International Journal of Innovative Research in Technology & Science
ISSN: 2321-1156 Volume VII Issue III, May 2019

Improve performance of Dataset Classification Method Using K-Means Clustering: Review

Tejaswini Priya [1], Prof. Surendra Chadokar[2], Department of CSE, LNCTS, Bhopal, India
priyatejaswini7@gmail.com, surendrachadokar1984@gmail.com

Abstract: *Data mining has created an excellent progress in recent year however the problem of missing information has remained an excellent challenge for data processing algorithms. It's an activity of extracting some helpful data from an oversized information base, by exploitation any of its techniques. Data processing is employed to find data out of information and presenting it during a type that's simply understood to humans. Data processing is that the notion of all ways and techniques which permit analyzing terribly giant information sets to extract and find out previously unknown structures and relations out of such large a lot of details. This paper studied the category speciation and cluster techniques on the idea of algorithms that is used to predict previously unknown class of objects. Various efforts are created to enhance the performance of the K-means cluster formula. During this paper we've been briefed within the type of a review the work administered by the various researchers using K-means cluster. They need mentioned the restrictions and applications of the K-means cluster formula in addition. Determination these problems is that the subject of the many recent analysis works. During this paper, they will be doing a review on k-means cluster algorithms*.

Keywords: *Data Mining, Classification, Clustering, K-mean Method, clusters and data mining.*

I. INTRODUCTION

The purpose of information mining technique is to mine information from a large data set and build over it into an inexpensive type for supplementary purpose. Data processing is additionally called the information discovery in databases (KDD). Technically, data {processing} is that the process of finding patterns among range of fields in large electronic information service. It's the most effective method to differentiate between information and data. Data processing consists of extract, transform, and cargo dealing information onto the information warehouse system, Store and manages the information in a very info system. But information can't be used directly. Its real worth is expected by extracting data helpful for call support. In most areas, information analysis was traditionally a manual method. Once the scale of information manipulation and exploration goes on the far side human capabilities, individuals search for computing technologies to modify the method. Data {processing} is process of extraction, transformation and loading of knowledge to from information or warehouse system. Storing and managing information, give access to information analyst [1]. The goal of cluster is to cluster information points that are close (or similar) to every alternative establish such groupings (or clusters) in an unsupervised manner. Varied definitions of a cluster will be developed, reckoning on the target of cluster. Generally, one could settle for the read that a cluster may be a cluster of objects that are a lot of kind of like one another each alternative than to members of other clusters. The term "similarity" ought to be understood as mathematical similarity, measured in some well-defined sense. In metric areas, similarity is usually outlined by means that of a distance norm. Distance will be measured among the information vectors themselves, or as a distance from a knowledge vector to some prototypal object (prototype) of the cluster. The prototypes are typically not better-known beforehand, and are wanted by the cluster algorithms at the same time with the partitioning of the information. The prototypes could also be vectors of an equivalent dimension because the information objects; however they'll even be outlined as "higher-level" geometrical objects, like linear or nonlinear subspaces or functions. The concept of information grouping, or cluster, is easy in its nature and is about to the human approach of thinking; whenever we are conferred with an oversized quantity of information, we tend to typically tend to summarize this large range of information into a little range of teams or classes so as to further facilitate its analysis. Cluster analysis is done by finding similarities between information in line with the characteristics found within the information and grouping similar information objects into clusters. Cluster is AN unsupervised learning method. Cluster is beneficial in many searching pattern analysis grouping deciding and machine learning things together with data processing, document retrieval image segmentation and pattern classification. The term cluster is used in many analysis communities to explain ways for grouping of unlabelled information. These communities have completely different terminologies and assumptions for the parts of the cluster method and also the contexts during which cluster are used. Samples of cluster are crusty and cluster genes, market segmentation. Typical pattern cluster activity involves the subsequent steps pattern illustration, Definition of pattern proximity live acceptable to the information domain, Clustering or grouping information abstraction if required and assessment of output if required. Pattern illustration refers to the quantity of categories the quantity out there of accessible patterns and also the number sorts and scale of the options available to the cluster algorithmic rule Feature choice is that the method of distinguishing the foremost effective set of the initial. Feature extraction is that the use of 1 or additional transformations of the input options to supply new salient options. Either or each of those techniques is accustomed acquire an acceptable set of options to use in cluster. Pattern proximity is typically measured by a distance operate outlined on pairs of patterns cluster. Information abstraction is that the

method of extracting an easy and compact illustration of information set. Within the cluster context a typical information abstraction could be a compact description of every cluster sometimes in terms of cluster prototypes or representative patterns like the centroid[2].K-Means cluster K-means cluster is most generally used cluster rule that is employed in several areas like data retrieval, pc vision and pattern recognition. K-means cluster assigns n information points into k clusters in order that similar information points are classified along. It's a reiterative methodology that assigns every purpose to the cluster whose centroid is that the nearest. Then it once more calculates the centroid of those teams by taking its average. The rule the essential approach of K-means cluster [4].

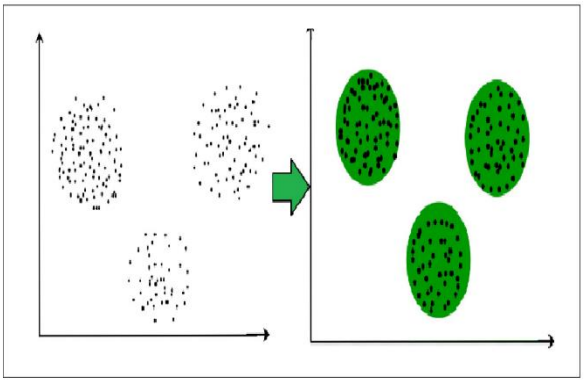

Figure1 clustering processes get three clusters

Graphical illustration for operating of K-means rule. Within the start there are 2 sets of objects. Then the centroids of each set are determined. In keeping with the centroid once more the clusters are shaped that gave the various clusters of dataset. This method repeats till the simplest clusters are achieved K-means bunch technique is wide used cluster rule, which is most well-liked bunch rule that's utilized in scientific and industrial applications. it's a technique of cluster analysis that is employed to partition N objects into k clusters in such the simplest way that every object belongs to the cluster with the closest mean. The normal Means rule is extremely easy .

1. Choose the worth of K i.e. Initial centroids.
2. Repeat step three and four for all information points in dataset.
3. Realize the closest purpose from those centroids within the Dataset.
4. Kind K clusters by assignment every purpose to its nighest centroid.
5. Calculate the new world centroid for every cluster.

Properties of k-means rule

1. efficient whereas process giant information set.
2. It works only on numeric values.
3. The shapes of clusters are lent form.

K-means is that the most typically used partitioning rule in cluster analysis owing to its simplicity and performance. However it's some restrictions once coping with terribly giant datasets owing to high machine quality, sensitive to outliers and its results depends on initial centroids that are hand-picked at random. Several solutions are planned to enhance the performance of K-Means. However nobody gives a world resolution. a number of planned algorithms are quick however they fail to keep up the standard of clusters. Some generate clusters of excellent quality however they're terribly pricey in term of machine quality. The outliers are major drawback which will result on quality of clusters. Some rule only works on only numerical datasets [5].

II.RELATED WORK

Yan Zhu et al. [6] has proposed a new method in which clustering initialization has been done using clustering exemplars produced by affinity propagation. They have also minimized the total squared error g the clusters.

S.Poonkuzhali et al. [7] propose a framework for an effective retrieval of medical records using data mining techniques. Their work focuses on retrieval of updated, accurate and relevant information from Medline datasets using Machine Learning approach. The proposed work uses keyword searching algorithm for extracting relevant information from Medline datasets and K-Nearest Neighbor algorithm (KNN) to get the relation between disease and treatment

G. Liu et al. [8] has presented a general K-means clustering to identify natural clusters in datasets. They have also shown high accuracy in their results

S. K. Wasan et al. [9] examine the impact of data mining techniques, including artificial neural networks, on medical diagnostics. They identify a few areas of healthcare where data mining and statistics can be applied to healthcare databases for knowledge discovery.

A K Dogra et al. [10] Data mining has made a great progress in recent year but the problem of missing data has remained a great challenge for data mining algorithms. It is an activity of extracting some useful knowledge from a large data base, by using any of its techniques. Data mining is used to discover knowledge out of data and presenting it in a form that is easily understood to humans. Data mining is the notion of all methods and techniques which allow analyzing very large data sets to extract and discover previously unknown structures and relations out of such huge heaps of details. This paper studied the classification and clustering techniques on the basis of algorithms which is used to predict previously unknown class of objects.

K. B. Sawan et al. [11] existing K-means clustering algorithm has a number of drawbacks. The selection of initial starting point will have effect on the results of number of clusters formed and their new centroids. Overview of the existing methods of choosing the value of K i.e. the number of clusters along with new method to select the initial centroid points for the K-means algorithm has been proposed in the paper along with the modified K-Means algorithm to overcome the deficiency of the classical K-means clustering algorithm. The new method is closely related to the approach of K-means clustering because it takes into account information reflecting the performance of the algorithm. The improved version of the algorithm uses a systematic way to find initial centroid points which reduces the number of dataset scans and will produce better accuracy in a smaller number of iteration with the traditional algorithm. The method could be computationally expensive if used with large data sets because it requires calculating the distance of every point with the first point of the given dataset as a very first step of the algorithm and sort it based on this distance. However this drawback could be taken care by using multi-threading technique while implementing it within the program. However further research is required to verify the capability of this method when applied to data sets with more complex object distributions.

Junatao Wang et al. [12] propose an improved means algorithm using noise data filter in this paper. The shortcomings of the traditional k-means clustering algorithm are overcome by this proposed algorithm. The algorithm develops density-based detection methods based on characteristics of noise data where the discovery and processing steps of the noise data are added to the original algorithm. By pre-processing the data to exclude these noise data before clustering data sets the cluster cohesion of the clustering results is improved significantly and the impact of noise data on k-means algorithm is decreased effectively and the clustering results are more accurate

D. T. Pham et al [13] has worked on the number of k used in K-means clustering. They have concluded different number of clusters for different datasets.

M. P. Sebastian et al. [14] proposes k-means algorithm, for different sets of values of initial centroids, produces different clusters. 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

G. Sahoo et al. [15] focused on K-Means initialization problems. The K-Means initialization problem of algorithm is formulated by two ways; first, how many numbers of clusters required for clustering and second, how to initialize initial centers for clusters of K-Means algorithm. This paper covers the solution for of the initialization problem of initial cluster centers. For that, a binary search initialization method is used to initialize the initial cluster points i.e. initial centroid for K-Means algorithm Performance of algorithm evaluated using UCI repository datasets.

III.EXPECT OUTCOME

In research in area of data mining-based clustering method using improve performance of dataset classification method using k-means clustering and find minimum error of medical health care dataset analysis and exceptional best solution.

IV. CONCLUSION

In study space of information mining aim at the problem of the information classification methodology of the classical K-means rule, this paper proposes the tactic of optimizing the initial cluster center to enhance the K-means rule, and exploitation the genetic rule to wash the information. The experimental results show that the projected methodology is additional correct than the classical information classification methodology. Analyses show that it's terribly tough to call one data processing rule because the best suited for the identification and/or prognosis of all diseases. Looking on concrete situations, sometime some algorithms perform higher than others, however there are cases once a mix of the simplest properties of a number of the same algorithms results simpler. During this paper, they need created a survey on work distributed by completely different researcher's exploitation K-means cluster approach. They additionally mentioned the evolution, limitations and applications of K-means cluster rule. It's determined that plenty of improvement has been created to the operating of K-means rule within the past years. Most work distributed on the development of efficiency and accuracy of the clusters. This field is usually open for enhancements. Setting applicable initial variety of clusters is usually a difficult task. At the tip it's all over that though there has been created lots of work on K-means cluster approach.

REFERENCES

[1]. E. A. Khadem, E. F. Nezhad, M. Sharifi, "Data Mining: Methods & Utilities", Researcher2013; 5(12):47-59. (ISSN: 1553-9865).

[2]. Fayyad, U. M. , Piatetsky-Shapiro, G., Smyth, P., Uthurusamy , R. G. R.: Advances in Knowledge Discovery and Data Mining. AAAI Press / The MIT Press, Menlo Park, CA. (1996)

[3]. E. Kijsipongse, S. U-ruekolan, "Dynamic load balancing on GPU clusters for large-scale K-Means clustering, " 2012 IEEE International Joint Conference on Computer Science and Software Engineering, vol., no., pp.346, 350, May 30 2012-June 1 2012.

[4]. M. Li and al. "An improved k-means algorithm based on Map reduce and Grid", International Journal of Grid Distribution Computing, (2015)

[5]. Saroj, Tripti Chaudhary, "Study on Various Clustering Techniques", International Journal of Computer Science and Information Technologies, Volume 6, Issue 3, 2015.

[6]. Y. Zhu, J. Yu, C. Jia, "Initializing K-means Clustering Using Affinity Propagation, " Ninth International Conference on Hybrid Intelligent Systems, 2009. HIS '09. vol.1, no., pp.338, 343, 12-14 Aug. 2009.

[7]. S. Poonkuzhali, T. Sakthimurugan, An Effective Retrieval of Medical Records using Data Mining Techniques, International Journal Of Pharmaceutical Science And Health Care. ISSN: 2249-5738. 2(2) (2012), pp 72-78

[8]. G. Liu ,Y. Sun; K. Xu, "A k-Means-Based Projected Clustering Algorithm, " 2010 Third International Joint Conference on Computational Science and Optimization (CSO), vol.1, no., pp.466, 470, 28-31 May 2010.

[9]. S.K. Wasan, V. Bhatnagar , H. Kaur, The Impact Of Data Mining Techniques On Medical Diagnostics, Data Science Journal, Volume 5, (2006) pp. 119-126

[10]. A K Dogra, TanujWala, "A Review Paper on Data Mining Techniques and Algorithms", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 5, May 2015.

[11]. Kedar B. Sawant, "Efficient Determination of Clusters in K-Mean Algorithm Using Neighborhood Distance "International Journal of Emerging Engineering Research and Technology Volume 3, Issue 1, January 2015.

[12]. Junatao Wang, Xiaolong Su,îAn Improved K-means Clustering Algorithm, Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on 27 may,2011 (pp. 44-46)

[13]. D. T. Pham, S. S. Dimov, and C. D. Nguyen, "Selection of K in K-means clustering", Proc. IMechE Vol. 219 Part C: J. Mechanical Engineering Science, IMechE 2005.

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

[15]. Yugal Kumar and G. Sahoo, "A New Initialization Method to Originate Initial Cluster Centers for K-Means Algorithm", International Journal of Advanced Science and Technology Vol.62, (2014).