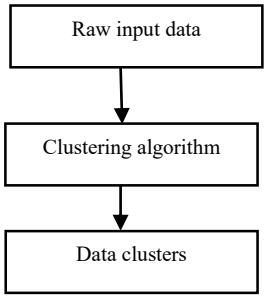

Fig. 1 clustering stages

There are two types of Clustering Partitioning and Hierarchical Clustering.

1. Hierarchical Clustering - A set of nested clusters organized in the form of tree.
2. Partitioning Clustering - A division of data objects into subsets (clusters) such that each data object is in exactly one subset.

IV. 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.

A. 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.

B. 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. Kmean algorithm is the most popular algorithm in partitioned clustering.

C. 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 [7].It is an iterative process. The similarity of each object with each of the currently existing clusters is calculated .Initially no clusters exist [7].if the similarity calculated is greater than the given threshold value. The object is placed in the relevant cluster otherwise new cluster is created.

D. 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. 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” .It is similar like linkage clustering method only difference is ,it is based on the connection points within the certain threshold [8] .

E. 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, they are

- Decision trees
- Neural networks

F. 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.

V. 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 [9].k-mean algorithm steps are as follows-

A. 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 ∈ 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. Assigned Center = di .center
6. For all center cj ∈ C do
7. Compute the squared Euclidean distance dist = dist (di,cj)
8. if dist < di .center Distance then
9. di .center Distance = dist
10. di .center = j
11. end if
12. end for
13. if di .center <> assigned Center then
14. Changed ++ 15 Recompute cj. new for next iteration
15. end if
16. end for
17. Until changed = = 0

The error sum of square is calculated by

$E=\sum_{i=1}^{k}\sum_{p\epsilon ci} dist(p, C_i)^2$

Where,

E = sum of square error

K = number of clusters

P = An object

C_i = i 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. 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 [10] .

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

REFERENCES

[1] Daqing Chen, Sai Laing Sain, Kun Guo," Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining", Springer, Vol. 19, Issue 3, pp 197–208, Sep 2012

[2] Takanobu Nakahara, Takeaki Uno, Yukinobu Hamuro," Prediction model using micro-clustering", 18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems - KES2014, Elsevier, ScienceDirect, vol. 35, pp. 1488 – 1494, 2014.

[3] V. Duon, M. Phayung. "Fast K-Means Clustering for very large datasets based on Map Reduce Combined with New Cutting

Method (FMR KMeans)", Springer International Publishing Switzerland, 2015.

[4] M. Li and al. "An improved k-means algorithm based on Map reduce and Grid", International Journal of Grid Distribution Computing, (2015)

[5] Omar Kettani, Faical Ramdani, Benaissa Tadili, "AKmeans: An Automatic Clustering Algorithm based on Kmeans ", Journal of Advanced Computer Science & Technology, vol. 4, issue 2 2015.

[6] Sk Ahammad Fahad "A Modified K-Means Algorithm for Big Data Clustering" April 2016.

[7] "IBM What is big data? — bringing big data to the enterprise". www.ibm.com. Retrieved 2013- 08-26.

[8] Francis, Matthew (2012-04-02). "Future telescope array drives development of exabyte processing". Retrieved 2012-10-24.

[9] Vladimír Holý, Ondřej Sokol, Michal Černý,"Clustering Retail Products Based on Customer Behaviour", Applied Soft Computing, Elsevier Vol 60, PP: 752-762, 2017.

[10] Zhexue Huang,CSIRO Mathematical and Information sciences ,Australia "clustering Large datasets with mixed Numeric And Categorical Values" * This Work was supported by the Cooperative Research Centre for Advanced Computational Systems (ACSys) established under the Australian Government's Cooperative Research Centres Program.