International Journal of Innovative Research in Technology & Science

Received 3 June 2018, Received in revised form, 24 June 2018, Accepted 25July 2018, Available online 25 July 2018

ISSN: 2321-1156 *Volume VI Issue IV, July 2018*

Performance Analysis Using Different Dataset Based on K-means Clustering: Survey

Yogita Mishra, M. Tech. Scholar Depart. Of CSE, SIRTS, RGPV, India; yogitamishra94@gmail.com;
Prof. Vijay Bhandari, Department of CSE, SIRTS, RGPV, India; vijaysirt@gmail.com;
Dr. Amit Shrivastava, Department of CSE SIRTS, RGPV, India; sagar.amitshri@gmail.com;

ABSTRACT

Clustering could be a very important task in methoding process. K-Means cluster could be a cluster methodology within which the given data set is split into K variety of clusters and information partitioning a group of objects into homogeneous clusters could be a basic operation in data processing. The operation is required in a very variety of information mining tasks. Cluster or information grouping is that the key technique of the information mining. It's an unattended learning task wherever one seeks to identify a finite set of classes termed clusters to explain the information. The grouping of information into clusters relies on the principle of increasing the intra category similarity and minimizing the put down category similarity. The goal of cluster is to see the intrinsic grouping in an exceedingly set of unlabelled information. K-mean cluster is wide used to minimize square distance between options values of two points reside within the same cluster. Planned to use the Principal element Analysis methodology to reduce the information set from high dimensional to low dimensional. The new methodology is employed to search out the initial centroids to create the formula simpler and efficient. The planned methodology is simply implemented in matlab tool and is suitable for big information sets, like those in data processing applications. Experimental results show that, with a small loss of quality, the planned Method will significantly reduce the time taken than the standard kernel k-means cluster methodology. The planned Methodology is additionally compared with different recent similar methodology.

Keywords: Data mining, Clustering, unsupervised learning classification, partitioning clustering, K-means clustering method, Knowledge discovery system, density-based clustering.

1. INTRODUCTION

Advancement in sensing and digital storage technologies and their dramatic growth within the applications starting from marketing research to scientific knowledge explorations have created several high-volume and high dimensional information sets. Most of the information keep in electronic media have influenced the event of economical mechanisms for information retrieval and automatic data processing tools for effective classification and cluster of high-dimensional data. Additionally to the present, the exponential growth of high-dimensional knowledge needs advanced methoding (DM) methodology to automatically perceive process and summarize information. DM may be a method of extracting previously unknown, potentially helpful and ultimately understandable data from the high volume of knowledge. Data processing techniques may be generally classified into 2 major classes. Knowledge clump may be a method of characteristic the natural groupings that exists in an exceedingly given knowledge set, specified the objects within the same cluster are additional similar and also the objects in numerous clusters are less similar (in alternative words, dissimilar). It's been considered as a very important tool in numerous applications like pattern recognition, image process, data processing, remote sensing, statistics, etc [1].

CLUSTERING

Clustering could be a form of unsupervised learning not supervised learning like Classification. In cluster technique, objects of the data set are classified into clusters, in such the way that teams are terribly completely different from one another and also the objects within the same cluster or cluster are terribly the same as one another.

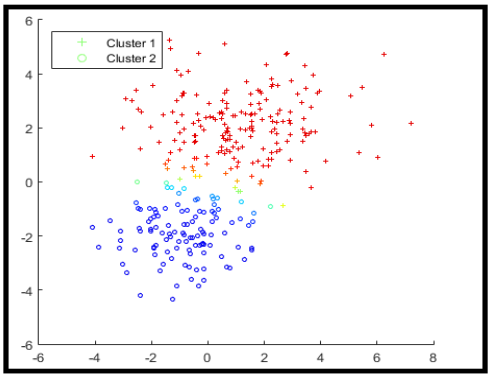

Figure1 Clustering

Not like Classification, during which predefined set of categories are given, however in cluster there aren't any predefined set of categories which implies that ensuing

Received 3 June 2018, Received in revised form, 24 June 2018, Accepted 25July 2018, Available online 25 July 2018
ISSN: 2321-1156 Volume VI Issue IV, July 2018

clusters don't seem to be best-known before the execution of cluster algorithmic program. During this these clusters are extracted from the data set by grouping the objects in it. Data mining involved with predictive information analysis involves totally different levels of learning, like (i) supervised (CLASSIFICATION) learning, involves with only labeled information (training patterns) and predicts the behavior of the unseen information sets and (ii) unsupervised (CLUSTERING) learning, involves with only unlabeled information, and (iii) semi-supervised learning. Cluster may be a tougher and difficult drawback than classification Forms of totally different levels of learning information [2].

Supervised learning: During this training information it contributes every the input and additionally the required results. These methods are fast and proper In supervised learning conjointly referred to as direct data processing the variables below Investigation are divided t into two groups: informative variables and one (or more) dependent variables. The aim of this analysis is to specify a relation between the variable quantity and instructive variables. The values of the variable quantity should be notable for a sufficiently massive a part of the information set to continue with directed data processing techniques. The correct results are best-known and are given in inputs to the model throughout the learning technique. Supervised models are neural network, many layers Perception, decision trees.

Unsupervised Learning: The model is not maintained with the proper results throughout the training. It should be used to cluster the computer file in classes on the support of their chance properties only. In unsupervised learning, all the variables are treated in same method, there's no distinction to the name purposeless data processing, still there's some target to realize. This target may well be as knowledge reduction as general or additional specific like cluster. The dividing line between unsupervised learning and supervised learning is that the same that distinguishes discriminate analysis from cluster analysis. Supervised learning needs, target variable ought to be outlined which an enough range of its values are given. Unsupervised learning generally either the target variable has between dependent and informative variables. However, in barely been recorded for too tiny variety of cases or the target variable is unknown .Unsupervised models are non identical kinds of cluster, amplitude and standardization, k-means, self organizing maps.

Semi-Supervised Learning: Semi-supervised learning some time additionally known as hybrid setting, involves partial labeled and unlabeled information sets for understanding the hidden behavior of the information sets. Cluster could be a harder and difficult drawback than classification. Kinds of completely different levels of learning information [3]

Types of Clusters

Well-separated clusters: cluster could be a set of purposes such any purpose during a cluster is nearer (or additional similar) to each different purpose within the cluster than to any point not within the cluster. A cluster may be a settle of purposes thus any purpose terribly} very cluster is nearest (or lots of similar) to every completely different purpose at intervals the cluster as differentiate to the other point that is not at intervals the cluster. Center-based clusters: once an object is additional near or almost like the cluster during which it resides then the opposite clusters then it's known as center-based cluster. Contiguous clusters (Nearest neighbor or Transitive): cluster could be a set of purposes such some extent during a cluster is nearer (or additional similar) to 1 or additional different points within the cluster than to any point not within the cluster. Density-based clusters: cluster could be a dense region of points that is separated by low-density regions, from different regions of high density. This sort of cluster used only if the clusters area unit irregular or tangled and when noise and outliers are present. Shared Property (Conceptual Clusters): it's the kind of clusters that share some common property or represent a specific concept [4].

2. LITERATURE REVIEW

AnkitaVimal et al [5]. A brief study of various distances measures and their effect on different clustering algorithms is carried out in this article. With the help of k-mean matrix partitioning and dominance based clustering algorithms; Euclidean distance measure and other four distance measure were studied to analyze their performance by accuracy of various techniques using synthetic datasets. Real-world data sets of cricket and synthetic datasets from Syndical software were used for cluster analysis. In this study it is found that the Euclidean distance measure performs better than the other measures.

Li et al. [6] proposed eliminating the issues of the traditional k-means clustering algorithm for clustering large Data. The k-means clustering algorithm, being efficiently able to handle the increasing size of data brings with it an increased time complexity. Authors proposed optimizing k-means according to the Hadoop cloud computing platform and Map

Received 3 June 2018, Received in revised form, 24 June 2018, Accepted 25July 2018, Available online 25 July 2018
ISSN: 2321-1156 Volume VI Issue IV, July 2018

Reduce Framework that allows distributed and parallel processing of data. The algorithm starts by initialization of the cluster centers followed by partitioning the dataset into equally sized small data blocks for parallel processing. The blocks are then exposed to the Map and Reduce tasks that run till the desired clustering results are achieved. Optimization of the k-means algorithm is also done in terms of initialization of cluster centers that otherwise cause instability in clustering results.

A k Patidar et al. [7] .used four standard similarity measure functions such as Euclidean, Cosine, Jaccard and Person correlation function in SNN clustering algorithm on a synthetic dataset, KDD Cup'99, Mushroom data set and some randomly generated database. In SNN technique generally data must be cleaned in order to find desired cluster. Here, they are inserting un-clustered data to desired core cluster discovered by SNN algorithm. Ultimately, they suggested in their studies that Euclidean measure performed well in SNN algorithm comparable to other three measures.

Feldman et al [8] proposed using coresets of large data instead of using large dataset for clustering purposes. A corset can be defined as a subset of the dataset with same properties of the dataset. Running the clustering algorithm on a corset helps in cutting down the query processing time though satisfying the exact constraints and optimality definitions as used by the dataset. The proposal relaxes the knowledge boundations of the previous algorithms of knowing the number of data points and the dimensions in advance and is limited to taking them in increasing order for each new inserted value. Using the merge and reduce paradigms, the k-means, PCA and projected clustering are made into parallel streaming clustering algorithms.

Carl Meyer et al. [9] proposed a methodology where cluster assembling is used to determine the number of clusters. They defined graph on similarity matrix by using different k values and also using different algorithms. A random walk then performed on graph to determine number of clusters from Eigen values of respective transition probability matrix. Through iteration of consensus clustering refinement is done to remove noisy data.

Agrawal et al. [10] has depicted about data mining applications and different methods of clustering documents. The objective of their work is to identify the clustering ability of the algorithms for identifying the clusters embedded in subspaces. These subspaces consist of high dimensional data and scalability. HPSO (Hybrid Particle Swarm Optimization) is a new and innovative clustering technique addressed in [2] which combines features of partitioned and hierarchical clustering techniques and proved to be very efficient and powerful for performing hierarchical clustering. It employs the swarm intelligence of ants in a decentralized environment.

ShiYao Liu et al. [11] researched similarity-based methods of clustering. They adopted weights into those methods so that priorities can be assigned. Then they used all this integration for cluster ensembleing with experimenting on real world data sets. According to authors results are proven to be valid and advantageous than other approaches.

Malay K. Pakhira et al. [12] .modified K-means algorithm has been proposed that solves the empty cluster problem. This modified K-means algorithm had produced effective result and experiments had proved this method better than the traditional clustering techniques

Shaohong Zhang et al. [13] targeted ensemble problem by stating that selection of suitable cluster ensemble method for specific data in unsupervised manner becomes critical because of unavailability of true information at hand before clustering. According to authors consensus affinity of cluster ensemble helps significantly improvement for ensemble solution selection and even for partition selection.

Wang et al. [14]. He focuses in his paper on introducing some novel criteria for determining the number of clusters. This new selection criterion measures the quality of clustering's through their instability from sample to sample. Here the clustering instability is estimated through cross validation, and the goal of the method is to minimize the instability. The data is divided into two training sets and one validation set to imitate the definition of stability. Then, a distance based clustering algorithm is applied on the independent and identically distributed training sets and the inconsistencies evaluated on the validation set. This method has been proven to be effective and robust on a variety of simulated and real life examples.

Von Luxburg et al. [15]. In this work, the authors propose some new definitions of stability and some related clustering notions. The results suggest that the existence of a unique minimize indicates stability, and the existence of a symmetry permuting such minimize indicates instability. The results indicate that stability does not reflect the validity or meaningfulness of the selection of the number of clusters.

Received 3 June 2018, Received in revised form, 24 June 2018, Accepted 25July 2018, Available online 25 July 2018
ISSN: 2321-1156 Volume VI Issue IV, July 2018

Instead, the parameters it measures are independent of clustering parameters.

3. EXPECT OUTCOME

Identify various challenges in the field of data mining in k-means clustering and following objective in unsupervised partitioning clustering algorithm. Proposed technique find following objectives.

1. Final optimal solution.
2. Find useful information extract in dataset and minimize cluster
3. Increase accuracy and minimize error in Extract reliable data
4. Minimize error-values in clustering and best answer.

4. CONCLUSION

Overall the goal of cluster is to be to determine the central grouping during a set of unlabeled information. Information this paper conclude that increasing efficiency of k mean algorithmic rule and Users realize higher results such as queries and execution time additionally reduced. The k-means rule is wide used for cluster massive sets of information. However the quality rule doesn't invariably guarantee smart results because the accuracy and efficiency is cut in spatial arrangement setting. Projected rule notice higher results and increasing potency queries and execution time additionally reduced. Projected rule and existing k-means cluster victimization e_coli dataset and yeast dataset and realize best solution.

5. REFERENCES

[1]. Jain a K, Murty M N and Flynn P J 1999 Data clustering: A review. ACM Computing Surveys 31(3):264–323.

[2]. Ricardo Baeza-Yates1, Carlos Hurtado1, and Marcelo Mendoza2, “Query Recommendation using Query Logs in Search Engines” , IEEE,2010.

[3]. Namrata S Gupta, Bijendra S. Agrawal, Rajkumar M. Chauhan , “ Survey on Clustering Techniques of Data Mining”, American International Journal of Research in Science, Technology, Engineering & Mathematics,pp-206-111,2015.

[4]. Sukhvir Kaur, “Survey of Different Data Clustering Algorithms”, IJCSMC, Vol. 5, Issue. 5, pg.584 – 588, May 2016.

[5]. Ankita Vimal, Satyanarayana R Valluri, Kamalakar Karlapalem (2008) “An Experiment with Distance Measures for Clustering” International Conference on Management of Data COMAD 2008,pp.241-244.

[6]. Zhihua Li, Xudong Song, Wenhui Zhu and Yanxia Chen, “K-means Clustering Optimization Algorithm Based on Map Reduce “, Proceedings of the International Symposium on Computers & Informatics (ISCI 2015), pp. 198- 203, 2015.

[7]. Anil Kumar Patidar, Jitendra Agrawal and Nishchol Mishra (2012). "Analysis of Different Similarity Measure Functions and their Impacts on Shared Nearest Neighbor Clustering Approach", International Journal of Computer Applications, Vol.40., No.16, February 2012.pp.1-5.

[8]. Dan Feldman, Melanie Schmidt, Christian Sohler, “Turning Big data into tiny data: Constant-size coresets for kmeans, PCA and projective clustering”, Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1434-1453, 2013.

[9]. Sadeghian, A. H. and Nezamabadi-pour H., “Gravitational ensemble clustering,” Proceeding of IEEE Iranian Conference on Intelligent Systems (ICIS), pp. 1-6, 2014.

[10]. Agrawal, Rakesh, Gehrke, Johannes, Gunopulos, Dimitrios, Raghavan and Prabhakar, “Automatic subspace clustering of high dimensional data”, Data Mining and Knowledge Discovery (Springer Netherlands) Vol. 11, pp. 5-33, DOI:10.1007/s10618-005-1396-1, 2005

[11]. Abu-Jamous, B, RuiFa, Nandi A.K., Roberts D.J., “Binarization of Consensus Partition Matrix for ensemble clustering,” Proceedings of the 20th European Signal Processing Conference (EUSIPCO), pp. 2193 – 2197, 2012.

[12]. Malay K. Pakhira, “A Modified k-means Algorithm to Avoid Empty”, International Journal of Recent Trends in Engineering, Vol. 1, No. 1, pp. 220-226, 2009.

[13]. Carl Meyer, Shaina Race and Kevin Valakuzhy, “Determining the Number of Clusters via Iterative Consensus Clustering,” August 6, 2014.

[14]. Wang, J., Consistent selection of the number of clusters via cross-validation. Biometrika, 2010.

[15]. Ben-David, S., Von Luxburg, U. & Pal, D., A sober look at stability of clustering. In Proc. 19th Ann. Conf. Learn. Theory (COLT 2006), Ed. G. Lugosi and H. Simon, pp. 5–19. Berlin: Springer, 2006.