International Journal of Innovative Research in Technology & Science (IJIRTS)

ISSN: 2321-1156 Volume V, Issue V, September, 2017

A REVIEW ON UNSUPERVISED LEARNING APPROACH BASE K-MEANS CLUSTERING ALGORITHM

Ankur Pathak1, Prof. Alok Sahelay2
Department of Computer Science & Engineering, OCT Bhopal

Abstract:-*Learning is that the method of generating helpful data from a large volume of information. Learning will be classified as supervised learning and unsupervised learning. Cluster may be a reasonably unsupervised learning. A pattern representing a typical behavior or characteristics that exist among every item will be generated. This paper provides a summary of various partition cluster algorithmic rule. In data processing, cluster may be a technique during which the set of objects are appointed to a group referred to as clusters. Cluster is that the most essential a part of data processing. K-means cluster is that the basic cluster technique and is most generally used algorithmic rule. It's additionally called nearest neighbor looking. It merely clusters the datasets into given range of clusters. Improve the performance of data set analysis victimization new unsupervised learning cluster approach get best data.*

Keywords: - Data Mining, Clustering, Supervised Learning, Unsupervised Learning, K-means clustering, Data, Dataset, Analysis.

I. INTRODUCTION

Due to the increased availability of computer hardware and software and the rapid computerization of business, large amount of data has been collected and stored in databases. Researchers have estimated that amount of information in the world doubles for every 20 months. However raw data cannot be used directly. Its real value is predicted by extracting information useful for decision support. In most areas, data analysis was traditionally a manual process. When the size of data manipulation and exploration goes beyond human capabilities, people look for computing technologies to automate the process [1]. 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 predefined 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. This paper describes about the general working behavior, the methodologies to be followed and the parameters which affects the performance of the partition clustering algorithms [2].

Overview of Different Clustering Algorithms: Clustering is a division of data into groups of similar objects Clustering algorithm can be divided into the following categories: (HCA) Hierarchical clustering algorithm, (PCA) Partition clustering algorithm, (SCA) Spectral clustering algorithm, (GCA) Grid based clustering algorithm, (DCA) Density based clustering algorithm [3].

HCA: Hierarchical clustering algorithm group's data objects to form a tree shaped structure. It can be broadly classified 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 the iteration clusters are merged based on a criteria. The merging can be done by using single link, complete link, and centroid or wards method. In divisive approach all data points are considered as a single cluster and they are splited into number of clusters based on certain criteria, and this is called as top down approach. Examples for this algorithms are LEGCLUST [4], BRICH [5] (Balance Iterative Reducing and Clustering using Hierarchies), CURE (Cluster Using Representatives) [6], and Chameleon [1].

PCA: Partition clustering algorithm splits the data points into k partition, where each partition represents a cluster. The partition is done based on certain objective function. One such criterion functions is minimizing square error criterion which is computed as, E = Σ || p – mi || 2 (1) where p is the point in a cluster and mi is the mean of the cluster. The cluster should exhibit two properties, they are (1) each group must contain at least one object (2) each object must belong to exactly one group. The main drawback of this algorithm [3] is whenever a point is close to the center of another cluster; it gives poor result due to overlapping of data points.

SCA: Spectral clustering refers to a class of techniques which relies on the Eigen structure of a similarity matrix. Clusters are formed by partition data points using the similarity matrix. Any spectral clustering algorithm will have three main stages [7]. They are 1. Preprocessing: Deals with the construction of similarity matrix. 2. Spectral Mapping: Deals with the construction of Eigen vectors for the similarity matrix 3. Post Processing: Deals with the grouping data points The following are advantages of Spectral clustering algorithm: 1. Strong assumptions on cluster shape are not made. 2. Simple to implement. 3. Objective does not consider local optima. 4. Statistically consistent. 5. Works faster. The major drawback of this approach is that it exhibits high computational complexity. For the larger dataset it requires O (n3) where n is the number of data points [8]. Examples for this algorithm are SM (Shi and Malik) algorithm, KVV (Kannan, Vempala and Vetta) algorithm, NJW (Ng, Jordan and Weiss) algorithm [4].

GCA: Grid based algorithm quantizes the object space into a finite number of cells that forms a grid structure [1].Operations are done on these grids. The advantage of this method is lower processing time. Clustering complexity is based on the number of populated grid cells and does not depend on the number of objects in the dataset. The major features of this algorithm are: 1. No distance computations. 2. Clustering is performed on summarized data points. 3. Shapes are limited to union of grid-cells. 4. The complexity of the algorithm is usually O (Number of populated grid-cells) STING is an algorithm [4].

DCA: Algorithm Density based algorithm continue to grow the given cluster as long as the density in the neighborhood exceeds certain threshold [1]. 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. We Needs only one scan of the input dataset. 4. Needs density parameters to be initialized. DBSCAN, DENCLUE and OPTICS [2] are examples for this algorithm

II.RELATED WORK

K. A. Abdul Nazeer et al [9] the major drawback of the k-means algorithm is about selecting of initial centroids which produces different clusters. But final cluster quality in algorithm depends on the selection of initial centroids. Two phases includes in original k means algorithm: first for determining initial centroids and second for assigning data points to the nearest clusters and then recalculating the clustering mean. But this enhanced clustering method uses both the phases of the original k-means algorithm. This algorithm combines a systematic method for finding initial centroids and an efficient way for assigning data points to clusters. But still there is a limitation in this enhanced algorithm that is the value of k, the number of desired clusters, is still required to be given as an input, regardless of the distribution of the data points.

Soumi Ghosh et al. [8] proposed a comparative discussion of two clustering algorithms namely centroid based K-Means and representative object based Fuzzy C-Means clustering algorithms. This discussion is on the basis of performance evaluation of the efficiency of clustering output by applying these algorithms. The result of this comparative study is that FCM produces closer result to the K-means but still computation time is more than k-means due to involvement of the fuzzy measure calculations.

According to Y. S. Thakare et al. [11], the performance of k-means algorithm which is evaluated with various databases such as Iris, Wine, Vowel, Ionosphere and Crude oil data Set and various distance metrics. It is concluded that performance of k-means clustering is depend on the data base used as well as distance metrics.

Cui, Xiaoli, et al. [12] proposed optimized big data K-Means using Map-Reduce in which they claimed to counter the iteration dependence of Map-Reduce jobs. They used a sequence of three Map-Reduce (MR) jobs. However, in their approach sampling technique is used in the first M-R joband in the final M-R Job the data set is mapped to centroids using the Voronoi diagram. Variety is an important feature in big data so using sampling techniques is questionable when applied to huge data sets in maintaining the quality of clustering.

Roy Kwang et al. [13].For Autonomous Cluster Initialization of Probabilistic Neural Network approach was demonstrated. In this approach statistical based Probabilistic Neural Network (PNN) was used for pattern classification problems with Expectation – Maximization (EM) chosen as the training algorithm. Global algorithm solves the problem of random initialization. Initially, user needs to predefine the number of clusters using trial and error method. Global K-means provides a deterministic number of clusters using a selection criterion. This model was well tested with Fast Global k-means to ensure their correct classification and computational times.

Siri KrishanWasan et al. [14].Global k-means is the incremental algorithm that allows us to add one cluster center at a time and uses each data point as a candidate for the k-th cluster center. Experimental results show that the global k-means algorithm considerably outperforms the k-means algorithms. New version of this algorithm is proposed in this paper, it uses minimizing an auxiliary cluster function to compute the starting point for the k-th cluster center. Numerical results of these experiments (i.e. 14 data sets) demonstrate the superiority of the new algorithm, however it required more computational time than global KMmean algorithm. Based on colon dataset, global k-means and x-means algorithms were analyzed. Comparison was made in respect of accuracy and convergence rate. Accuracy of global k-means is slightly more than accuracy of x-means. Number of trials to reach a global and a stable optimum solution is less for both the algorithms. Speed of execution is the fastest for xmeans in comparison to global k-means.

Bagirov et al. [15].The modified global k-means algorithm was developed for clustering in gene expression data sets which is effective for solving clustering problems in gene expression data sets. This algorithm computes clusters incrementally and to compute kpartition of a data set it uses k − 1 cluster centers from the previous iteration. Computation of the starting point for the k-th cluster center is the key point. Starting point is calculated by minimizing so-called auxiliary cluster functions.

III. EXPECTED OUTCOME

1. Decrease error in dataset
2. Smart Data Analysis
3. Reliable data
4. Best possible solution

IV. CONCLUSION

In this paper we study completely different KM clustering algorithms and examine their benefits and limitation. The paper describes completely different methodologies and parameters related to partition clustering algorithms. The disadvantage of k-means algorithmic program is to search out the best k value and initial centroid for every cluster and KMC provides resolution to overcome the drawbacks of suggests that algorithmic program however it's its own limitations like slow execution and huge area demand. To scale back these draw-backs of metric linear unit range of resolution and strategies had been projected that was economical as com-pared to km. they summarized most of them in our critical analysis section. They additionally projected a newly clustering algorithmic program a quicker improve the performance of data set analysis exploitation new unsupervised learning agglomeration approach get best data. Our algorithmic program needs less computing time and fewer distance calculations. It'll additionally take low memory space.

REFERENCE

[1]. E. A. Khadem, E. F. Nezhad, M. Sharifi, "Data Mining: Methods & Utilities", Researcher; 5(12):47-59. (ISSN: 1553-9865), 2013.

[2]. Jiawei Han, Micheline Kamber, "Data Mining Concepts and Techniques" Elsevier Publication.

[3]. P. Berkhin, Survey of Clustering Data Mining Techniques. Technical report, Accrue Software, San Jose, Cailf, 2002.

[4]. Santos, J.M, de SA, J.M, Alexandre, L.A, 2008. LEGClust- a Clustering Algorithm based on Layered Entropic sub graph. Pattern Analysis and Machine Intelligence, IEEE Transactions: 62- 75.

[5]. M. Livny, R. Ramakrishna, T. Zhang, 1996. BIRCH: An Efficient Clustering Method for Very Large Data-bases. Proceeding ACMSIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery: 103-114.

[6]. S. Guha, R. Rastogi, and K. Shim, 1998. CURE: An Efficient Clustering Algorithm for Large Databases. Proc. ACM Int'l Conf. Management of Data: 73-84.

[7]. M Meila, D Verma, Comparison of spectral clustering algorithm. University of Washington, Technical report, 2001.

[8]. Cai X. Y. et al, Survey on Spectral Clustering Algorithms. Computer Science: 14-18, 2008.

[9]. K. A. Abdul Nazeer, M. P. Sebastian, Improving the Accuracy and Efficiency of the k-means Clustering Algorithm, Proceedings of the World Congress on Engineering, VOL.1 2009, July 1 - 3, 2009, London.

[10]. Soumi Ghosh, Sanjay Kumar Dubey, Comparative Analysis of K-Means and Fuzzy C-Means Algorithms, International Journal of Advanced Computer Science and Applications, Vol. 4, No.4, 2013.

[11]. Y. S. Thakare, S. B. Bagal, Performance Evaluation of K-means Clustering Algorithm with Various Distance Metric, International Journal of Computer Applications (0975 n 8887) Volume 110 ñ No. 11, Jan., 2015.

[12]. Cui, Xiaoli and Zhu, Pingfei and Yang, Xin and Li, Keqiu and Ji, Changqing, Optimized big data K-means clustering using Map Reduce, The Journal of Super-computing, 70, 1249-1259 ,2014.

[13]. Chang, Roy Kwang Yang, Chu Kiong Loo, and M. V. C. Rao. "A Global k-means Approach for Autonomous Cluster Initialization of Probabilistic Neural Net-work." Informatics (Slovenia) 32, no. 2 2008 219-225.

[14]. Kumar, Parvesh, and SiriKrishan Wasan. "Analysis of X-means and global k-means USING TUMOR classification." In Computer and Automation Engineering (ICCAE), 2010 the 2nd International Conference on, vol. 5, pp. 832-835. IEEE, 2010.

[15]. Bagirov, Adil M., and KarimMardaneh. "Modified global k-means algorithm for clustering in gene expression data sets." In Proceedings of the 2006 workshop on intelligent systems for bioinformatics Volume 73, pp. 23-28. Australian Computer Society, Inc., 2006.