International Journal of Enhanced Research in Management & Computer Applications
ISSN: 2319-7471, Volume 7 Issue 2, February-2018, Impact Factor: 3.578

A Descriptive Study on Clustering Methods for Big Data

Parul

Research Scholar, Department of Computer Science and Application, M.D.U., Rohtak, Haryana

ABSTRACT

Today's age is the age of data. Nowadays the data is being produced at a huge rate. Researchers are facing many challenges in machine learning and data mining communities to use this massive data and to gain useful knowledge. Today's main challenge is to manage and analyze such a large data at a given time. Clustering helps in analyzing the data and helps in decision-making. Clustering is widely used in the diversity of applications such as marketing, insurance, monitoring, fraud detection and scientific discovery to extract widely useful information. In this letter, their classification provides an overview of the different clustering methods with normal work and a comparison (from theoretical point of view) between them. Paper presents properties, advantages and drawbacks of various clustering methods.

Key Words: cluster, clustering methods, outlier

I. INTRODUCTION

Large data reflects the strong growth of odd data flows due to the increasing use of new technologies. In fact, with the development of the web, the use of social network, mobile, connected and communicating objects is now more abundant than information and it is growing fast every day. Some studies argue that using this huge data handling and intelligently can be a new pillar of economics as well as scientific research, experimentation and simulation. In fact, many opportunities for Big Data are such health (increased efficiency of some treatment), biotherapy, marketing (increasing sales), transportation (cost reduction), business, finance (reducing risk), management (decision high With efficiency and speed), social, media, and government services.

Thus, clustering is the division of data in a group of equal objects. Clustering can be classified into the following categories:

A Partitioning Clustering
B. Hierarchical Clustering
C density based clustering
D. Model based clustering
E. Grid Based Clustering

Paper is organized in this way: Section II is a related study; Completing briefly the ideas of researchers on the clustering techniques given in their research papers. Section III includes details of various clustering methods. The fourth section describes the challenges of Big Data: Paper ends in section V in the end.

II. RELATED STUDY

• Zomaya et al [1] present surveys of existing clustering algorithms of different categories (segmentation-based, hierarchical-based, density-based, grid-based and model-based). In his work he compared his five-categories with his most representative algorithm, his goal was to perform best for Big Data.
• Another recent research [2] presents a general view of data mining algorithms and platforms, which can be used in the field of Big Data by discussing various challenges and characteristics.
• In [3] researchers have reviewed some older algorithms that can handle large data sets as closest neighbor discovery, decision tree and neural network.
• Reference [4] has classified the clustering algorithm in four types, has shown its professionals and the opposition and compares them to different factors.
• Reference [5] has described the various limitations of Kashmir algorithm and the different techniques used to remove them.

III. CLUSTERING METHODS

A. Partitioning method

Partition-based method divides the data objects into several partitions (clusters). In this method, the data object is divided into non-overlapping subsets, such that all the data objects in the same cluster are close to the center mean. In this method, all groups are determined immediately. Partitioning methods relocate instances by moving them from one cluster to another, starting from an initial partitioning, starting with the initial partition.

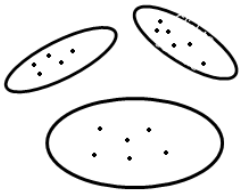

Figure1: shows split clustering

For such methods, generally the number of groups should be determined by the user. Unless a (optical) optimal partition is obtained, this method reduces the clustering criteria given by transferring data points between groups. K-mean and K-medoids are examples of partitioning-based methods [2] the groups of this method should meet these two requirements: -

(1) Each group or cluster should have at least one object.
(2) Each object must be related to exactly one group or cluster.

The advantages and disadvantages of partitioning clustering methods are as follows:
Benefit

1. Easy to understand and implement.
2. Produce more dense clusters than hierarchical method, especially when the groups are spherical.
3. Able to process large datasets.

Drawbacks

1. Poor handling of noise data and outlets.
2. Works only on numerical data.
3. Blank Cluster Generation Problems

B. Hierarchical method

As the name suggests, this method creates groups in a hierarchical order i.e. this is the form of nested cluster arranged in the hierarchical tree. It forms clusters in a top-down or bottom-up fashion. The hierarchical clustering method is of two types -

1. Agglomerative - This is a downward approach, in this approach, initially each object is considered as separate cluster. After this, it is merged into two or more suitable groups to create new groups, this merger of the cluster is recursively, as long as the desired cluster structure or stop criteria (K) of desired groups does not reach.

2. Divisive - this is an top down approach. In this approach, initially, the whole dataset is considered to be a cluster. The cluster is divided into sub-clusters, which in turn is divided into more sub-clusters. This process is repeated until the stopping criterion is completed (desired cluster k).

Benefit

1. It is more versatile
2. Less sensitive for noise and outliers.
3. Any number of groups can be obtained by cutting dendrogram at the desired level.
4. Apply for any specialty type.

Drawbacks

1. Backtracking is not possible.

C. Density based method

Density-based clustering method is based on the density, connectivity and range concepts. This method creates groups based on the density of data points in an area and continues to increase a given cluster until the density in the neighborhood (number of objects or data points) is greater than some threshold .Therefore, each data instance in the cluster the neighborhood of a given radius has to contain at least a minimum number of objects.. This method makes arbitrarily shaped groups because the cluster moves towards density in any direction. Since this method takes the form of groups based on the density of data points, it naturally eliminates the outliers or noise data points. DBSCAN , OPTICS and DENCLUE are examples of density based algorithms.

Benefit

1. Outliers Resistant
2. Forms clusters of arbitrary shapes.
3. Does not require the number of clusters.
4. Insensitive to order data objects

Drawbacks

1. Unsuitable for high-dimensional datasets due to the curse of dimensionality phenomenon.
2. Its quality depends on threshold set.

D. Model based

Model-based clustering method optimizes fit between given data and some (predefined) mathematical models. It assumes that the data is generated by a model or underlying probability distribution mix and attempts to recover the original model from the data. Models that we recover from data define clusters and provide objects to clusters. It creates a way to automatically determine the number of groups by taking noise (outliers) based on standard data and thus provides a strong clustering method. . EM (which uses a mixture density model), COBWEB (conceptual clustering) and neural network approaches (such as self-organizing feature maps) are examples of model based clustering methods.

Benefit
1. Robust to noisy data or outlier.
2. Fast processing speed
3. It determines the number of clusters to generate.

Drawbacks

1. Complex in nature

E. Grid Based

Grid-based clustering uses a multi-resolution grid data structure. It is used to create groups in a large multi-dimensional space, where groups are considered to be more dense regions than their environments. This method divides the space into cells of a finite number which creates a grid structure on which all the functions of clustering are performed. It is different from the traditional clustering algorithm in which there is no connection with the data point but with the value space around the data point. Deposit grid-data separates the grid-based clustering techniques from the number of data objects that employ a similar grid to collect regional statistical data, and then do clustering directly to the grid instead of the database. Grid based methods assist in expressing data at different levels of expansion, based on all the characteristics, which have been selected as dimensional features. In this approach, cluster data is represented more meaningfully. A specific grid-based clustering algorithm consists of the following five basic steps:

1. Creating grid structure that is, dividing the data space into a limited number of cells.
2. Calculating the cell density for each cell.
3. Sorting cells according to their density.
4. Identifying Cluster Centers.
5. Traversal of neighboring cells.

Benefit

1. Fast processing time.
2. Independent of the number of data objects.

Drawbacks

1. Depending on the number of cells in each dimension in the space of quantity.

IV. BIG DATA CHALLENGES

Generally, the data available on the web grows faster in the mass. In general, we find that the data can be classified into three basic types: structured (basic data types such as integers, letters and integers or array of characters. Is used in the database), unstructured (no predefined format: emails, books, magazines, documents, videos, photos) and semi-structured (two pages the type is a combination of data, they are usually used in XML). Most of the data produced are unstructured and conventional database management tools are unable to handle this kind of information.

Indeed, the extreme challenge of Big Data is to make heterogeneous data (weather, logistics, geolocation, car traffic) and associate them to extract useful information and thus improve the various sectors exploiting this huge amount of data very wide and dispersed.

Volume: Too many data (which I have labeled a "Ton Bytes") to suggest that the actual numerical scale on which data volume in a particular setting becomes challenging is domain-specific, but we all agree that Now we have a ton of "bytes")

Variety: complexity, thousands or more features in every data item, curtail of dimensions, connective explosions, many data types and many data formats.

Velocity: High data in our systems, real-time, incoming and outgoing data and information:

Veracity: A lot of training samples for essential and substantial data, rich micro-scalable model-building and model verification for testing many different concepts, fine-tuning "truth" in your data collection, from which "whole population analytics"

CONCLUSION

In this paper, we studied various clustering methods and algorithms based on these methods. Clustering is widely used in many applications, each clustering system has its own pros and cons. K means (division based) is the simplest of all algorithms. But its use is limited to numerical data values only. Increasing the performance of the k-mean algorithm increases the number of cluster growth. The hierarchical method uses nested clusters by splitting or merging data points. No backtracking is allowed in this method. Density based method is designed to create groups of arbitrary shapes. It automatically creates clusters, that is, there is no need to mention the number of groups and naturally removes the outliers. The grid-based method mainly focuses on spatial data.

Generally, in order to manage the large quantities of data keeping in mind the requirements of acceptable resources, we have to improve the clustering algorithm, reducing their complexity in terms of time and memory.

REFERENCES

[1]. A. Fahad, N. Alshatri, Z. Tari, A. ALAmri, A. Y. Zomaya, I. Khalil, F. Sebti, and A. Bouras, “A Survey of Clustering Algorithms for Big Data: Taxonomy & Empirical Analysis,” IEEE transactions on emerging topics in computing, 2014.
[2]. A. Sherin, S. Uma, K.Saranya and M. Saranya Vani ”Survey On Big Data Mining Platforms, Algorithms And Challenges”. International Journal of Computer Science & Engineering Technology, Vol. 5 No, 2014.
[3]. C. YADAV, S. WANG, et M. KUMAR, “Algorithm and approaches to handle large Data-A Survey,” International Journal of computer science and network, vol 2, issue 3, 2013.
[4]. Preeti Baser, Dr. Jatinderkumar R. Saini “A Comparative Analysis of Various Clustering Techniques used for Very Large Datasets”
[5]. Kehar Singh, Dimple Malik and Naveen Sharma "Evolving limitations in K-means algorithm in data mining and their removal" IJCEM International Journal of Computational Engineering & Management, Vol. 12, April 2011,
[6]. Keshav Sanse, Meena Sharma ”Clustering methods for Big data analysis” International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 3, March 2015
[7]. Btissam Zerhari, Ayoub Ait Lahcen, Salma Mouline “Big Data Clustering: Algorithms and Challenges” Research Gate Conference Paper May 2015
[8]. Yi Wang, Qixin Chen, Chongqing Kang, Qing Xia “Clustering of Electricity Consumption Behavior Dynamics Toward Big Data Applications “ in IEEE TRANSACTIONS ON SMART GRID, VOL. 7, NO. 5, SEPTEMBER 2016
[9]. Apache Hive. Available at http://hive.apache.org
[10]. http://blogs.worldbank.org/voices/meet-winners-and-finalists-firstwbg-big-data-innovation-challenge.