INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN TECHNOLOGY & SCIENCE

ISSN: 2321-1156 *Volume VIII Issue III, May 2020*

A SURVEY PAPER ON IMPROVING FUZZY C-MEANS CLUSTERING TECHNIQUES

Sonam Raj, Kedar Nath Singh; Department of CSE , TITS, RGPV, Bhopal, India;
sonamraj1970@gmail.com, cseknsingh@gmail.com

Abstract— *the most objective of the information mining process is to extract information from an outsized data set and transform it into a clear structure for further use. Clustering may be a main task of exploratory data analysis and data processing applications. Clustering are main task of grouping of dataset into different group of objects in such some way that objects within the same group or called a cluster are more almost like one another than to those in other groups. This paper may be a survey of recent clustering techniques for detection of non communicable diseases (NCD) like breast cancer dataset etc. These fuzzy clustering algorithms are widely studied and applied during a kind of application areas. During this paper, they're replacing standard fuzzy c-means (FCM) algorithm more error in data analysis and overcome noise sensitivity using with improved proposed algorithm.In order to enhance the efficiency of the searching process clustering techniques recommended. These are discussed with few of the foremost dominant algorithms in their respective sub domains. Finally a model is proposed at the side of various algorithms. Our proposed algorithms are determining estimated error rate of non communicable diseases dataset.*

Keywords: Clustering Techniques, data mining, Fuzzy C-Means, Machine Learning, NCD, Dataset, ER, and Datasets.

I. INTRODUCTION

Clustering may be a division of knowledge into groups of comparable objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to things of other. Representing data by fewer clusters necessarily loses certain fine details (akin to loss data compression), but achieves simplification. It represents many data objects by few clusters, and hence, it models data by its clusters. Data modeling puts clustering during a historical perspective rooted in mathematics, statistics, and numerical analysis. From a machine learning perspective clusters correspond to hidden patterns, the look for clusters is unsupervised learning, and therefore the resulting system represents a knowledge concept. Therefore, clustering is unsupervised learning of a hidden data concept. Data processing deals with large databases that impose on clustering analysis additional severe computational requirements. These challenges led to the emergence of powerful broadly applicable data processing clustering methods surveyed below. Data processing is that the exploration and analysis of huge data sets, so as to get meaningful pattern and rules. The key idea is to seek out effective way to combine the computer's power to process the information with the human eye's ability to detect patterns. The target of knowledge mining is meant for, and work best with large data sets. Data processing is that the component of wider process called knowledge Discovery from database [1]. Data processing may be a multi-step process, requires accessing and preparing data for a mining the information, data processing algorithm, analyzing results and taking appropriate action. The data, which is accessed, are often stored in one or more operational databases. In data processing the information are often mined by passing various processes [2].

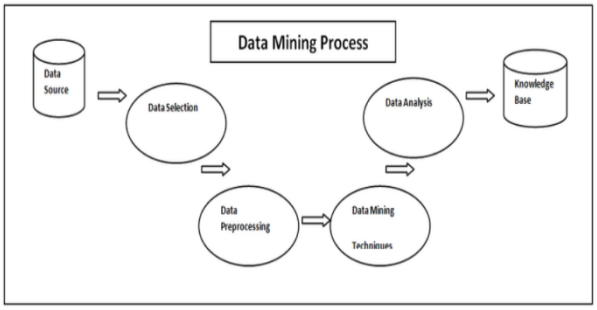


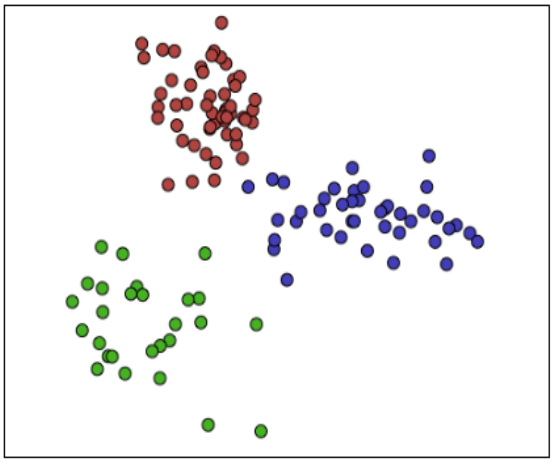
Fig 1. Data Mining Process

Clustering is usually one among the primary steps in data processing analysis. It identifies groups of related records which will be used as a start line for exploring further relationships. This system supports the event of population segmentation models, like demographic-based customer segmentation. Additional analyses using standard analytical and other data processing techniques can determine the characteristics of those segments with reference to some desired outcome. There are general kinds of clusters [3].

1. Well-separated clusters: - A cluster may be a set of points such any point during a cluster is Closer (or more similar) to each other point within the cluster than to any point not within the cluster.
2. Center-based clusters A cluster may be a set of objects such an object during a cluster is closer (more similar) to the "center" of a cluster, than to the middle of the other cluster the middle of a cluster is usually a centroid, the typical of all the points within the cluster, or a medoid, the foremost "representative" point of a cluster.
3. Contiguous clusters: - A cluster may be a set of points such some extent during a cluster is closer (or more similar) to at least one or more other points within the cluster than to any point not within the cluster.
4. Density-based clusters: - A cluster may be a dense region of points, which is separated by

Low-density regions, from other regions of high density. Used when the clusters are irregular or intertwined, and when noise and outliers are present.

Cluster Analysis :Finding groups of objects such the objects during a group are going to be similar (or related) to at least one another and different from (or unrelated to) the objects in other groups Cluster Analysis is extremely useful without proper analysis implementation of clustering algorithm won't provide good Results Cluster analysis is beneficial to know group related documents for browsing, group genes and proteins that have similar functionality, or group stocks with similar price fluctuations and also reduces the dimensions of huge data sets[4].



Classification of Clustering: Traditionally clustering techniques are broadly divided in hierarchical and Partitioning and density based clustering. Categorization of clustering is neither straightforward, nor canonical. Actually, groups below overlap.

Partitioning Methods: The partitioning methods generally end in a group of M clusters, each object belonging to at least one cluster. Each cluster could also be represented by a centroid or a cluster representative; this is often some kind of summary description of all the objects contained during a cluster. The precise sort of this description will depend upon the kind of the thing which is being clustered. just in case where real-valued data is out there, the first moment of the attribute vectors for all objects within a cluster provides an appropriate representative; alternative kinds of centroid could also be required in other cases, e.g., a cluster of documents are often represented by an inventory of these keywords that occur in some minimum number of documents within a cluster. If the amount of the clusters is large, the centroids are often further clustered to produces hierarchy within a dataset [5].

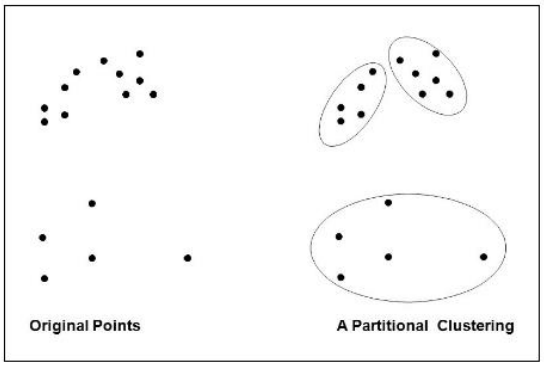


Fig 3. Different cluster analysis

The fuzzy c means algorithm also mentioned as fuzzy ISO data is one among the foremost habitually used methods in pattern recognition, fuzzy c means may be a method of clustering which follows one piece of knowledge to belong to 2 or more clusters. This method was developed by Dunn in 1973 and enhanced by Bezdek in 1981 and it's recurrently utilized in pattern recognition, classification, medical image segmentation, etc. FCM is an iterative algorithm, which is employed to seek out cluster centre that minimize a dissimilarity function. FCM uses fuzzy partition such given information can belong to many groups. To realize a high-quality classification may be a squared error clustering criterion and solutions of minimization are a minimum of squared error immobile point of J within the following equation Fuzzy portioning is administered through an iterative optimization of the target. The clustering method for both k means and FCM is same but in k means algorithm when it cluster , it takes the mean of the weighted cluster so on easy to spot masses or the origin point of cancer or tumor. In FCM, it considers that every point has weighted value related to cluster. To search out what proportion breast cancer has opened up, this system helped to doctors or radio logistic. The performance is predicated on initial cluster centers. FCM also suffers from the presence of outliers and noises in order that it's difficult to identify the initial partitions. FCM gives better results than hard k-means algorithm [6].

II. Literature Survey

This section extensively represents the research work and developments that has taken place in past years.

Ramani et al. [7] Cancer is one among the foremost leading causes of deaths among the ladies within the world. Among the cancer diseases, carcinoma is particularly a priority in women. Mammography is one among the methods to search out tumor within the breast, which is useful for the doctor or radiologists to detect the cancer. Doctor or radiologists can miss the abnormality because of inexperience's within the field of cancer detection. Segmentation is extremely valuable for doctor and radiologists to analysis the information within the mammogram. Accuracy rate of breast cancer in mammogram depends on the image

segmentation. This paper may be a survey of recent clustering techniques for detection of breast cancer. These fuzzy clustering algorithms are widely studied and applied during a kind of application areas. So as to enhance the efficiency of the searching process clustering techniques recommended. During this paper, we've presented a survey of clustering techniques.

.

Moh'd Belal et al. [8]. Clustering algorithms are utilized during a big variety of application areas. One among these algorithms is that the Fuzzy C-Means algorithm (FCM). One among the issues with these algorithms is that the time needed to converge. During this paper, a quick Fuzzy C-Means algorithm (FFCM) is proposed supported experimentations, for improving fuzzy clustering. The algorithm is predicated on decreasing the amount of distance calculations by checking the membership value for every point and eliminating those points with a membership value smaller than a threshold value. We applied FFCM on several data sets. The experiments demonstrate the efficiency of the proposed algorithm. Clustering involves dividing data points into homogeneous classes or clusters in order that points within the same cluster are similar as possible, and points in several clusters are as dissimilar as possible. In non-fuzzy or hard clustering, data is split into crisp clusters, where each datum belongs to precisely one cluster. In practice, however, there are many situations during which the info points might be classified as belonging to at least one cluster almost also on another. Such a situation can't be catered by hard clustering.

Fu. Huaiguo et al. [9] within the field of cluster analysis, most of existing algorithms assume that every feature of the samples plays a consistent contribution for cluster analysis. Feature-weight assignment may be a special case of feature selection where different features are ranked consistent with their importance. The feature is assigned a worth within the interval [0, 1] indicating the importance of that feature, we call this value "feature-weight". During this paper we propose a replacement feature weighted fuzzy c-means clustering algorithm during a way which this algorithm be ready to obtain the importance of every feature, and then use it in appropriate assignment of feature-weight. These weights incorporated into the space measure to shape clusters supported variability, correlation and weighted feature. The Goal of cluster analysis is to assign data points with similar properties to equivalent groups and dissimilar data points to different groups. Generally, there are two main clustering approaches i.e. crisp clustering and fuzzy clustering. Within the crisp clustering method the boundary between clusters is clearly defined.

Cannon et al. [10]. This paper reports the results of a numerical comparison of two versions of the fuzzy c-means (FCM) clustering algorithms. Especially , we propose and exemplify an approximate fuzzy c-means (AFCM) implementation based upon replacing the required ``exact'' variates within the FCM equation with integer-valued or real-valued estimates. This approximation enables AFCM to take advantage of a lookup table approach for computing Euclidean distances and for exponentiation. Internet effect of the proposed implementation is that CPU time during each iteration is reduced to approximately one sixth of the time required for a literal implementation of the algorithm, while apparently preserving the general quality of terminal clusters produced. The 2 implementations are tested numerically on a nine-band digital image, and a pseudo code subroutine is given for the convenience of applications-oriented readers. Our results suggest that AFCM could also be wont to accelerate FCM processing whenever the feature space is comprised of tuples having a finite number of integer-valued coordinates.

Pal et al. [11] in proposed the fuzzy-possibility c-means (FPCM) model and algorithm that generated both membership and typicality values when clustering unlabeled data. FPCM constrains the typicality values in order that the sum over all data points of typicalities to a cluster is one. The row sum constraint produces unrealistic typicality values for giant data sets. During this paper, we propose a replacement model called possibilistic-fuzzy c-means (PFCM) model. PFCM produces memberships and possibilities simultaneously, alongside the standard point prototypes or cluster centers for every cluster. PFCM may be a hybridization of possibility c-means (PCM) and fuzzy c-means (FCM) that always avoids various problems of PCM, FCM and FPCM. PFCM solves the noise sensitivity defect of FCM, overcomes the coincident clusters problem of PCM and eliminates the row sum constraints of FPCM. We derive the first-order necessary conditions for extreme of the PFCM objective function, and use them because the basis for a typical alternating optimization approach to finding local minima of the PFCM objective functional. Several numerical examples are as long as compare FCM and PCM to PFCM. Our examples show that PFCM compares favorably to both of the previous models. Since PFCM prototypes are less sensitive to outliers and may avoid coincident clusters, PFCM may be a strong candidate for fuzzy rule-based system identification.

Hung et al. [12] The Fuzzy C-Means (FCM) algorithm is usually used for clustering. The performance of the FCM algorithm depends on the choice of the initial cluster center and/or the initial membership value. If a decent initial cluster center that's on the brink of the particular final cluster center are often found, the FCM algorithm will converge very quickly and therefore the time interval are often drastically reduced. The authors propose a completely unique algorithm for efficient clustering. This algorithm may be a modified FCM called the psFCM algorithm, which significantly

reduces the computation time required to partition a dataset into desired clusters. We discover the particular cluster center by employing a simplified set of the first complete dataset. It refines the initial value of the FCM algorithm to speed up the convergence time. Our experiments show that the proposed FCM algorithm is on the average fourfold faster than the first FCM algorithm. We also demonstrate that the standard of the proposed FCM algorithm is that the same because the FCM algorithm.

Kolen et al.[13] during this paper, we present an efficient implementation of the fuzzy c-means clustering algorithm. The first algorithm alternates between estimating centers of the clusters and therefore the fuzzy membership of the info points. The dimensions of the membership matrix are on the order of the first data set, a prohibitive size if this system is to be applied to very large data sets with many clusters. Our implementation eliminates the storage of this arrangement by combining the 2 updates into one update of the cluster centers. This alteration significantly affects the asymptotic runtime because the new algorithm is linear with reference to the quantity of clusters, while the first is quadratic. Elimination of the membership matrix also reduces the overhead related to repeatedly accessing an outsized arrangement . Empirical evidence is presented to quantify the savings achieved by this new method**.**

Havens et al.[14] Very large (VL) data or big data are any data that you simply cannot load into your computer's memory . This is often not an objective definition, but a definition that's easy to know and one that's practical, because there's a dataset too big for any computer you would possibly use; hence, this is often VL data for you. Clustering is one among the first tasks utilized in the pattern recognition and data processing communities to look VL databases (including VL images) in various applications, and so, clustering algorithms that scale well to VL data are important and useful. This paper compares the efficacy of three different implementations of techniques aimed to increase fuzzy c-means (FCM) clustering to VL data. Specifically, we compare methods that are supported 1) sampling followed by no iterative extension; 2) incremental techniques that make one sequential go through subsets of the data; and 3) kernel zed versions of FCM that provide approximations supported sampling, including three proposed algorithms. We use both loadable and VL datasets to conduct the numerical experiments that facilitate comparisons supported time and space complexity, speed, quality of approximations to batch FCM (for loadable data), and assessment of matches between partitions and ground truth. Empirical results show that sampling plus extension FCM, bit-reduced FCM, and approximate kernel FCM are good choices to approximate FCM for VL data. We conclude by demonstrating the VL algorithms on a dataset with 5 billion objects and presenting a group of recommendations regarding the utilization of various VL FCM clustering schemes.

Wang et al.[15] the fuzzy c-means (FCM) is one among the algorithms for clustering supported optimizing an objective function, being sensitive to initial conditions, the algorithm usually results in local minimum results. Aiming at above problem, we present the worldwide fuzzy c-means clustering algorithm (GFCM) which is an incremental approach to clustering. It doesn't depend upon any initial conditions and therefore the better clustering results are obtained through a deterministic global search procedure. We also propose the fast global fuzzy c-means clustering algorithm (FGFCM) to enhance the converging speed of the worldwide fuzzy c-means clustering algorithm. Experiments show that the worldwide fuzzy c-means clustering algorithm can give us more satisfactory results by escaping from the sensibility to initial value and improving the accuracy of clustering; the fast global fuzzy c-means clustering algorithm improved the converging speed of the worldwide fuzzy c-means clustering algorithm without significantly affecting solution quality.

Duan et al.[16] Sensitive to the initial number and centers of clusters is one shortcoming of fuzzy c-means clustering method. Getting to reduce the sensitivity, a partial supervision-based fuzzy c-means clustering method is proposed during this paper. During this method, the information is first clustered with standard fuzzy c-means algorithm. If the clustering result doesn't accord with the structure of information, there must be one or more clusters that are wrongly separated leading to some clusters on the point of one another . The close clusters are often found by investigating the partition matrix. Those close clusters should be divided or merged. In both situations, approaches are then proposed during this new method to update the acceptable cluster number and cluster centers. With the updated cluster centers as labeled patterns, partially supervised fuzzy clustering is carried to offer the acceptable clusters. Experiments on four synthetic datasets and a true dataset show that the proposed clustering method has good performance by comparing to the quality fuzzy c-means clustering method.

III EXPECTED OUTCOME

In field of information mining find different challenges but identified better solution and optimal data. Improve data identifying and minimization error.

IV. CONCLUSION

In this paper, they found that there are approaches having their own class of work. present are several fuzzy methodologies existing like fuzzy c-means with threshold, probabilistic fuzzy c-means etc. while working on dataset analysis apportionment procedure, they found that suboptimal solution but our proposed method found better solution than traditional FCM.

ISSN: 2321-1156 Volume VIII Issue III, May 2020

Data processing is widely utilized in medical field. Healthcare providers utilize the information mining tools to form effective decision regarding the way to enhance the patient health, the way to provide health care services at low cost and the way to predict fraud in insurance etc. Healthcare area of examiner also face several challenges while using data processing in health field like several data processing techniques required parameters from user. These techniques are sensitive to user's parameters. Its results vary consistent with the parameters which are given by users. Sometime users don't have sufficient information about selection and usage of parameter. Data processing technique like separation of various classifiers, separation of clustering with classification and clustering. For achieving better data processing in this field.

REFERENCES

[1]. Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. "Data mining cluster analysis: basic concepts and algorithms." Introduction to data mining : 487-533, 2013.

[2]. Han, Jiawei, Micheline Kamber, and Jian Pei. "Data mining concepts and techniques third edition." Morgan Kaufmann, 2011.

[3]. Sahu, Hemlata, Shalini Shrma, and Seema Gondhalakar. "A brief overview on data mining survey." International Journal of Computer Technology and Electronics Engineering (IJCTEE) 1, no. 3: 114-121, 2011.

[4]. Dunn, Joseph C. "Well-separated clusters and optimal fuzzy partitions." Journal of cybernetics 4, no. 1 : 95-104, 1974.

[5]. Lin, Yung-Shen, Jung-Yi Jiang, and Shie-Jue Lee. "A similarity measure for text classification and clustering." IEEE transactions on knowledge and data engineering 26, no. 7 : 1575-1590,2013.

[6]. Windham, Michael P. "Cluster validity for fuzzy clustering algorithms." Fuzzy Sets and Systems 5, no. 2 : 177-185, 1981.

[7]. Ramani, R., S. Valarmathy, and N. Suthanthira Vanitha. "Breast cancer detection in mammograms based on clustering techniques-a survey." International Journal of Computer Applications 62, no. 11 2013.

[8]. Moh'd Belal, AL-Zoubi, Amjad Hudaib, and Bashar Al-Shboul. "A fast fuzzy clustering algorithm." In Proceedings of the 6th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases, pp. 28-32. 2007.

[9]. Fu, Huaiguo, and Ahmed M. Elmisery. "A new feature weighted fuzzy c-means clustering algorithm." Algarve, Portugal : 11, 2009.

[10]. Cannon, Robert L., Jitendra V. Dave, and James C. Bezdek. "Efficient implementation of the fuzzy c-means clustering algorithms." IEEE transactions on pattern analysis and machine intelligence 2 : 248-255, 1986.

[11]. Pal, Nikhil R., Kuhu Pal, James M. Keller, and James C. Bezdek. "A possibilistic fuzzy c-means clustering algorithm." IEEE transactions on fuzzy systems 13, no. 4: 517-530,2005.

[12]. Hung, Ming-Chuan, and Don-Lin Yang. "An efficient fuzzy c-means clustering algorithm." In Proceedings 2001 IEEE International Conference on Data Mining, pp. 225-232. IEEE, 2001.

[13]. Kolen, John F., and Tim Hutcheson. "Reducing the time complexity of the fuzzy c-means algorithm." IEEE Transactions on Fuzzy Systems 10, no. 2 (2002): 263-267.

[14]. Havens, Timothy C., James C. Bezdek, Christopher Leckie, Lawrence O. Hall, and Marimuthu Palaniswami. "Fuzzy c-means algorithms for very large data." IEEE Transactions on Fuzzy Systems 20, no. 6 : 1130-1146,2012.

[15]. Wang, Weina, Yunjie Zhang, Yi Li, and Xiaona Zhang. "The global fuzzy c-means clustering algorithm." In 2006 6th World Congress on Intelligent Control and Automation, vol. 1, pp. 3604-3607. IEEE, 2006.

[16]. [16] Duan, Lingzi, Fusheng Yu, and Li Zhan. "An improved fuzzy c-means clustering algorithm." In 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), pp. 1199-1204. IEEE, 2016.