International Journal of Innovative Research in Technology & Science

Received 6 August 2018, Received in revised form, 20 August 2018, Accepted 23 September 2018, Available online 30 September 2018

ISSN: 2321-1156 Volume VI Issue V, September 2018

Performance Analysis of Big Data Using K-Means Clustering Algorithm

Pankaj S Deshmukh, Dept. of CSE, SRK University, Bhopal, India; Pank1980@gmail.com;

Harishchadra Vasantrao Parmar, Dept. of CSE, JTMCOE, Faizpur, MH, India; parmarharishv@gmail.com;

ABSTRACT

Data clustering is a classification technique. Clustering aims at making groups of objects, or clusters, in such a way that objects within the same cluster are very similar and several objects are separate several clusters though k-means clustering . It is very popular clustering, clustering algorithm primarily based on the K-means clustering and clustering algorithm improvement center in the cluster. Initial cluster centers and the local optimum each are base clustering's. Modifications that improve clustering algorithms, the focus is on finding better solutions for a number of clusters, it gets the more clustering centers. The algorithm is applied to the various dataset for clustering analysis. Proposed algorithm achievement information in the dataset its additional analysis of the most effective centroids, in clustering information as correctness. In clustering best information and minimize cluster fault through proposed algorithm. Clustering algorithm primarily based on K-means Clustering improves information as correctness.

Keywords: Data Mining, Clustering, K-means Clustering, Initialization, Partitioning Clustering, Information Extraction, Correctness Information, Hierarchical Clustering Algorithm.

I. INTRODUCTION

The field of information mining and knowledge discovery is rising as a replacement, elementary analysis space with necessary applications to Science, engineering, medicine, business, and education. Data processing makes an attempt to formulate analyze and implement basic induction processes that facilitate the extraction of significant data and information from unstructured knowledge. Size of databases in scientific and business application is large wherever the amount of records in an exceedingly dataset will vary from some thousand to thousands of millions. Clustering could also be outlined as an information reduction tool i.e. used to produce subgroups that are a lot of and a lot of manageable than individual data point. Basically, agglomeration is justified as a method used for grouping a large variety of knowledge into significant teams or clusters supported some similarity between data. Clusters are the teams that have knowledge similar on basis of common options and dissimilar to knowledge in different clusters. Data mining is that the process of extracting hidden, previously unknown and helpful data from giant knowledge bases and data warehouses. Data mining process involve steps like knowledge cleansing, integration, selection, transformation, data processing technique, and pattern analysis and information illustration. various data processing techniques are used like classification, clustering, association rules, successive patterns, Prediction, call trees, etc. are used in various applications. Here we tend to discuss regarding agglomeration algorithms like fuzzy c-means, k-means. Agglomeration is that the method of organizing knowledge objects into a collection of disjoint categories known as clusters. Clustering is an example of classification. Classification refers to a procedure that assigns knowledge objects to a collection of categories. Unattended means that clustering doesn't depend on predefined categories and training examples whereas classifying the information objects. Cluster analysis seeks to partition a given knowledge set into teams supported such that options so the information points among a bunch are a lot of kind of like one another than the points in numerous teams. Therefore, a cluster could be a collection of objects that are similar among themselves and dissimilar to the objects belonging to different clusters. Clustering is a crucial space of analysis that finds applications in several fields together with bioinformatics, pattern recognition, image process, marketing, data processing, economics, etc. Cluster analysis could be a one in every of the first knowledge analysis tool within the data processing. Clustering algorithms are primarily divided into two categories: hierarchical algorithms and Partition algorithms. A hierarchical clustering algorithm divides the given knowledge set into smaller subsets in hierarchical fashion. A partition clustering algorithm partition the information set into desired variety of sets in a very single step [1]. Clustering could also be defined as an information reduction tool i.e. used to produce subgroups that are a lot of and a lot of manageable than individual data point. Basically, clustering is justified as a method used for grouping a large variety of knowledge into significant teams or clusters supported some similarity between data. Clusters area unit the teams that have knowledge similar on basis of common options and dissimilar to knowledge in different clusters Cluster analysis teams knowledge objects primarily based solely on data found in knowledge that describes the objects and their relationships. The objects among a bunch are kind of like one another each different and totally different from the objects in other teams. Cluster analysis is one in every of the key data processing techniques, wide used for several sensible applications in varied rising areas like Bioinformatics. clustering is an unsupervised methodology that subdivides associate input file set into a desired variety of subgroups so the objects of a similar subgroup are going to be similar (or related) to at least one another and totally different from (or unrelated to) the objects in different teams [2].

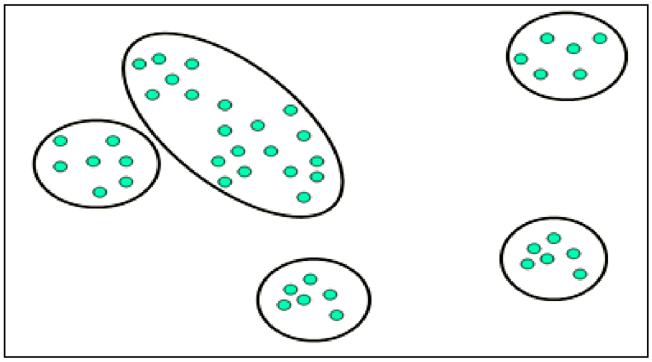

Figure 1 clustering

International Journal of Innovative Research in Technology & Science
Received 6 August 2018, Received in revised form, 20 August 2018, Accepted 23 September 2018, Available online 30 September 2018
ISSN: 2321-1156 Volume VI Issue V, September 2018

Clustering

Clustering is that the task of segmenting a various cluster into variety of comparable subgroups or clusters. Clusters of objects are shaped so objects at intervals a cluster have high similarity as compared to one another, however are terribly dissimilar to things in alternative clusters. Cluster is usually used to hunt for distinctive groupings at intervals an information set. The identifying issue between cluster and classification is that in cluster there are not any predefined categories and no examples. The objects are sorted along supported self-similarity. Typical business question that may be answered using cluster are: What are the groupings hidden in your information. That client ought to be sorted along for target promoting purposes and Cluster is sorted below descriptive data processing tasks.

(a) Varieties of Clustering

1.Hierarchical Clustering: a collection of nested clusters organized as a hierarchical tree hierarchical algorithms decompose an information D of n objects into many levels of nested partitioning till every set consists of just one object. There are two varieties of hierarchical algorithms; an agglomerative that builds the tree from the leaf nodes up, whereas a divisive builds the tree from the highest down [3].

2. Partitioning Clustering: A division knowledge objects into non-overlapping set (Clusters) specified every knowledge object is in precisely one set. Partitioning algorithms construct one partition of an information D of n objects into a collection of k clusters, specified the objects in a very cluster are a lot of kind of like one another than to things in numerous clusters.

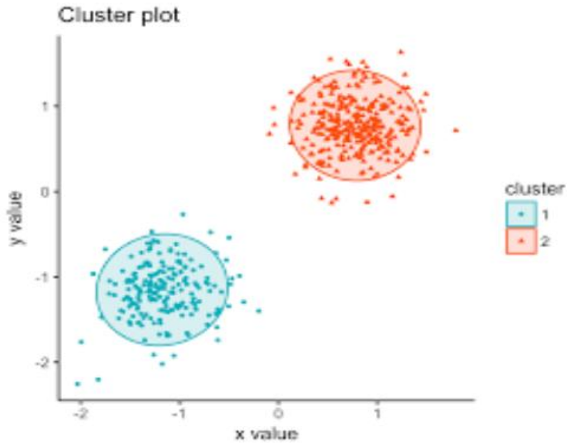

Figure 2 Partitioning cluster analysis

K-means could be a normally used partitioning primarily based clustering technique that tries to search out a user such that variety of clusters (k), that area unit depicted by their centroids, by minimizing the sq. error operates. The agglomeration methodology aims at optimizing the price operate to reduce the dissimilarity of the objects among every cluster, whereas increasing the dissimilarity of various clusters. Being a repetitious and hill-climbing methodology, it's quite sensitive to initial positions of cluster centers. what is more, since the associated price functions area unit nonlinear and multimodal, typically these algorithms converge to a neighborhood minimum i.e. it produces totally different clusters for various sets of values of the initial centroids [4,5].

III. RELATED WORK

In the section of related we mentioned the work that has done in the field of surface detect in image processing

Baolin Yi et al. [6] .The traditional k-means algorithm has sensitivity to the initial start center. To solve this problem, this paper proposed a new method to find the initial center and improve the sensitivity to the initial centers of k-means algorithm. The algorithm first computes the density of the area where the data object belongs to; then it finds *k* data objects, which are belong to high density area, as the initial start centers. Experiments based on the standard database UCI show that the proposed method can produce a high purity clustering results and eliminate the sensitivity to the initial centers to some extent.

.

Parvesh Kumar et al. [7]. Clustering is an important data mining technique to extract useful information from various high dimensional datasets. A wide range of clustering algorithms is available in literature and still an open area for researcher. K-means algorithm is one of the basic and most simple partitioning clustering technique is given by Mac Queen in 1967 and aim of this clustering algorithm is to divide the dataset into disjoint clusters. After that many variations of k-means algorithm are given by different authors. Here in this paper we make analysis of two variant of k-means namely global k-means and x-means over colon dataset.

Wong et al. [8]. This paper proposes an image clustering algorithm using Particle Swarm Optimization (PSO) with two improved fitness functions. The PSO clustering algorithm can be used to find centroids of a user specified number of clusters. Two new fitness functions are proposed in this paper. The PSO-based image clustering algorithm with the proposed fitness functions is compared to the K-means clustering. Experimental results show that the PSO-based image clustering approach, using the improved fitness functions, can perform better than K-means by generating more compact clusters and larger inter-cluster separation.

Mitchell Yuwono et al. [9]. Clustering can be especially effective where the data is irregular, noisy and/or not differentiable. A major obstacle for many clustering techniques is that they are computationally expensive, hence limited to smaller data volume and dimension. We propose a lightweight swarm clustering solution called Rapid Centroid Estimation (RCE). Based on our experiments, RCE has significantly quickened optimization time of its predecessors, Particle Swarm Clustering (PSC) and Modified Particle Swarm Clustering (mPSC). Our experimental results show that on benchmark datasets, RCE produces generally better clusters compared to PSC, mPSC, K-means and Fuzzy C-means. Compared with K-means and Fuzzy C-means which produces clusters with 62% and 55% purities on average respectively, thyroid dataset has successfully clustered on average 71% purity in 14.3 seconds.

Received 6 August 2018, Received in revised form, 20 August 2018, Accepted 23 September 2018, Available online 30 September 2018

ISSN: 2321-1156 Volume VI Issue V, September 2018

Yujun Lin et al. [10]. In this paper, an improved clustering method based on k-means is proposed. The proposed method consists of two major stages split and merge stages. Initially k-means method is employed in the dataset, and in the split stage, each cluster will be split into smaller clusters with k-mean repeatedly if they are sparse. Furthermore, in the merge stage the average distance is employed for merging standard. Experiments are tested on real and synthetic datasets. Experimental results demonstrate the proposed clustering method can detect clusters with different sizes, shapes and densities. Moreover, it outperforms the traditional k-means and single-link clustering method.

Liu Guoli et al. [11]. This paper deeply works over the aspect that the k-means c1usteriug algorithm is very sensitive to the initial values. In order to improve the dependence on the initial values, it proposes a new algorithm called K-means clustering algorithm based on iterative density (hereinafter referred to as IDKM). Through continuous modification to density threshold, it gets the more clustering centers, and merges them until the specified number of clustering center is met. IDKM algorithm is applied to the IRIS data set for clustering analysis, and then the result proves that the improved algorithm optimizes the dependence; Finally, IDKM is applied to Student achievement data set, the analysis of the clustering results guides students to study, it realizes the application of K-means clustering algorithm on data mining.

Zhang Min et al. [12]. K-means in the field of clustering analysis algorithms is a kind of more traditional algorithm. It exists many shortcomings. For example, K value is easily affected by manmade subjective factors, and the algorithm is easy to fall into a local optimal solution, and the clustering result is not stable, etc.; And K-means++ algorithm as the classic improved algorithm of K-means algorithm, but there is still a phenomenon of unstable cluster center. This paper is a kind of improvement aimed at the shortcoming of K-means++ algorithm, which introduces the concept of the variance in probability and mathematical statistics. Variance reflects the degree of density between samples and other samples. In the K-means++ algorithm when you select the first initial clustering center, you need to select minimum variance of sample points, which is in the position of the largest sample density, then you select the next cluster centers based on the weight method of D2 which is described in the K-means++ algorithm. Experimental results show the accuracy is higher and stability is better.

Altug Akay et al. [13].A novel data mining method was developed to gauge the experience of the drug Sitagliptin (trade name Januvia) by patients with diabetes mellitus type 2. To this goal, we devised a two-step analysis framework. Initial exploratory analysis using self-organizing maps was performed to determine structures based on user opinions among the forum posts. The results were a compilation of user's clusters and their correlated (positive or negative) opinion of the drug. Subsequent modeling using network analysis methods was used to determine influential users among the forum members. These findings can open new avenues of research into rapid data collection, feedback, and analysis that can enable improved outcomes and solutions for public health and important feedback for the manufacturer.

Ahmed Alsayat et al. [14]. The increasing influence of social media and enormous participation of users creates new opportunities to study human social behavior along with the capability to analyze large amount of data streams. One of the interesting problems is to distinguish between different kinds of users, for example users who are leaders and introduce new issues and discussions on social media. Furthermore, positive or negative attitudes can also be inferred from those discussions. Such problems require a formal interpretation of social media logs and unit of information that can spread from person to person through the social network. Once the social media data such as user messages are parsed and network relationships are identified, data mining techniques can be applied to group different types of communities. However, the appropriate granularity of user communities and their behavior is hardly captured by existing methods. In this paper, we present a framework for the novel task of detecting communities by clustering messages from large streams of social data. Our framework uses K-Means clustering algorithm along with Genetic algorithm and Optimized Cluster Distance (OCD) method to cluster data. The goal of our proposed framework is twofold that is to overcome the problem of general K-Means for choosing best initial centroids using Genetic algorithm, as well as to maximize the distance between clusters by pair wise clustering using OCD to get an accurate clusters. We used various cluster validation metrics to evaluate the performance of our algorithm. The analysis shows that the proposed method gives better clustering results and provides a novel use-case of grouping user communities based on their activities. Our approach is optimized and scalable for real-time clustering of social media data.

Patrick Breen et al. [15]. Pre-Exposure Prophylaxis (PrEP) is a groundbreaking biomedical approach to curbing the transmission of Human Immunodeficiency Virus (HIV). Truvada, the most common form of PrEP, is a combination of tenofovir and emtricitabine and is a once-daily oral mediation taken by HIV-seronegative persons at elevated risk for HIV infection. When taken reliably every day, PrEP can reduce one's risk for HIV infection by as much as 99%. While highly efficacious, PrEP is expensive, somewhat stigmatized, and many health care providers remain uninformed about its benefits. Data mining of social media can monitor the spread of HIV in the United States, but no study has investigated PrEP use and sentiment via social media. This paper describes a data mining and machine learning strategy using natural language processing (NLP) that monitors Twitter social media data to identify PrEP discussion trends. Results showed that we can identify PrEP and HIV discussion dynamics over time, and assign PrEP-related tweets positive or negative sentiment. Results can enable public health professionals to monitor PrEP discussion trends and identify strategies to improve HIV prevention via PrEP.

Received 6 August 2018, Received in revised form, 20 August 2018, Accepted 23 September 2018, Available online 30 September 2018

Wei Du et al. [16]. As a partition based clustering algorithm, K-Means is widely used in many areas for the features of its efficiency and easily understood. However, it is well known that the K-Means algorithm may get suboptimal solutions, depending on the choice of the initial cluster centers. In this paper, we propose a projection-based K-Means initialization algorithm. The proposed algorithm first employ conventional Gaussian kernel density estimation method to find the highly density data areas in one dimension. Then the projection step is to iteratively use density estimation from the lower variance dimensions to the higher variance ones until all the dimensions are computed. Experiments on actual datasets show that our method can get similar results compared with other conventional methods with fewer computation tasks.

IV CONCLUSION

The simplicity of k-means algorithm makes it the choice for many clustering tasks. However k-means suffers from the problems of random initialization, and the predetermined of number of clusters k .Enhance Clustering Algorithm based on K-means Clustering**.** The focus of the thesis is mainly on the algorithms which decrease the fault in data and based on the k-means algorithm .Clustering has a crucial role in different applications. The commonly used efficient clustering algorithm is k-means clustering. K-means clustering is an important topic in data mining. However k-means is still at the stage of exploration and development. Find and concludes that many improvements are basically required on k-means algorithm to improve problem of cluster initialization, cluster quality and error less data of our propose algorithm**.** Enhance Clustering Algorithm based on K-means Clustering get the optimize number of clusters. Both algorithm are simple to understand and can be applicable for various type of dataset. Minimum cluster and get useful information**,** Increase correctness in clustering technique. Reliable data and, Minimize fault values in dataset.

REFERENCES

[1]. Madhu Yedla,Srinivasa Rao Pathakota,TM Srinivasa, “Enhancing K-means Clustering Algorithm with Improved Initial Center”, International Journal of Computer Science and Information Technologies(IJCSIT),Vol.1(2) ,pp. 121-125, 2010.

[2]. Vaishali R. Patel, Rupa G. Mehta, “Clustering Algorithms: A Comprehensive Survey”, International Conference on Electronics, Information and Communication Systems Engineering, MBM Engineering College, JNV University, Jodhpur, 2011.

[3]. Kapil Joshi, Himanshu Gupta, Prashant Chaudhary, Punit Sharma, “Survey on Different Enhanced K-Means Clustering Algorithm”, International Journal of Engineering Trends and Technology, Volume 27 Number 4 - September 2015.

[4]. Mrs. Bharati M. Ramager, “Data Mining Techniques and Applications”, Indian Journal of Computer Science and Engineering Vol. 1 No. 4 301-305, 2010.

[5]. Mohammed J. Zaki, Limsoon Wong, “Data Mining Techniques”, Research gate, August 9, 2003.

[6]. Baolin Yi, Haiquan Qiao, Fan Yang, Chenwei Xu, “An Improved Initialization Center Algorithm for K-means Clustering”, 978-1-4244-5392-4/10/, IEEE, 2010.

[7]. Parvesh Kumar, Siri Krishan Wasan “Analysis of X-means and global k-means USING TUMOR Classification”, 978-1-4244-5586-7/10, Volume 5, 2010 IEEE.

[8]. Man To Wong, Xiangjian He , Wei-Chang Yeh , “Image Clustering Using Particle Swarm Optimization”, 978-1-4244-7835-4/11, IEEE, 2011.

[9]. Mitchell Yuwono, Steven W. Su, Bruce Moulton, Hung Nguyen, “Method for increasing the computation speed of an unsupervised learning approach for data clustering”, WCCI 2012 IEEE World Congress on Computational Intelligence ,June, 10-15, Brisbane, Australia, 2012.

[10]. Yujun Lin,Ting Luo, Sheng Yao, Kaikai Mo, Tingting Xu, Caiming Zhong, “An Improved Clustering Method Based on K-means”, International Conference on Fuzzy Systems and Knowledge Discovery,IEEE,2012.

[11]. Liu Guoli, HeBei LangFang, Wang Tingting, Li Yanping, YuLimei, Gao Jinqiao, “The Improved Research on K-Means Clustering Algorithm in Initial Values”, 2013 International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC),Dec 20-22 , Shenyang, China, IEEE,2013.

[12]. Zhang Min, Duan Kai-fei, “Improved research to k-means initial cluster centers”, 978-1-4673-9295-2/15, IEEE, 2015.

[13]. Altug Akay, Andrei Dragom and Bjorn-Erik Erlandsson, “A Novel Data-Mining Approach Leveraging Social Media to Monitor Consumer Opinion of Sitagliptin”, IEEE Journal Of Biomedical And Health Informatics, VOL. 19, NO. 1, January 2015.

[14]. Ahmed Alsayat, Hoda El-Sayed, “Social Media Analysis using Optimized K-Means Clustering”, 978-1-5090-0809-4/16,SERA, June 8-10, 2016, Baltimore, USA, IEEE, 2016.

[15]. Patrick Breen, Jane Kelly, Timothy Heckman, Shannon Quinn, “Mining Pre-Exposure Prophylaxis Trends in Social Media”, IEEE International Conference on Data Science and Advanced Analytics, 978-1-5090-5206-6/16, 2016.

[16]. Wei Du, Hu Lin, Jianwei Sun, Bo Yu and Haibo Yang, “A New Projection-based K-Means Initialization Algorithm”, Proceedings of 2016 IEEE Chinese Guidance, Navigation and Control Conference August 12-14, Nanjing, China, 2016.