K-means algorithm based Clustering for Big data

Khevana Shah

I.T. Department , L D College of Engineering

University Area, Ahmedabad, Gujarat, India. 380015

email:khevanashah19@gmail.com

Abstract

Clustering is a data mining technique used to place data elements into related groups without advance knowledge of the group definition. Clustering is a pro-cess of partitioning a set of data in a set of meaningful sub-classes, called cluster. In this paper, we propose to give a review of the most used clustering methods. First, we give an introduction about clustering methods, how they work and their main challenges. Second, we present the clustering methods with some comparisons including mainly the classical partitioning clustering methods like well-known k-means algorithms, Gaussian Mixture Modals and their variants, the classical hierarchical clustering methods. Clustering algorithms can be categorized into partition-based algorithms, hierarchical-based algorithms, density-based algorithms and grid-based algorithms. Partitioning clustering algorithm splits the data points into k partition, where each partition represents a cluster. Hierarchical clustering is a technique of clustering which divide the similar dataset by constructing a hierarchy of clusters. Density based algorithms and the cluster according to the regions which grow with high density. It is the one-scan algorithms. Grid Density based algorithm uses the multi resolution grid data structure and use dense grids to form clusters. Its main distinctiveness is the fastest processing time. In this survey paper, an analysis of clustering and its different techniques in data mining is done.

Keywords - *Clustering, Types of clustering, Classification, Data mining, big data.*

INTRODUCTION

Clustering is a process of grouping objects with similar properties. Any cluster should exhibit two main properties; low inter-class similarity and high intra-class similarity. Clustering is an unsupervised learning i.e. it learns by observation rather than examples. There are no preened class label exists for the data points. Cluster analysis is used in a number of applications such as data analysis, image processing, market analysis etc. Clustering helps in gaining, overall distribution of patterns and correlation among data objects. [8]

The notion of a "cluster" varies between algorithms and is one of the many decisions to take when choosing the appropriate algorithm for a particular problem. At last the terminology of a cluster seems obvious: a group of data objects. How-ever, the clusters found by different algorithms vary significantly in their properties, and understanding these "cluster models" is key to understanding the differences between the various algorithms. Typical cluster models include: Connectivity models, Centroid models, Distribution models, Density Models, Subspace models, Group models and Graph-based models. A "clustering" is essentially a set of such clusters, usually containing all objects in the data set. Additionally, it may specify the relationship of the clusters to each other, for example a hierarchy of clusters embedded in each other. Data Clustering is one of the challenging mining techniques exploited in the knowledge discovery process. Clustering huge amounts of data is a difficult task since the goal is to and a suitable partition in an unsupervised way (i.e. without any prior knowledge) trying to maximize the similarity of objects belonging to the same cluster and minimizing the similarity among objects in different clusters? Many different

Clustering techniques have been de need in order to solve the problem from different perspective, i.e. partition based clustering, density based clustering, hierarchical methods and grid-based methods etc. [10]

In this paper we represent a survey of recent clustering approaches for data mining research. In Data Mining the two types of learning sets are used, they are supervised learning and unsupervised learning. [9]

a) Supervised Learning In supervised training, data includes together the input and the preferred results. It is the rapid and perfect technique. The accurate results are recognized and are given in inputs to the model through the learning procedure. Supervised models are neural network, Multilayer Perceptron and Decision trees. b) Unsupervised Learning The unsupervised model is not provided with the accurate results during the training. This can be used to cluster the input information in classes on the basis of their statistical proper-ties only. Unsupervised models are for dissimilar types of clustering, distances and normalization, k-means, self-organizing maps.

OVERVIEW OF DIFFERENT CLUSTERING ALGORITHMS

Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with ending a structure in a collection of unlabelled data. Clustering is a division of data into groups of similar objects. Clustering algorithm can be divided into the follow-in categories: A. Hierarchical clustering algorithm B. K-means clustering algorithm C. Density Based Clustering algorithm D. Partition clustering algorithm E. Grid based clustering algorithm

A. HIERARCHICAL CLUSTERING Hierarchical clustering algorithm group's data objects to form a tree shaped structure. It can be broadly classic end into agglomerative hierarchical clustering and divisive hierarchical clustering. In agglomerative approach which is also called as bottom up approach, each data points are considered to be a separate cluster and on each iteration clusters are merged based on a criteria. The merging can be done by using single link, complete link, centroid or wards Method. In divisive approach all data points are considered as a single cluster and they are spliced into number of clusters based on certain criteria, and this is called as top down approach.

Advantages of hierarchical clustering 1. Embed-deed exibility regarding the level of granularity. 2. Ease of handling any forms of similarity or distance. 3. Applicability to any attributes type.

Disadvantages of hierarchical clustering 1. Vagueness of termination criteria. 2. Most hierarchal algorithm do not revisit once constructed clusters with the purpose of improvement.

B. K-Mean Clustering Algorithm K-means clustering is a partitioning method. K-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.

The k-means algorithm has the following important properties: 1. it is efficient in processing large data sets. 2. It often terminates at a local optimum

3. It works only on numeric values. 4. The clusters have convex shapes

DENSITY BASED CLUSTERING ALGORITHM

Density based algorithm continue to grow the given cluster as long as the density in the neighbourhood exceeds certain threshold [6]. This algorithm is suitable for handling noise in the dataset. The following points are enumerated as the features of this algorithm. 1. Handles clusters of arbitrary shape 2. Handle noise 3. Needs only one scan of the input dataset. 4. Needs density parameters to be initialized.

DBSCAN, DENCLUE and OPTICS are examples for this Algorithm.

PARTITIONING ALGORITHMS

partitioning algorithms divide data into several subsets. The reason of dividing the data into several subsets is that checking all possible subset systems is computationally not feasible; there are certain greedy heuristics schemes are used in the form of iterative optimization. Specifically, this means differant relocation schemes that iteratively reassign points between the k clusters. Relocation algorithms gradually improve clusters.

Partitioned Clustering There are many methods of partitioning clustering; they are k-mean, Bi-secting K Means Method, Medoids Method, PAM (Partitioning around Medoids), CLARA (Clustering Large Applications) and the Probabilistic Clustering. We are discussing the k-mean algorithm as: In k-means algorithm, a cluster is represented by its centroid, which is a mean (average pt.) of points within a cluster. This works efficiently only with numerical attributes. And it can be negatively affected by a single outlier. The k-means algorithm is the most popular clustering tool that is used in scientific and industrial applications. It is a method of cluster analysis which aims to partition "n? observations into k clusters in which each observation belongs to the cluster with the nearest mean

GRID BASED ALGORITHMS

Grid-based clustering where the data space is quantized into nine number of cells which form the grid structure and perform clustering on the grids. Grid based clustering maps the in nine number of data records in data streams to nine numbers of grids. Grid based clustering is the fastest processing time that typically depends on the size of the grid instead of the data. The grid based methods use the single uniform grid mesh to partition the entire problem domain into cells and the data objects located within a cell are represented by the cell using a set of statistical attributes from the objects. These algorithms have a fast processing time, because they go through the data set once to compute the statistical values for the grids and the performance of clustering depends only on the size of the grids which is usually much less than the data objects. The grid-based clustering algorithms are STING, Wave Cluster, and CLIQUE. All these methods use a uniform grid mesh to cover the whole problem. For the problems with highly irregular data distributions, the resolution of the grid mesh must be too many to obtain a good clustering quality. A mesh can result in the mesh size close to or even exceed the size of the data objects, which can significant increase the computation load for clustering.

BIG DATA CLUSTERING METHODS

For some applications like medicals, aerospace we can have very large databases called also "Big Data". For such databases, a huge amount of memory and time complexity (computation) are required for the corresponding clustering which makes the classic clustering methods unable to handle those huge amounts of data, so new methods are required. To reduce execution time and memory, some methods replace real data by integer values, other methods use samples of the data or summary of the data, etc. Due to its low computational cost and easily parallelized process, k-means algorithm is still used for big data clustering, but improvements and variants are needed to support multiple core parallelism, to handle outliers, to reduce complexity for very big data and for more convergence efficiency. To solve scalability issue for kernel k-means, approximate kernel k-means is proposed which use sampling to reduce execution time. [5] A. Incremental mining Those techniques as DIGNET and siibFCM based on k-means algorithm, but use only one pass, thus allows reducing execution time and handling noise but also affects cluster quality besides the clustering strongly depends on data or-der. B. Probability based for clustering big data CLIQUE is an EM algorithm variant with much higher scalability. It can handle high dimensional data and big data requirements. C. Squashing techniques Those techniques (ex: BIRCH, Bubble, Bubble-FM) scan data to compute certain data summaries then cluster the summaries using height balanced tree of nodes. D. Distributed methods Distributed or parallel methods distribute computation on many computers to improve scalability, like DBCURE-MR , P-PIC Google Map Reduce algorithm, HDFS, Hadoop, etc.

ANALYSIS AND EVALUATION

In this section, we discuss the two methods used for evaluation of our experimental outcomes: (i) manual evaluation using a human assessor and (ii) Evaluation using purity.

A. Human Assessors In order to judge the clusters manually, the clustering program developed for this work outputs the original review text of the cluster centroid plus a sample of ?. random re-views in the same cluster. From this, a human assessor can view the output and decide whether or not the sample reviews primarily focus on the majority class topic. Results were inconsistent. Some clusters generated were successful, while others were not as. A study with a variety of human assessors was not per- formed. Only one human performed the judgments for this work. Based on this analysis, it was determined that K-means clustering performed slightly better in regards to clustering reviews based on topics.

B. Purity 1) Definition of Purity: Purity is an external metric for evaluating multiclass clusters. It computes a value for each individual cluster based on how accurate the classification was for the majority class found in the cluster. Purity is defined as follows: where T is the set of cluster labels, ? is the set of clusters and N is the number of reviews. The purity for an individual cluster can be calculated by simply applying the formula inside the summation to a singular cluster in a clustering instance.

$$\text{purity}(!;>) = \frac{1}{N}\ {}^{K}_{X}\max_j\, !_k \, {}^{\backslash}\!>_j \qquad (1)$$

V. CONCLUSIONS The paper describes different methodologies and parameters associated with different clustering algorithms used in larger data sets. And it gives an overview of different clustering algorithms used in large data sets. Then de-scribes about the general working behaviour, and the methodologies followed on these approaches and the parameters which used in these algorithms with large data sets. The overall goal of the data mining process is to extract information from a large data set and transform it into an understand-able form for further use. Clustering is important in data analysis and data mining applications. It is the task of grouping a set of objects so that objects in the same group are more similar to each other than to those in other groups (clusters). Clustering can be done by the different no. of algorithms such as hierarchical, partitioning, grid and density based algorithms. Hierarchical clustering is the connectivity based clustering. Partitioning is the centroid based clustering, the value of k-mean is set. Density based clusters are de ned as area of higher density then the remaining of the data set. Grid based clustering is the fastest processing time that typically depends on the size of the grid instead of the data. The grid based methods use the single uniform grid mesh to partition the entire problem domain into cells.

REFERENCES

[0] Yu Hu, Hans Hellendoorn, "Mixture-Model-Based Clustering for Daily Tra c Volumes" in 2015 IEEE 18th International Conference on Intelligent Transportation Systems.

[2] BINGJING CAI, HAIYING WANG, HUIRU ZHENG, HUI WANG "An Improved Random Walk Based Clustering Algorithm for Community Detection in Complex Networks" in 978-1-4577-0653-0/11//2011 IEEE

[3] Koustuv Dasgupta, Konstantinos Kalpakis and Parag Namjoshi "An E cient Clustering-based Heuristic for Data Gathering and Aggregation in Sensor Networks" in 2003 IEEE.

[4] Chantal Fry, Sukanya Manna "Can we Group Similar Amazon Reviews: A Case Study with Dif-ferent Clustering Algorithms" in 2016 IEEE Tenth International Conference on Semantic Computing.

[5] Abdelkarim Ben Ayed, Mohamed Ben Halima, Adel M. Alimi "Survey on clustering methods: Towards fuzzy clustering for big data" in Interna-tional conference of soft computing and pettern recognization.

[6] Amandeep Kaur Mann, Navneet Kaur "Survey Paper on Clustering Techniques" in International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 4, April 2013.

[7] Siuly, Y. Li, and P. Wen "Analysis and classi-cation of EEG signals using a hybrid clustering technique" in The 2010 IEEE

[8]ICME International Conference on Complex Medical Engineering July 13-15,2010, Gold Coast, Australia.M.Vijayalakshmi,MCA,M.Phil, M.Renuka Devi,MCA,M.Phil,(Phd) " A Survey of Di erent Issue of Di erent clustering Algorithms Used in Large Data sets" in International Journal of Advanced Research in Computer Science and Software Engineering.

[9] Amandeep Kaur Mann Navneet Kaur "Review paper on clustering technique" in Global Journal of Computer Science and Technology Software Data Engineering Volume 13 Issue 5 Version 1.0 Year 2013

[10] Anoop Kumar Jain and Satyam Maheswari "Survey of Recent Clustering Techniques in Data Mining" in International Archive of Applied Sciences and Technology Volume 3 [2] June 2012: 68 - 75