International Journal of Current Trends in Engineering & Technology
www.ijctet.org, ISSN: 2395-3152
Volume: 06, Issue: 02 (March-April, 2020)

A SURVEY PAPER ON IMPROVEMENT OF DATA ANALYSIS BASED ON K-MEANS ALGORITHM WITH INCOMPLETE DATA CLUSTERING

Priyanka Verma, Khushbu Rai
Computer Science and Engineering Department,
LNCTS, Bhopal, India

Abstract:- *Data mining might be a way of extracting desired and helpful data from the pool of data. Clustering processing is that the grouping of data points with some common similarity. Cluster may be a vital aspect of data mining. It simply clusters the knowledge sets into given no. of clusters. Various numbers of the way is used for the knowledge cluster among that K-suggests that's that the foremost generally used cluster formula. During this paper, they briefed within the type of review work done by completely different researcher's victimization K-means cluster formula. As a partition based cluster algorithm, K-Means is wide employed in several areas for the choices of its efficiency and easily understood. However, it's documented that the K-Means algorithm could get suboptimal solutions, depending on the choice of the initial cluster centres. During this paper, they propose a projection-based K-Means initialization formula. The planned formula initially uses standard mathematician kernel density estimation techniques to look out the extremely density information areas in one dimension. Then the projection step is to iteratively use density estimation from the lower variance dimensions to the upper variance ones till all the size units of measurement computed. Experiments on actual datasets show that our technique will get similar results compared with different standard ways with fewer computation tasks. Proposed clustering algorithm to decrease error in dataset analysis and reduce clusters data and reduce similarity types of data in clusters. Therefore discovering the distribution pattern of data becomes tough using a projected vector performs cluster algorithm. The proposed method decreases error and improves the accuracy of the dataset, implementation using Mat lab software, and overcome challenges in the varied application of knowledge mining and implementation of the clustering method.*

Keywords: data processing, Unsupervised Learning, Clustering, data processing Techniques, K-means Clustering, Intelligent Data Analysis, Error Rate, Data Analysis.

I. INTRODUCTION

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems, data processing consists of extract, transform, and cargo dealings information onto the knowledge warehouse system processing includes the anomaly detection, association rule learning, classification, regression, summarization, and cluster. Processing is one of the foremost vital analysis fields that are due to the expansion of every component and package technologies that has imposed organizations to depend heavily on these technologies. Processing ideas and methods could also be applied in varied fields like promoting, medicine, property, client relationship management, engineering, web mining, etc. varied cluster algorithms according to totally different techniques are designed and applied to numerous processing issues successfully. During this paper, bunch analysis is completed by exploitation easy k mean cluster and changed k mean cluster. Standardization and classification may be a vital preprocessing step in to standardize the values of all variables from dynamic vary into specific vary. Cluster analysis is a kind processing technique that's wont to obtain information segmentation and pattern info. By cluster, the knowledge individuals get the knowledge distribution, observe the character of each cluster, and build an additional study on explicit clusters. The aim of cluster analysis is that the objects during a cluster need to be reasonably like one another and totally different from the objects in other groups. Bunch is much higher once there's larger similarity at intervals a gaggle and bigger the excellence between the groups. Thus we'll say that information possesses to be used with the rule to extract helpful data from it. Varied bunch algorithms, according to totally different techniques, are designed and applied to numerous processing issues successfully. Describes the numerous processing techniques that allow extracting unknown relationships among the knowledge things from massive data assortment that are helpful for deciding. The wide-spread use of distributed data systems finishes up within the development of big information collections in business, science, and on the online [1]. These information collections contain a wealth of knowledge, that but has got to be discovered.

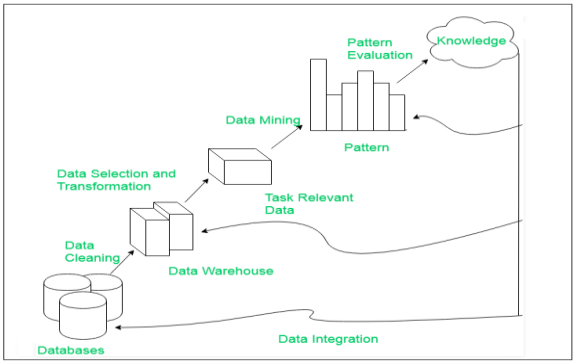

Fig1: Steps in the KDD process

Businesses will learn from their dealings information additional regarding the behaviour of their customers and thus will improve their business by exploiting this information. Science will get from data-based information (e.g., satellite data) new insights on analysis queries. Internet usage data is analyzed and exploited to optimize data access. Therefore processing generates novel, unknown interpretations of data [2]. In recent years, there's a tremendous increase in the usage of the web. The usage of the web generates much information. This information is gaining its size because the year passes. The knowledge is generated at a record rate on a day today. To research that information and cluster into a cluster is a tedious task. The matter additionally lies in storing and retrieving of data. The analysis of these information points into a totally different cluster is additionally a difficult task. Researchers have calculable that the quantity of knowledge within the globe doubles for every twenty months. But data can't be used directly. Its real worth is predicted by extracting data helpful for call support. In most areas, information analysis was historically a manual method. Once the dimension of data manipulation and exploration goes on the far side human capabilities, individuals explore for computing technologies to change the strategy [3].

Clustering

Clustering might be a way of grouping information objects into disjointed clusters so as that the knowledge within an equivalent cluster are similar; however, information happiness to require issuing completely different cluster differ. A cluster is collections of a data object that are like one another are in the same cluster, and dissimilar to the objects are in other clusters. The demand for organizing the sharp increasing information and learning valuable data from information that creates agglomeration techniques are widely applied in several application areas like AI, biology, client relationship management, information compression, processing, data retrieval, image process, machine learning, marketing, medicine, pattern recognition, psychology, statistics than on. Cluster analysis might be a tool that's accustomed observes the characteristics of the cluster and focusing on a specific cluster for any analysis. Agglomeration is unattended learning and do not place confidence in predefined categories. In agglomeration, we tend to measure the unsimilarity between objects by activity the space between every combination of objects. These measures embrace the Euclidian [4]. It's centroid based clustering during which information points split into k partition, and each partition represents a cluster. Completely alternative ways of partitioning cluster are k-means, bisecting k-means technique, and thus the Probabilistic centroid and FCM. K-means cluster technique is often a way of the cluster that's widely used. This formula is that the foremost popular cluster tool that's utilized in scientific and industrial applications. It's a way of cluster analysis that aims to partition 'n' observations into k clusters throughout which each and each observation belongs to the cluster with the nearest-mean. K-means rule might be a knowledge processing rule that performs cluster. It divides the knowledge set into the type of teams specified similar items comprise the same teams .K suggests that takes the quantity of desired clusters like num cluster=4 and initial suggests that as input by using k suggests that ++ methodology. Euclidian distance is chosen as distance operates. It's a repetitive method for cluster the dataset.

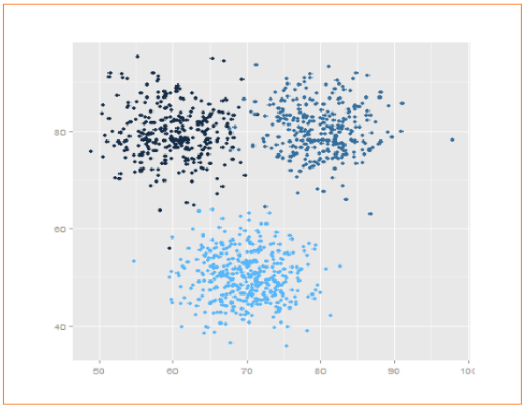

Fig2: three clusters in clustering

II.RELATED WORK

Shafeeq et al. [5] present a changed K-means algorithm to spice up the cluster quality and to fix the optimum range of clusters. As an input range of clusters (K) given to the K-means algorithm by the user. However, within the sensible state of affairs, it's terribly tough to repair the number of clusters before. The strategy projected during this paper works for every case i.e., for a celebrated range of clusters before likewise as the unknown range of clusters. The user has the flexibleness either to fix the range the number of clusters or input the minimum number of clusters needed. The new cluster centres are computed by the algorithm by incrementing the cluster counters by one in every iteration until it satisfies the validity of cluster quality. This algorithm can overcome this drawback by finding the optimum range of clusters on the run.

Soumi Ghosh et al. [6] present a comparative discussion of two cluster algorithms, particularly the centre of mass-based K-Means and representative object-based FCM (Fuzzy C-Means) cluster algorithms. This discussion is on the premise of performance analysis of the potency of cluster output by applying these algorithms.

F.A. Ramadan et al. [7] propose a cheap increased k-means algorithm to beat issues in existing k-means. Original means is understood due to its ease, simplicity, speed of convergence and adaptability to thin information. In spite of its sizable amount of advantages, it suffers from sure disadvantages. These drawbacks are the formatting of centroids, problem to converge to a

native minimum, updating of centroids until the native minimum isn't fount & execution of recurrent whereas loops of these issues are handled by the projected k-means cluster algorithm the improved algorithm first assigns datasets to its highest centre of mass then compute distance with different centroids. In next step, the two distances are compared, and if the new distance tiny is little than the previous distance, then the information point is touched to new cluster otherwise it is small then it's allotted to the same cluster. This method can save an excellent deal of some time and improve the potency. This algorithm uses two new functions. The first one is the distance () perform that's accustomed compute distance between every information and it's nearest cluster head. The second is distance new () perform accustomed compute the distance between data points and different remaining clusters. The experimental result shows that the improved k-means algorithm is way quick and economical than the primary k-means.

Binu Thomas et al. [8] gave a comparative analysis between the k-means cluster algorithm and fuzzy cluster algorithm. During this paper, the researcher additionally discusses the advantages and limitations of fuzzy c-means algorithms means could also be a partial primarily based cluster algorithm, whereas Fuzzy c-means is known partial based cluster algorithmic program. Fuzzy c-means principally works in 2 methods. Within the initial method, cluster centres are calculated, and in second, the knowledge points are assigned to the calculated cluster centre with the help of Euclidian distance. This method is almost a bit like typical k-means with a slight distinction. In fuzzy c-means algorithm membership worth starting from zero to at least one is assigned to knowledge item in cluster.0 membership indicates that the knowledge purpose isn't a member of cluster whereas one indicates the degree thereto information represents a cluster. The matter round-faced by fuzzy c-means algorithm is that the ad of membership worth of data points in every cluster is restricted to at least one Algorithm conjointly face drawback in addressing outliers. On the other hand comparison with k-means shows that the fuzzy algorithm is economical in getting hidden patterns and knowledge from natural data with outlier points.

Kohei Arai et al. [9] have projected hierarchical k-means which mixes k-means and hierarchical algorithm. The strategy executes k-means for a couple of mounted ranges of times then apply the hierarchical algorithm on centroids obtained as a result of executions of k-means. The centroids, therefore, obtained from the hierarchical algorithm are then used as initial centroids for K-means. However, authors have recommended that their technique works higher (in terms of speed) as compared to ancient k-means for advanced cluster task (large numbers of data set and much of dimensional attributes)

T. Gonzalez et al.[10] technique picks up an initial centre of mass arbitrarily and, also, the remaining centroids are selected because the knowledge that has the simplest minimum-distance to the antecedently designated centroid. This system was originally developed as a 2-approximation to k-centre cluster drawback.

Ismail Bin Mohamad et al.[11] applying standardization before cluster finishes up in higher quality, economical and proper cluster result. The author has experimented on min-max, z-score and decimal scaling techniques and complete that among the three techniques, z-score provides the best result for infectious diseases dataset with improved accuracy over ancient k-means. But the author has commented that the selection of standardization technique needs to be tired in accordance with the character of the chosen dataset.

Rakesh Kumar et al. [12] planned a ranking mechanism that uses the various ratings of a review and calculates the mixture score of the product. The ranking of various product is completed by suggests that of their reviews rating through rank selection methodology. The planned product-ranking approach victimization reviews rating establish the very best list of product and facilitate the client in selecting the only product during this framework, and the collected information is preprocessed and transformed for feature choice when omitting the unimportant options the classification method train the knowledge set to induce the last word model. Currently, the ranking approach picks the very best k-products. The planned approach considerably reduces the user time in choosing the right product

Utkarsh Gupta et al. [13] planned a singular recommender system supported a hierarchical cluster formula. The Item specific or user-specific data is assessed into a set of clusters victimization hierarchical cluster formula mentioned as Chameleon. Following this, a system is used to predict the rating of a specific item given by users. The tactic started with the set of users with their options, supported that cluster is completed victimization ranked cluster formula. Then for a given item and a user, the mapping is completed to predict rating The prediction is completed by mapping a user into a specific cluster then selection theme is applied for all user present therein cluster for the actual item. The performance of the Chameleon based recommender system is evaluated by comparing it with an existing technique supported K-means cluster formula. The results showed that the Chameleon based recommender system primarily reduce errors considerably as compared to K-means based Recommender System. The dataset used might be a picture rating dataset with a sample of 80k ratings with data regarding users and things. Type of users is 943, with feature set (age, gender, occupation, pin code). The number of things is 1682 with

feature set (release year, picture type). The planned approach is best than the prevailing K-Means based approach in terms of low Mean Absolute Error.

Joy deep Das et al. [14] present a Recommender System supported information cluster techniques. This approach affects the quantifiability drawback associated with the recommendation task. Totally different vote systems' algorithms are used to mix opinions from multiple users for recommending things of interest to the new user. During this work, authors used the DBSCAN cluster formula for bunch the users. Depending on the cluster thereto the item belongs vote algorithms suggest things to the user. The thought behind this approach is "clusters –then apply to vote" that partitions the users of the RS into teams then apply the recommendation formula one by one to each cluster. The planned system recommends an item to a user of a cluster supported rating statistics of the other users of that cluster. This approach avoids computations over the entire information, limits it to the targeted information and reduces the amount of your time of the formula. The formula is tested on the Netflix prize dataset. Netflix with 17770 rating files specified one per picture is taken under consideration. The image rating file consists of the rating data with the attribute set (movie id, year of unfairness, title, average rating, genre) given by the patrons to that picture. The rating of each picture given by all the patrons is used to calculate a median rating. The system recommends, per the user's preference of picture genres for selecting the foremost well-liked things in an exceedingly cluster, a vote based formula is applied one by one to the clusters.

H. Altay Guvenir et al. [15] has planned a fresh classification formula VFI5 and has applied to the drawback of diagnosis of erythematic squalors. There are several authors WHO have used medicine dataset from UCI (University of CA at Irvine) starting from his work wherever he applied his new developed formula VFI5. This represents a thought description by a bunch of feature intervals. The classification of a fresh instance is based on a vote among the classification created by the values of each feature one by one. All training examples are processed quickly. The VFI5 formula constructs intervals for each feature from the training examples for each interval, one price and thus the votes of each category therein interval are maintained. Thus, an interval could represent many categories by sorting the vote for each category. This formula has obtained ninety six.25% of classification accuracy.

Wang et al. [16].Clustering has been intensively studied in machine learning and data processing communities. Although demonstrating promising performance in various applications, most of the prevailing clustering algorithms cannot efficiently handle clustering tasks with incomplete features which is common in practical applications to deal with this issue, and we propose a completely unique K-means based clustering algorithm which unites the clustering and imputation into one single objective function. It makes these two processes be negotiable with one another to realize optimality. Furthermore, we design an alternate optimization algorithm to unravel the resultant optimization problem and theoretically prove its convergence. The great experimental study has been conducted on nine UCI benchmark datasets and real-world applications to evaluate the performance of the proposed algorithm, and therefore the experimental results have clearly demonstrated the effectiveness of our algorithm which outperforms several commonly-used methods for incomplete data clustering.

III.EXPECT OUTCOME

The infield of data mining and determine the number of the challenge of K-Mean clustering method based on unsupervised clustering algorithm using improve performance of dataset analysis, error minimizations using medical health care dataset analysis and best solution.

IV. CONCLUSION

Clustering algorithms are very important of large data analysis process using unsupervised learning method and may be thought of as an area of an overall data processing framework. In fact, several algorithms were specifically designed to handle a number of these problems and k-means is concentrated on these problems, which may be self-addressed in the next analysis. Medical data processing will facilitate to arrange some strategies for identification and deciding activities. Data mining using k-means clustering-based cluster centre and find a minimum error of medical dataset analysis but get a suboptimal solution. In the study of some well-known algorithms concerned with data processing technique. Under the clustering techniques of knowledge mining, varied algorithms specifically like k-means, hierarchal clustering and k-mean algorithms are studied. The results are compared between KMCA and PVSM and analyzed in accordance with their efficiencies. For classification, the hybrid k-means clustering approach and PVSM formula were implemented and compared under the clustering techniques of information mining varied algorithms specifically k-mean data processing could also be a broad area that deals among the analysis of the big volume of data by the mix of techniques from several fields like machine learning, pattern recognition, statics, technology and direction system. Medical mining is one major application space wherever accuracy is very important. They've got observed an outsized kind of algorithms to perform data analysis tasks. A hybrid approach is to partition the information, avoiding the necessity to run algorithms on very large datasets. Different types of healthcare information set analysis in processing supported bunch algorithms k-means clustering approach (KMCA) and P PVSM with their much dataset

analysis and remove copy data and overcome existing technique issues. Clustering algorithmic program could also be a developed cluster scheme, autonomous of any initial circumstances and grants outstanding outcomes in terms of the whole of the square error traditional. Our projected genetic approach executes gets higher accuracy as compare to the KMCA and show in results. A planned approach is improved information optimization, and clusters size minimizes quick supported time and more iteration; however error minis and similar time providing best solutions of approximately the equal cluster error. The above work is simulated using MATLAB simulation tool.

REFERENCES

[1]. S. Anupama Kumar and M. N. Vijayalakshmi "Relevance of data mining techniques in edification sector", International Journal of Machine Learning and Computing, Volume 3, Issue 1, February 2013.

[2]. Fayyad, U. Data Mining and Knowledge Discovery: Making Sense Out of Data. IEEE Expert, v. 11, no. 5, pp. 20-25, October 1996

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

[4]. D. Napoleon, P. Ganga Lakshmi "An Efficient K-Means Clustering Algorithm for Reducing Time Complexity using Uniform Distribution Data Point" IEEE 2010, pp. 42-45.

[5]. Shafeeq, A., Hareesha, K. Dynamic Clustering of Data with Modified K-Means Algorithm, International Conference on Information and Computer Networks, vol. 27, 2012

[6]. 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.

[7]. FAHIM, SALEM A.M, TORKEY FA, RAMADAN MA "An efficient enhanced k-means clustering algorithm" Journal of Zhejiang University SCIENCE A ISSN 1009-3095 (Print); ISSN 1862-1775 (Online)

[8]. Raju G, Binu Thomas, Sonam Tobgay and Th. Shanta Kumar, "Fuzzy Clustering Methods in Data Mining: A comparative Case Analysis" 2008 International Conference on advanced computer theory and engineering, 2008 IEEE

[9]. Kohei Arai, Ali Ridho Barakbah, "Hierarchical K-means: an algorithm for centroids initialization for K-means ", Reports of the Faculty of Science and Engineering, Saga University, Vol. 36, No.1, 2007.

[10]. T. Gonzalez, "Clustering to minimize the maximum inter cluster distance". Theoretical Computer Science, Vol. 38, pp. 293–306, 1985.

[11]. Ismail Bin Mohamad, Dauda Usman, "Standardization and Its Effects on K-Means Clustering Algorithm", Research Journal of Applied Sciences, Engineering and Technology, Vol. 6, 2013.

[12]. Rakesh Kumar, Aditi Sharan, Payal Biswas, "Framework for Ranking Products Using Ranked Voting Method", 2016, Second International Conference on Computational Intelligence & Communication Technology© 2016 IEEE DOI 10.1109/CICT.2016.138,

[13]. Utkarsh Gupta and Dr Nagamma Patil, "Recommender System Based on Hierarchical Clustering Algorithm Chameleon", 2015 IEEE International Advance Computing Conference (IACC), 978-1-4799-8047-5/15/$31.00 c_2015 IEEE.

[14]. Joydeep Das, Partha Mukherjee, Subhashis Majumder, and Prosenjit Gupta, "Clustering-Based Recommender System Using Principles of Voting Theory", 2014 International Conference on Contemporary Computing and Informatics (IC3I) @2014 IEEE.

[15]. Güvenir, H., Demiröz, G., &Ilter, N. "Learning differential diagnosis of erythematosquamous diseases using voting feature intervals". Artificial Intelligence in Medicine, 13(3) 147-165, 1998.

[16]. Wang, Siwei, Miaomiao Li, Ning Hu, En Zhu, Jingtao Hu, Xinwang Liu, and Jianping Yin. "K-means Clustering with Incomplete Data." IEEE Access, 2019.