International Journal of Research in Engineering, Science and Management
Volume-2, Issue-7, July-2019
www.ijresm.com | ISSN (Online): 2581-5792

A Comparative Study of Clustering Algorithms

Silky Chourasia[1], Brij Kishore[2]
[1]*Student, Dept. of Computer Science and Engg., Apex Institute of Engineering and Technology, Jaipur, India*
[2]*Assistant Professor, Dept. of Computer Science and Engg., Apex Inst. of Engg. and Technology, Jaipur, India*

***Abstract*: Clustering algorithms are quite popular with the big data analysis and is an essential tool for segregating data into small sections. Analyzing big data is a tough task but can be done easily if the data is segregated into sections on the basis of some parameters or on the basis of similarity. The purpose of clustering and classification algorithms is to make sense of and extract value from large sets of structured and unstructured data. If you're working with huge volumes of unstructured data, it only makes sense to try to partition the data into some sort of logical groupings before attempting to analyze it.**

***Keywords*: K-Means, Fuzzy Logic, Clustering, DB-Scan, Data Mining, Big Data.**

1. Introduction

This papers deals the theory and the comparative analysis of the clustering techniques. For a given data set of items, with certain features, and values for these features (like a vector). The task is to categorize those items into groups. k-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more.

Clustering refers to the division of data into groups of similar objects. Each group, or cluster, consists of objects that are similar to one another and dissimilar to objects in other groups. When representing a quantity of data with a relatively small number of clusters, we achieve some simplification, at the price of some loss of detail (as in lossy data compression, for example). Clustering is a form of data modeling, which puts it in a historical perspective rooted in mathematics and statistics. From a machine learning perspective, clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. Therefore, clustering is unsupervised learning of a hidden data concept. Clustering as applied to data mining applications encounters three additional complications: (a) large databases, (b) objects with many attributes, and (c) attributes of different types. These complications tend to impose severe computational requirements that present real challenges to classic clustering algorithms. These challenges led to the emergence of powerful broadly applicable data mining clustering methods developed on the foundation of classic techniques.

2. Literature survey

The detection of outliers and their removal plays an important role for processing meaningful and important data. Zhongxiang Fan proposed an improved K-Means algorithm by detecting outliers based on grid density which reduced the influence of outliers on the results [1].

Surasit Songama et.al proposed a two-phase classification method. In the first phase, the data patterns were clustered by K-Means algorithm and in second phase, outliers were constructed by a distance-based technique and a class label was assigned to each pattern [2].

Fabrizio Angiulli et.al provided distance-based outlier detection method to find out the top outliers in an unlabeled data set [3].

Harshada C. Mandhare performed a comparative study of cluster based outlier detection, distance based outlier detection and density based outlier detection techniques to find out most efficient outlier detection method [4].

Priority scheduling method is one amongst many conventional techniques for selecting a task from a lot. In this technique the priority of a task is defined or assigned dynamically by using quality of service parameters. It has been observed that assigning the priority to a task statically brings out many difficulties [5]. In [5], [6] the author assigns the priority by considering various constraints such as cost of application and execution time, based on QoS value.

In [7], [8] authors have insinuated an advanced scheduling algorithm for execution of tasks. This algorithm takes several dynamic decisions while considering many/multiple criteria in choosing a task, which would be mapped on to a specific VM. This approach has reduced the make span time of the whole

process very effectively.

In [9] authors have suggested an algorithm that improves the traditional cost based scheduling which improves the overall execution cost by allocating appropriate resources. In this technique various cloudlets have been categorized and are mapped on to resources as per their processing capabilities.

Many authors have considered priority to be one of the main factors. In [10] the authors have proposed a priority algorithm for finding the optimal choices. Priority based method is more beneficial as compared to FCFS technique. This algorithm can be improved even further by providing more information on regular fashion of its usage.

In [11] authors have coined a new technique which is Priority-based and handles Earliest-Deadline-First. Actually here two task scheduling techniques have been used in combination. In this approach the major focus is on memory utilization, optimization of existing resources and its allocation.

The K-means algorithm is a widely used basic clustering algorithm. The basic thought is to divide the data set X containing n data into k subsets (where k is user-specified), and each subset represents a cluster. The similarity in the same cluster is high, and the similarity between different clusters is low. The specific operation is that the user selects k points from n samples randomly as the initial clustering center through prior knowledge, and then calculates the distance between the remaining sample points and the initial clustering center by using Euclidean distance. The sample points are divided into the nearest initial clustering centers, and the set of points assigned to the same initial clustering center belongs to one cluster, forming a total of k clusters. When the partition is completed, the center point of each cluster is calculated again and the center point is the mean sample of each cluster again, and the process is repeated until the result of the partition does not change, that is, the objective function converges.

Arshad Muhammad Mehar et al. [12] built up another technique in view of internal validation measures, in order to locate an ideal estimation of k which, can give more steady groups. The proposed bunch legitimacy measure is utilized to figure the extent of basic articles in every pair of groups.

KA Abdul Nazeer et. al. [13] introduced an upgraded k-means that includes sorting the information set and parceling the sorted information set into "k" number of sets which, brought about better starting centroids along these lines enhancing the precision of this algorithm. The algorithm converges speedier contrasted with traditional algorithm of K-Means. The main drawback of this algorithm is the estimation of k (number of sought groups) still should be given as an information.

Shi Na et. al. [14] discussed an enhanced k-means calculation keeping in mind the end goal to tackle the issue of ascertaining the Euclidean separation between every information article and all group focuses in every emphasis, which expands the running time. In this approach a straightforward information structure is utilized to store some data in each emphasis, which can then be utilized as a part of the following cycle.

Shuhua Ren et. al. [15] displayed a calculation CV-k-means i.e. coefficient of variety k-means algorithm. This helped in lessening the impacts of immaterial qualities brought on by taking Euclidean separation as the comparability measure by presenting variety coefficient weight vector. The main problem is that the quantity of craved clusters (k) is to be given as an information.

Kunhui Lin et. al. [16] displayed an upgraded k-means clustering paper that optimized the starting focuses in light of data dimensional density which, affirm that these underlying focuses have the greatest contrast between groups. This algorithm is implemented on the Hadoop platform (MapReduce programming model). This methodology helped in enhancing the steadiness of the K-means clustering.

Anupama Chadha et. al. [17] exhibited a calculation that does not require K (number of bunches) as an information. It expelled the reliance on K which, is in some cases exceptionally hard to foresee as it requires domain knowledge. The work is restricted to numeric information set as it were.

Madhu Yedla et. al. [18] proposed a paper on K-means keeping in mind the end goal to locate the better initial centroids and thus lessens time unpredictability. The primary thought was that if the information point stays inside same bunch then the required complexity lessens from O(k) to O(1). Subsequently the aggregate time unpredictability diminishes to half i.e. for allotting the information directs it decreases toward O(nk) rather than O(nkl) which, brought about aggregate time taken to be O(nlogn). The limitation of this methodology was that the initializing the value of K was still required.

Z. Min et al.[19] acquainted a calculation with conquer the impediment of k-means++ approach by picking least change test as the principal starting grouping focus which, won't just dispose of the effect of confined focuses, additionally explores demonstrate that the enhanced calculation have bunching consequence of a moderately steady and better dependability and precision. The issues of such algorithm contains (a) Time utilization issue brought by the complexity of the approach used (b) how to keep away from regular computation issues if there should be an occurrence of a lot of information, and so on.

Soumi Ghosh et. al. [20] showed a comparative study between KM and FCM based on the number of samples and K. The experimental results show that the K-means algorithm is far better than FCM as it takes more time in performing fuzzy measure calculations which, results in increase in its time complexity and hence effects the result. Hence, no doubt FCM produces as good results as produced by KM close results but the time complexity is comparatively still high.

Shreya Banerjee et. al. [21] has done an evaluative study to compare between the sundry variants of KM such as Bisecting KM, FCM & Genetic KM. Genetic KM outperforms as compared to the rest of clustering techniques for both the internal and external indices & provides the best performance.

3. Conclusion

The above studied papers helps in understanding the concept of Data Mining algorithms specially K-means implementation and application. K-Means algorithm is basic clustering technique which can be customized to any level as per the application and requirements.

References

[1] Zhong xiang Fan, San Yun, "Clustering of College Students Based in Improved K-means Algorithm," IEEE International Computer Symposium, 2016.
[2] Surasit Songma, Witcha Chimphlee, Kiattisak Maichalernnukul, Parinya Sanguansat, "Classification via k-Means Clustering and Distance-Based Outlier Detection," IEEE Tenth International Conference on ICT and Knowledge Engineering, 2012.
[3] Fabrizio Angiulli, Stefano Basta, and Clara Pizzuti 2006. "Distance based detection and prediction of outliers," IEEE Transactions on Knowledge and Data Engineering, 2006 18(2), pp. 145-160.
[4] Harshada C. Mandhare. Prof. S.R. Idate, "A comparative Study of Cluster Based Outlier Detection, Distance Based Outlier Detection and Density Based Outlier Detection Techniques," IEEE International Conference on Intelligent Computing and Control Systems, 2017.
[5] Wei-Jen Wang, Y. Chang, Win-Tsung lo and Yi-Kang lee, "Adaptive Scheduling for Parallel Task with QoS Satisfaction for hybrid Cloud Environments", Journal of Supercomputing, vol. 66, no. 2, Nov. 2013.
[6] J. Moses, R. Iyer, R. Illikkal, S. Srinivasan and K. Aisopos, "Shared Resource Monitoring and Throughput Optimization in Cloud-Computing Datacenters International Parallel & Distributed Processing Symposium, IEEE, Anchorage, AK, 2011, pp. 1024- 1033.
[7] S. Ghanbari and M. Othman "A Priority Based Job Scheduling Algorithm in Cloud Computing", Procedia Engineering, vol. 50, pp. 778–785, 2012.
[8] H. Lawrance, and S. Silas "Efficient QoS Based Resource Scheduling Using PAPRIKA Method for Cloud Computing", (IJEST) International Journal of Engineering Science and Technology vol. 5, no. 3 pp. 638-643, March 2013.
[9] S. Selvarani and G. S. Sadhasivam, "Improved cost-based algorithm for task scheduling in cloud computing," International Conference on Computational Intelligence and Computing Research, Coimbatore, IEEE, 2010, pp. 1-5.
[10] J. Xiao and Z. Wang, "A Priority Based Scheduling Strategy for Virtual Machine Allocations in Cloud Computing Environment," International Conference on Cloud and Service Computing, Shanghai, 2012, pp. 50-55.
[11] G. Gupta, V. K. Kumawat, P. R. Laxmi, D. Singh, V. Jain and R. Singh, "A simulation of priority based earliest deadline first scheduling for cloud computing system," 2014 First International Conference on Networks & Soft Computing (ICNSC2014), Guntur, 2014, pp. 35-39.
[12] A.M. Mehar, K Matawie and A Maeder, "Determining an Optimal Value of K in K-means Clustering" in IEEE International Conference on Bioinformatics and Biomedicine,2013: pp. 51-55.
[13] K A Abdul Nazeer, S D Madhu Kumar, "Enhancing the k-means clustering algorithm by using a O(n logn) heuristic method for finding better initial centroids" in Second International Conference on Emerging Applications of Information Technology, 2011: pp. 261-264
[14] S. Na, L. Xumin and G.yong, "Research on k-means Clustering Algorithm" in IEEE Third International Symposium on Intelligent Information Technology and Security Informatic,2010:pp. 63-67.
[15] S. Ren, A. Fan, "K-means Clustering Algorithm Based On Coefficient of Variation" in IEEE 4th International Congress on Image and Signal Processing, Vol. 4, 2011, pp. 2076-2079.
[16] K. Lin, X. Li, J. Chen, Z. Zhang, "A K-means Clustering with Optimized Initial Center Based on Hadoop Platform" in The 9th International Conference on Computer Science & Education, 2014, pp. 263-266.
[17] A. Chadha, S. Kumar, "An Improved K-Means Clustering Algorithm: A Step Forward for Removal of Dependency on K" in IEEE International Conference on Reliability, Optimization and Information Technology, 2014, pp. 136-140.
[18] M. Yedla, S. R. Pathakota, T. M. Srinivasa, "Enhancing K-means Clustering Algorithm with Improved Initial Center" in International Journal of Computer Science and Information Technologies, Vol. 1 (2), 2010, pp. 121-125.
[19] Z. Min, Kai-fei, "Improved research to k-means initial cluster centers" in Ninth International Conference on Frontier of Computer Science and Technology, 2015, pp. 349-353.
[20] S. Ghosh, S.K. Dubey, "Comparative Analysis of K-Means and Fuzzy C-Means Algorithms" in International Journal of Advanced Computer Science and Applications, Vol. 4, 2013, pp. 35-39.
[21] S. Banerjee, A. Choudhary, S. Pal, "Empirical Evaluation of K-Means, Bisecting K-Means, Fuzzy C-Means and Genetic K-Means Clustering Algorithms" in IEEE International WIE Conference on Electrical and Computer Engineering, 2015, pp. 168-172.