Appraising Research Direction & Effectiveness of Existing Clustering Algorithm for Medical Data

Sudha V
Asst. Prof: Dept of Information Science & Engg.
RNS Institute of Technology
Bangalore, India

Girijamma H A
Prof.: Dept. of Computer Science & Engg.
RNS Institute of Technology
Bangalore, India

Abstract—**The applicability and effectiveness of clustering algorithms had unquestioningly benefitted solving various sectors of real-time problems. However, with the changing time, there is a significant change in forms of the data. This paper briefs about the different taxonomies of the clustering algorithm and highlights the frequently used techniques to understand the research popularity. We also discuss the existing direction of the research work and find that still there is a significant amount of open issues when it comes to clustering medical data. We find that existing techniques are quite symptomatic in nature on local problems in clustering while problems associated with complex medical data are yet to be explored by the researchers. We believe that this manuscript will give a good summary of the effectiveness of existing clustering techniques towards medical data as a contribution.**

Keywords—Medical Data; Clustering Algorithm; k-Means Clustering; Fuzzy; Classification

I. Introduction

In the list of challenges about unsupervised learning techniques, clustering is one of the biggest challenges till date [1] [2]. Clustering deals with exploring an elite structure from a given set of the raw database. Theoretically, the technique of organizing the objects into the group where the member of the group's bears certain similarity score with each other is known's as clustering. A good clustering technique always identifies the internal grouping from a given set of raw data. The user frames the effectiveness of the clustering performance. The user provides such forms of converging criterion. The applications of the clustering algorithm observed in many places e.g. biology, city planning, libraries, marketing, studying natural calamities, etc. [3]. For a clustering algorithm to be robust, needed that it should explore random-shaped clusters, should possess scalability, and should have high dimensionality. It should have better usability and interoperability characteristics along with insensitive features towards inputs. Most important, it should also have the potential to counter-measure the adverse effect of noise as well as outliers. A robust clustering algorithm can also state if it possesses the capability to manage higher and diversified number of attributes. Finally, it should have lower demands for domain knowledge in order to evaluate input attributes [4] [5]. However, there are also certain pitfalls associated with conventional clustering techniques, for example:

1) Higher dependencies of the spatial feature is the prime criteria of effectiveness (usually, such forms observed over distance-based clustering.

2) All the clustering and classification demands cannot be fulfilled using existing clustering techniques.

3) Defining a specific measure of the distance in case of multi-dimensional spaces is quite a challenging task,

4) Existing techniques suffers from problems with the larger dimension of the data owing to the greater extent of time complexity.

Although the outcomes of any clustering algorithm can have multiple inferences, it is hardly possible to even identify the correct number of outcomes for higher dimensional data. The clustering algorithms used over the various field but the applicability of the clustering in medical science is highly challenging. The input for clustering techniques could be any form of medical data, where the purpose could be anything right from segmentation to the classification of a specific disease condition. The original of medical data could have diversified forms (signal, image, dataset, wavelet, etc.). The medical images as quite different from the natural images as they captured from a specific data capturing device. Hence, their formats are very different that causes to implement specific forms of medical image processing. There is also a possibility of inclusion of the higher amount of noises and distortion that potentially affect the data quality. Hence, performing clustering of the medical data is one of the challenging problems in medical image processing.

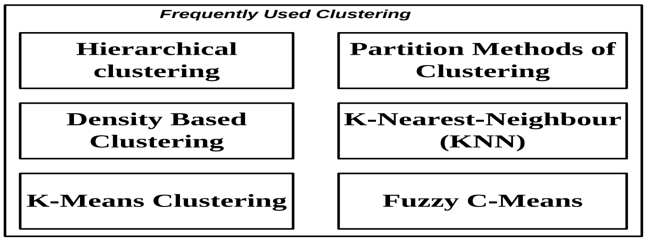
In most recent times, there has been a significant amount of research work being carried out in introducing clustering techniques using various forms of data. However, with the evolution of complex medical data capturing devices and analysis, the inputs of medical data are no simpler than ten years ago. They will be required to analyze in the perfect manner to assist in effect clustering algorithm. The prime aim of this is to present a discussion about the effectiveness of the existing clustering techniques towards medical data. The discussion has been carried out using standard research papers and its contribution towards solving clustering problems.

Section II discusses the fundamental briefing of the clustering techniques followed by existing research trends in Section III. Section IV discusses the recent techniques about clustering techniques and studying its effectiveness. The open research problems have been discussed in Section V while the summary of the work and future direction of the work is briefed in Section VI.

II. About Clustering Techniques

Clustering is a mechanism that allows the grouping of the data in the form of logical groups of certain significance [6]. One of the prime beneficial characteristics of clustering is its adaptability feature. The prime goal of the clustering algorithm is to carry out a transformation of the group of data into further meaningful data in order to ensure that the data residing in the similar group or cluster offers certain logic [7] [8]. The majority of the clustering algorithms aims to reduce the distance between two similar clusters (intra-cluster distance) and increase the distance between two different clusters (inter-cluster distance) [9]. The mechanism of clustering also termed as data segmentation owing to its characteristics of differentiating objects that also results in the identification of outliers. The usage of clustering observed in various fields e.g. machine learning, pattern recognition, false detection, analysis of the business market, etc. In the majority of the analysis, clustering tree is represented using dendrogram. The clustering technique is also frequently called as data mining technique using unsupervised approach applied for clustering (or grouping) data. For a given set of unlabeled data, mainly clustering technique explores the internal grouping.

As per theory, there are five types of clustering techniques i.e. i) Hierarchical methods, ii) Partitioning Methods, iii) Grid-based methods, iv) Machine Learning methods, v) algorithms for high dimensional data [10]. Hierarchical clustering methods are again divided into two types i.e. Agglomerative Algorithms and Divisive Algorithm. The partitioning methods of clustering are mainly of 5 type's viz. relocation algorithms, probabilistic clustering, k-medoids methods, k-means methods, and density-based algorithms. The density-based algorithms are further divided into density based connectivity clustering and density functions clustering. The clustering techniques using machine learning are again classified into type's viz. gradient descent and Artificial Neural Network and evolutionary methods. Finally, the clustering algorithms for high dimensional data is divided into three types i.e. subspace clustering, projection techniques, and co-clustering techniques. Although there are five types of clustering techniques, Fig.1 shows the frequently used clustering techniques practiced in existing research work.



Frequently Used Clustering

Hierarchical clustering

Partition Methods of Clustering

Density Based Clustering

K-Nearest-Neighbour (KNN)

K-Means Clustering

Fuzzy C-Means

Fig. 1. Frequently used Clustering Techniques

The frequently exercised clustering mechanisms shown in Fig.1 are briefed as follows:

- ***Overlapping Clustering:*** - Normally, such technique uses fuzzy logic over grouped data in order to incorporate fuzzy membership function for the clustered data.
- ***Exclusive Clustering:* -** This technique uses a particular technique to perform grouping of data. It ensures that a data should occupy only one cluster during the grouping operation.
- ***Probabilistic Clustering:* -** Normally, such techniques are applied for optimization techniques in order to ensure the best fit between the experimental value and framework. It uses probabilistic theory. A parametric distribution is used to represent each cluster.
- ***Hierarchical Clustering:* -** This forms of clustering normally constructs or agglomerates or performs breaking up or performs the divisive operation to form a cluster hierarchy.

The brief discussions of the different forms of the clustering techniques are as follows:

A. Hierarchical clustering

A pre-determined order of cluster is formulated either from top to bottom (divisive) or vice-versa (Agglomerative) in hierarchical clustering. It is normally represented by the dendrogram. Fig.2 shows the two forms of hierarchical clustering technique.

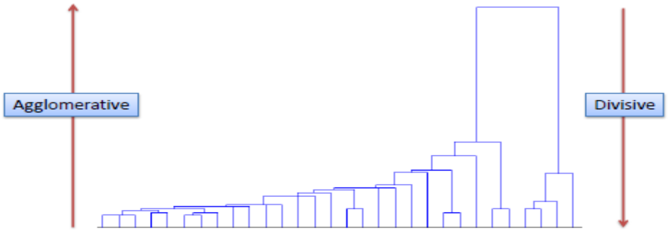


Fig. 2. Hierarchical Clustering

An agglomerative clustering initiate with the one-point group and then iteratively combines two or more precisely determined clusters. It performs the computation of all pair wise patterns for evaluating similarity coefficient. After analyzing it's each pattern in one class, it than combines the clusters to form new clusters and compute the respective distances of similarity score. This step is repeated until it ends up in k-cluster that can be one also. Similarly, divisive clustering initiates with one cluster and then iteratively divide the precise cluster. It starts its divisive operation from the top of the cluster that is distributed with the aid of flat clustering algorithm [11]. This mechanism is repeated till it reaches the singleton pattern of a cluster. Research papers use the terms called as Agglomerative Hierarchical clustering algorithm (AGNES) and Divisive Hierarchical clustering algorithm (DIANA) for agglomerative nesting and divisive analysis) respectively [12]. Both AGNES and DIANA are opposite of each other. The existing research studies that have discussed AGNES and DIANA [12]. The beneficial attributes of such algorithms are - i) simplified implementation to offer better outcomes and ii) Lesser pre-defined information about demanded number of clusters. The limitation of such techniques would be -

1) The algorithms cannot be effectively controlled to return to its prior state if required.

2) Increase of computational resources with an increase of data points.

3) Due to the form of the spatial factor selected for combining, this algorithm is witnessed with troubleshooting while splitting larger size of clusters, higher sensitivity to outliers, challenging to manage heterogeneity in clusters. In many cases, determining the precise number of clusters is highly difficult one.

B. Partition Methods of Clustering

This technique of clustering is used for partitioning database consisting of a specific number of clusters and objects. An optimization of iterative nature is used in partitioning technique between k-number of clusters. Such technique is further classified in the form of k-means as well as k-medoids approaches. Usage of *k*-means is seen in maximum research work as it is quite simple to be incorporated in a majority of research problems. It is also one of the simple algorithms for extracting the demanded cluster number using centroid. Fig. 3 shows the conventional representation of partitioning process. The technique doesn't have any pitfalls on the types of parameters that are governed by the location of the predetermined fraction of the coordinates within the cluster location. Therefore, the grouping of the nationalities by food habit as shown in Fig.3 can be easily done using k-means clustering.

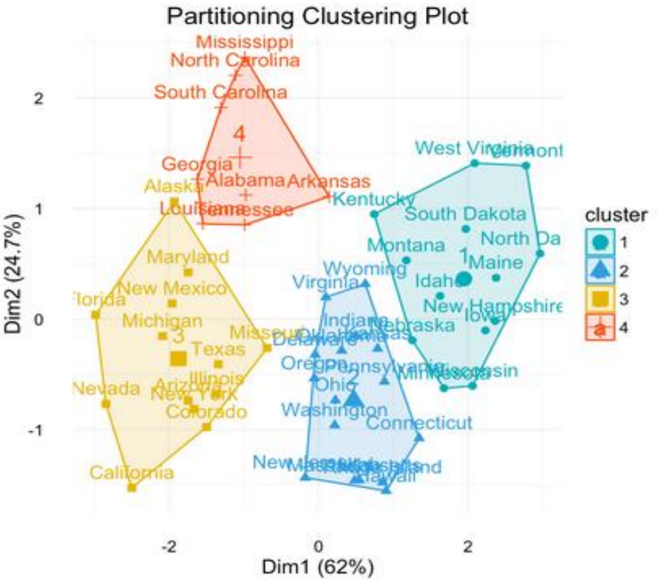


Fig. 3. Partitioning Clustering

C. Density Based Clustering

This is another frequently used clustering technique of existing system that is more inclined towards densities of the data point. This technique is more interested in exploring the random shapes cluster along with noise over a distance-based dataset. It always ensures that neighbor quantity is more than minimum data points in case the cluster is constructed. Fig.4 highlights a typical case of density-based clustering. It uses iterative processes for forming a cluster. One of the prominent pitfalls of this technique is that it cannot perform grouping of the data over the dataset of the larger dimension of differences in the cluster densities. The technique uses three different classified forms of objects e.g. classified, non-classified, and noise. A respective id of a cluster is always used for every classified object as well as noise object. However, this technique doesn't use any form of cluster id for non-classified objects. The example cited in Fig.4 shows implication of density-based clustering technique to categorize unhealthy tissue or a lesion from health tissue. It could further explore the sub-regions of different colors within the unhealthy tissue that could be again benefitted for association or classification operation. The advantage of using density-based clustering is to identify the cluster number as apriori in order comfortably manage the clusters with random dimension. However, it also suffers from the pitfalls as its inapplicability in heterogeneous densities. Moreover, its outcome highly depends on spatial measures.

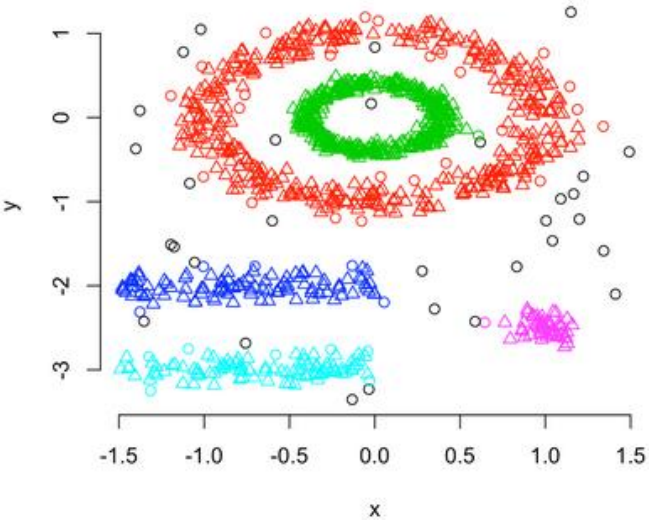


Fig. 4. Density-based Clustering

D. K-Nearest-Neighbour (KNN)

KNN algorithm is also known as memory-based clustering technique as it needs prior feeding of the samples required for training while performing processing at run time. The algorithm used in the mining operation. Different forms of the continuous parameters can be managed by the KNN algorithm although it can also work with similar capability over discrete-based properties during clustering. All the parameter in this algorithm associates distance and considers the maximum of them as far as possible. However, relationships of the parameters are not considered in this technique for computing similarity metric. This is the prime cause of errors in distance measures that significantly affects the classification accuracy. The beneficial factor of KNN algorithm is its simple implementation procedure accompanied by faster training steps. The issues in this algorithm are its dependencies of the larger database, slower validation process, and have higher noise sensitivity.

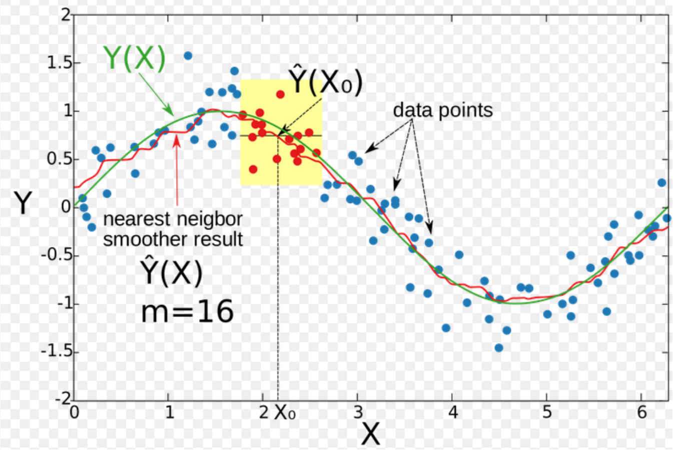


Fig. 5. KNN Clustering

E. K-Means Clustering

Usage of k-means clustering is seen in the majority of the clustering techniques. This technique is quite iterative in nature that classifies the given data in order to form k-disjoint clusters. Fig.6 shows the technique of KNN clustering for a given set of original data. The effectiveness of k-means clustering is normally assessed using squared error factor within a cluster.

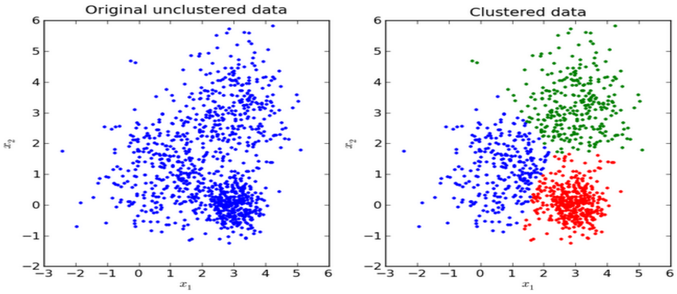


Fig. 6. K-Means Clustering

It was noted that adoption of k-means clustering forms cluster of compact form but it choose not to consider the distance between two clusters. From theoretical viewpoint, adoption of squared l_2-normalization leads to higher sensitivity in the case of maximized errors. This will eventually mean that such formulation is quite less robust from the statistical viewpoint. Only because of its simple implementation and efficiency towards computational performance, k-means algorithm is frequently used clustering technique. It also has very low memory utilization and relatively easier to understand compared to other existing clustering techniques. For distinct dataset, it offers higher precision result and offers better compactness in the cluster as compared to hierarchical clustering technique. However, it also suffers from limitations e.g. it doesn't resolve any overlapping clusters, higher dependencies of pre-determined information, random selection of clusters, the applicability only in case of presence of the mean value, and it cannot be used for outliers as well as noisy data.

F. Fuzzy C-Means

Usage of fuzzy logic over clustering has been started witnessing since last decade. Such form of the algorithm uses spatial attribute for assigning membership function mapping with the data points which is considered equivalent to the center of each group. If the nearness of the data is more towards the center of the cluster than the ability of the membership function is also more towards the cluster center. Using probabilistic approach, the sum of all the involved membership function is equivalent to 1. Fig.7 shows the mechanism of clustering in this case. The beneficial factors of using fuzzy c-means clustering are that its applicability of assigning membership functions at the center of the cluster. Moreover, fuzzy c-means algorithm is highly applicable for the dataset that is in overlapping form, and it works better as compared to a conventional k-means algorithm. However, the limitation of this technique does also exist e.g. usage of Euclidean's distance is not proportionate with the unequal weight and it involves more iterative steps. Predetermined information dependencies are another pitfall of this algorithm.

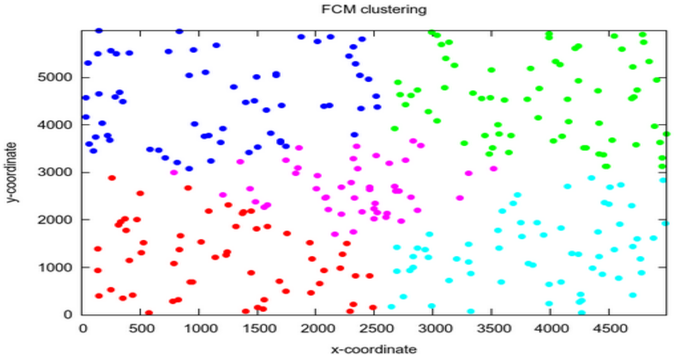


Fig. 7. Fuzzy c-Means Clustering

III. Existing Research Trend

This section discusses the existing research trends towards clustering techniques. For this purpose, we prune the research papers published between 2010 to till date from IEEE Xplore. We find that there are 14,730 conference paper and 2090 Journals associated with the problems and enhancement techniques of clustering. For an elaborated understanding, we use Fig.8 that basically furnishes two types of information i.e. i) complete classification of clustering algorithm and ii) a total number of research papers specific to each type of the clustering algorithm. It is widely known that clustering is specifically useful for performing pattern recognition, Spatial Data Analysis, Image Processing, Economic Science, document classification, data mining, etc. [13]. The survey outcome basically shows that k-means (No. of published Journal: 399, No. of Published Conference: 4226) algorithm is the highly adopted technique in classification followed by probabilistic technique (No. of published Journal: 216, No. of Published Conference: 804). Although there are other significant types of a clustering algorithm, they are less explored by the research community since 2010. Another significant trend is that all the investigation was carried out by diversified forms of the data, where maximum data is in the form of an image. There is also less specialization work of clustering towards detection and diagnosis of the complex medical condition. Most recently, there are certain standard review papers e.g. [14] [15] [16] that has reviewed over different research work being carried out over clustering techniques. But nowhere it is found how strongly clustering technique is used over medical data or any other form of complex data. With the increasing usage of the dynamic user, the formation, processing, and distribution process of such data would be quite complex to solved. Even the frequent usage of the k-means algorithm was not much seen to address the complicated problems associated with medical images. On the other hand, there has been considerable amount of work being carried out using Artificial Neural Network (No. of published Journal: 174, No. of Published Conference: 1237), Evolutionary technique (No. of published Journal: 174, No. of Published Conference: 1229) on clustering problems, probabilistic techniques (No. of published Journal: 216, No. of Published Conference: 804), and Co-clustering technique (No. of published Journal: 216, No. of Published Conference: 1723). Hence, it can be easily said that maximum research work till date from 2010 has used k-means clustering algorithm followed by the probabilistic approach, co-clustering approach, neural network, and evolutionary techniques. Apart from this, other techniques have received less attention till date. Therefore, it can be said that usage of machine learning and portioned-based clustering techniques are predominantly used in the existing system and can also be represented as existing research trends. However, the existing survey papers don't speak about predominant clustering techniques of recent time, and hence it is quite challenge to understand the effectiveness of existing clustering techniques.

The next section discusses the existing research techniques accompanied by brief highlights of existing problems, the technique adopted to solve them with associated advantages and limitation of existing techniques.

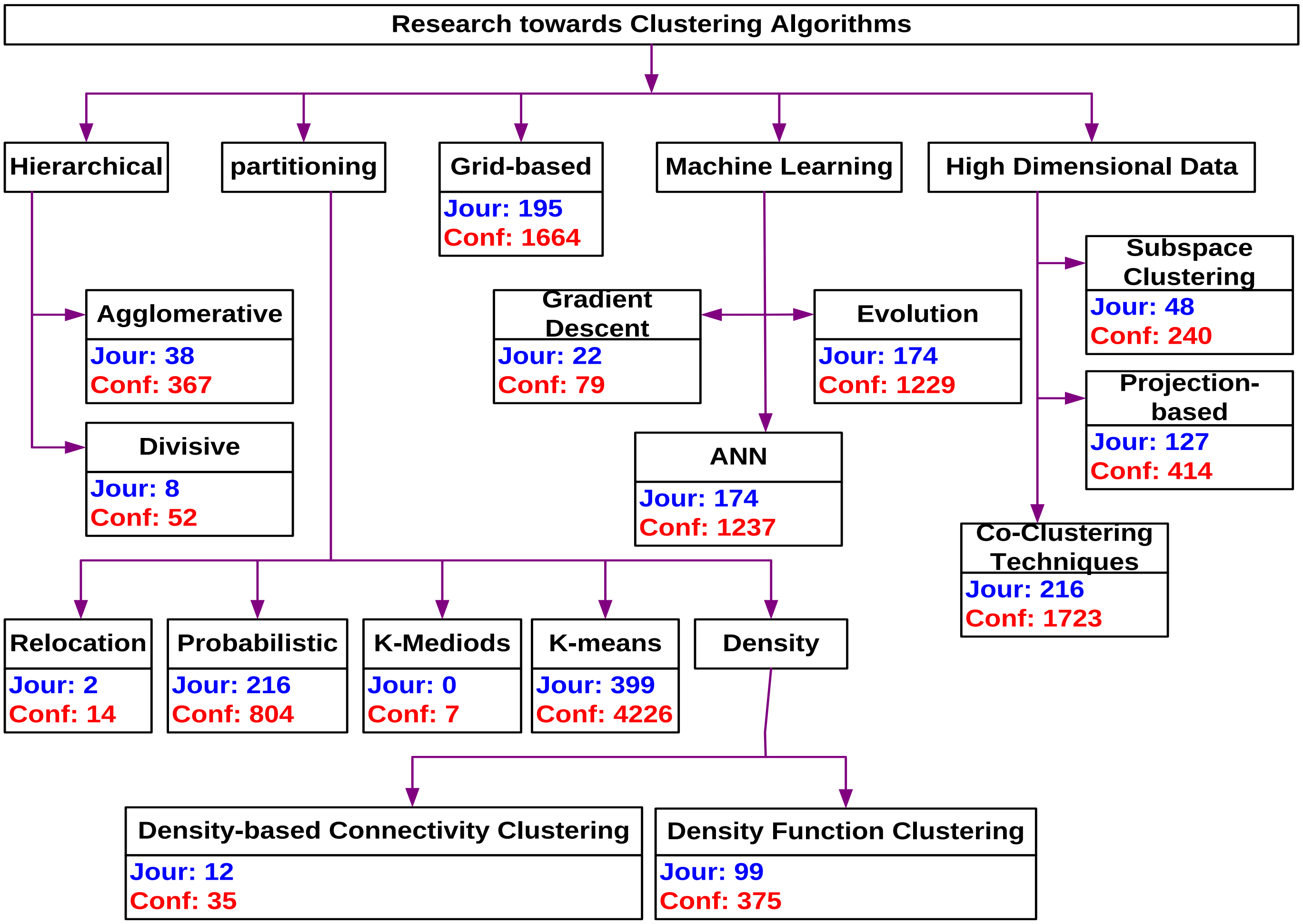


Fig. 8. Research Trend towards Clustering Algorithm

IV. Existing Research Work

This section discusses the existing research technique that has been used for enhancing clustering performance. Usage of clustering technique towards medical data is mainly associated with improving the image processing operations e.g. segmentation. The work carried out by Al-Dmour and Al-Ani et al. [17] have combinedly used unsupervised and semi-supervised classification approach in order to perform involuntary segmentation. The authors have also used the median filter as well as Fuzzy-c-means attributes for performing clustering. A technique called as subtractive clustering is used for minimizing computational complexity. Adoption of fuzzy clustering technique was also seen in the work of Proietti et al. [18] that applies membership function of kernel-based. The study claims to extract unconstrained structure. Clustering also plays a significant role in maintaining the resolution of an image. Al-Qizwini et al. [19] have used similarity of the subspace as well as manifold clustering. Applying subspace clustering assists in extracting low ranks clusters along with usage of Principal Component Analysis (PCA). Finally, training and testing are carried out on natural images where the outcomes were testified using Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). Although, clustering techniques is beneficial with the abundance of data, but could encounter a significant problem if data is incomplete. One of such investigation towards implementing clustering operation for a given set of impartial data was carried out using Li et al. [20]. The authors have used K-means clustering algorithm as well as a k-median method in order to perform clustering. The technique also performs minimax optimization technique for reduced complexities. The study outcome was assessed using numbers of wrongly estimated values for different clustering mechanism. Ahmad [21] have applied fuzzy clustering algorithm for breast cancer detection. The technique has applied fuzzy c-means clustering and applied an existing technique for computing the distance between two values of features. Clustering approaches were also studied with respect to the transfer function. Such direction of research work was carried out by Zhang et al. [22] where affinity-based propagation is studied over histograms of intensity gradient magnitude in order to generate transfer function. The study proved that such clustering technique assists in accomplishing better accuracy in clustering outcomes as well as it also achieves convergence point faster over medical images.

El-Khamy et al. [23] have presented a study that performs clustering of brain images in order to identify the suspected mass. The technique uses the fuzzy c-means algorithm as well as conformed threshold in order to enhance the clustering performance. The study outcome shows higher accuracy and lower processing time. Kitrungrotsakul et al. [24] have used clustering approach in order to perform segmentation that significantly minimizes the graph scale for increasing the optimization speed. Shabanzadeh et al. [25] have used biogeography-based optimization in order to perform data clustering over the real-life dataset. The study outcome was found to have better performance compared to existing clustering and optimization techniques. Haraty et al. [26] have enhanced k-means clustering for extracting diversified patterns from the medical data. The algorithm also uses greedy approach, where the outcomes of the study have been evaluated with respect to a number of items in dataset and f-measure, a coefficient of variance, etc. Hou and Lin [27] have used subspace clustering in order to carry out image retrieval. The technique uses low-rank representation and a matrix completion algorithm for performing involuntary tag completion. Usage of sparse subspace clustering was also seen in the work carried out by Wen et al. [28]. The technique also utilizes total variation method and forms a non-convex optimization model. The technique is mainly used for recovering image as well as performing clustering over the images that has incomplete information. Zhan et al. [29] have presented a clustering technique for medical images using graph-based theory. The technique uses the weighted representation of the medical image to give a shape of a completed graph that is further subjected to pruning. The study outcome was assessed using f-score. Usage of subspace clustering was also seen in the work carried out by Ziko et al. [30] where a visual descriptor was created. A supervised data is added during the clustering process that further minimizes the errors in the results. Aghabozorgi et al. [31] have used time-series data to formulate a unique hybrid clustering algorithm along with k-medoids. The study outcome of the presented work is assessed using accuracy over the cardinality of the datasets. Harchaoul et al. [32] have used the fuzzy c-means algorithm for overcoming the problems of overlapping clustering. Schultz and Kindlmann [33] have presented a technique for three-dimensional image analyses using spectral clustering. Using medical images, the technique was implemented. Boulemnadjel and Hachouf [34] have presented a technique of subspace clustering considering medical images. Paul et al. [35] have presented a simplified clustering technique that assists in the detection of specific diseases. The authors have used constraint k-means and k-mode clustering technique to achieve this. Sulaiman and Isa [36] have presented a technique of image segmentation using fuzzy k-means clustering. The interesting point is its applicability on different forms of images.

TABLE I. SUMMARY OF EXISTING CLUSTERING TECHNIQUE

Authors	Problem	Technique	Advantages	Limitation
Al-Dmour and Al-Ani et al. [17]	Segmentation	Median filter, fuzzy c-means, subtractive clustering	Low complexity	Not applicable to complex medical data.
Proietti et al. [18]	Minimizing error rate	Kernel-based construction of fuzzy members	Minimal Error rate	-No benchmarking -cannot perform classification of a complex disease condition.
Al-Qizwini et al. [19]	To attain super-resolution	Subspace clustering, PCA	Good Image quality	-Not testified over medical data -Not compared with existing techniques.
Li et al. [20]	Impartial data availability	Minimax Optimization, k-means, k-median	Minimizes complexity	Less likely to work of complex medical data.
Ahmad [21]	Enhancing clustering performance	Fuzzy c-means clustering	Only5% error in clustering	-Less likely applicable on complex medical data -computational performance not stated.
Zhang et al. [22]	Enhancing clustering performance	Affinity propagation clustering	Faster convergence	-Discussion restricted to volume visualized data.
El-Khamy et al. [23]	Adaptive clustering	Fuzzy c-means conformed thresholding	Higher accuracy, and faster processing time	Less effective benchmarking, complexity performance not discussed.
Kitrungrotsakul et al. [24]	Segmentation	Linear iterative clustering	Lower computational time	-Less likely applicable on complex medical data -Higher complexities for multi-modal image
Shabanzadeh et al. [25]	Clustering optimization	Biogeography-based optimization	Lower error rates	-uses of the recursive function to increase the complexity -less scalable approach
Haraty et al. [26]	Enhancing clustering performance	k-means, Greedy approach	Higher and stable F1-score	No-benchmarking
Hou and Lin [27]	Image retrieval	Subspace clustering	Benchmarked outcome	No tested over a medical image.
Wen et al. [28].	Clustering performance	Sparse Sub-space clustering	Efficient image recovery	No tested over a medical image.
Zhan et al. [29]	Medical image clustering	Undirected graph, sparsification	Minimal run time involved	Less Effective benchmarked outcomes.
Ziko et al. [30]	Constructing visual dictionary	Subspace clustering	Minimized Error	Complexity performance not stated.
Aghabozorgi et al. [31]	Cluster enhancing	Hybrid clustering	Good accuracy in classification	No tested over a medical image.
Harchaoul et al. [32]	Data analysis	Fuzzy C-means, probabilistic	Achieved good clustering accuracy, applicable to brain MRI	Not tested over a complex form of data.
Schultz and Kindlmann [33]	3D Medical Image Analysis	Spectral Clustering	Simplified usage, extensive operation	-N/A-
Boulemnadjel and Hachouf [34]	Clustering enhancement	Subspace Clustering	Applicable on original data	-Not benchmarked -More iteration leads to complexity.
Paul et al. [35]	Disease detection	k-means, k-mode	Involves for discrete and continuous data	-Less effective benchmarking -Less likely to work on complex medical data.
Sulaiman and Isa [36]	Image segmentation	Fuzzy k-means clustering	Applicable for post image processing stage.	-Doesn’t address complexity problems -Applicability on complex medical image is not discussed

V. Open Research Issues

This section discusses the open research issues after reviewing the standard clustering techniques as well as some of the significant research being carried out by recent times.

- *Less emphasis on classifying medical data:* Without precise classification, complex medical data cannot be subjected for diagnosis. Such complex medical data are often High-dimension and difficult to perform clustering. Owing to data complexity, existing classification techniques cannot be applied. The biggest challenge is to select one smaller set of highly precise data (suitable for diagnosis) from the massive volume of complex medical data.

- *Few works towards Multi-tier Clustering:* The majority of the Advanced Radiological images (e.g. MRI etc.) are gray scale and not a true color which poses challenges towards the investigation. The majority of the existing diagnosis from medical data is based on a region of interest. Accurate labeling of a region of interest is not feasible in real-time, and hence it demands multi-tier clustering. Existing clustering techniques uses single-tier approach (i.e. using the single template). Some of the challenges in the medical data that are not addressed are i) automatic detection of metastatic stages in medical images, ii) large-scale evaluation of disease detection followed by classification, and iii) automated segmentation

- *Less work towards Clustering Complex Disease Pattern:* Frequently used medical data doesn't exhibit heterogeneous symptoms associated with the particular disease. Existing cluster analysis is not effective towards identifying disease heterogeneity.

- *More inclination towards recursive-based approach:* It has also been seen that maximum studies in the existing system have been used the recursive function which calls for more number of iterative steps to achieve the stage of convergence or meet the objective function. Existing studied has been only testified with respect to time complexity and very few studies to be testified for space complexities. There is less availability of studies that considers using the non-recursive approach in the process of optimization.

Although there is the certain level of work being carried out towards enhancing clustering techniques, it can be easily seen that majority of them are associated with limitations (Table 1 of Section IV). Classification of the disease condition with faster response time and lower computational complexity is the critical demands of clustering techniques over medical images. There is a less number of analytical modeling designed using any of the existing clustering techniques for enhancing the classification performance. Moreover, usage of multi-dimensional technique can further leverage the disease classification while formulating novel clustering technique. Such technique can be used for performing clustering of the medical data with the complex disease condition. However, in an existing system, the term medical data is found maximum corresponding to image only. A closer look at the existing system also shows that there are various clustering techniques that offer lower time complexities. However, there is no such evidence if such claims will be applicable while changing the environments. It will also mean lower applicability in a physical world and more on research work. A closer look at the existing system shows that adoption of the complex medical dataset is few to find. Even with the general medical data, the multiple modalities among the images are quite less to find.

It was also observed that there had been various clustering techniques presented in the past with MRI image that are normally bigger in size using k-means clustering algorithm. In fact, the majority of such scheme is similar to this. Maximum of such techniques are found to provide non-intuitive outcomes of classification. Such outcomes are never considered to be understood completely by the radiologist or attending physician in real-time practices. For better outcomes, it is necessary to perform inference of the clinical outcomes using simple rules. Unfortunately, the complex medical data e.g. that of gene expression data are normally collected in the form of high dimensional format. Such data not only have the higher value of veracity but it also has a greater extent of outliers and noise. Therefore, it is quite a challenging task to design and develop a technique that can deal with such complicated issues of clustering.

VI. Conclusion & Future Work

Clustering is the better way to deal with the classification of the higher number of data by performing logical groups. This paper discusses the theoretical aspects of clustering and its applications and taxonomies. By reviewing the existing clustering schemes, we find that it uses the common database with no clustering algorithm to represent disease heterogeneity. Moreover, existing algorithms are quite specific to the medical database. However, to cope up the rising demands of clustering, it is required that it should start analyzing the database of complex disease condition as well as it should also address disease heterogeneity. It is also required that the algorithm should be working on multiple forms of a complex dataset with nearly similar outcomes. Finally, the paper highlights the open research problem associated with clustering of medical data. Our future work will be in the direction to find the certain robust solution for open resource issues. Our first approach will be to develop a novel prioritizing scheme to select the best sub-cluster from complex medical data followed by application of an enhanced fuzzy logic on the informative sub-cluster extracted from complex medical data. A novel labeling technique would be formulated to assists in extraction of normal as well as the abnormal region. Its consecutive approach will be to formulate a framework for disease classification to address problems associated with multi-tier clustering. A novel multi-modal scheme will be developed for extracting significant features from complex medical data. A study-specific optimal pattern selection strategy will be designed to obtain multiple patterns from data. This step could be further enhanced by performing extraction of multi-modal regional feature representation for each subject from multiple pattern spaces. We will also develop a new technique of Sub-class Clustering-Based Feature Selection by applying supervised learning to perform classification. Our final phase of the study will be to formulate clustering framework for complex disease pattern to address the problem of clustering complex disease pattern. A generative scheme with probability theory is the best way to start this for designing an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of complex disease patterns and parameters. Finally a novel clustering technique can be designed for extracting the patterns of complex disease.

REFERENCES

[1] M. E. Celebi, K. Aydin, Unsupervised Learning Algorithms, Springer-Technology & Engineering, 2016

[2] Y. Yang, Temporal Data Mining via Unsupervised Ensemble Learning, Elsevier, 2016

[3] C. C. Aggarwal, C. K. Reddy, Data Clustering: Algorithms and Applications, CRC Press, 2016

[4] C. Karlsson, Handbook of Research on Innovation and Clusters: Cases and Policies, Edward Elgar Publishing, 2008

[5] J. N. Sheth, Lawrence Sherman, Cluster Analysis and its Applications in Marketing Research, Marketing Classics Press, 2011

[6] S. Dua, P. Chowriappa, Data Mining for Bioinformatics, CRC Press, 2012

[7] R. Lee, Applied Computing & Information Technology, Springer, 2015

[8] S. N. Bhaduri, D. Fogarty, Advanced Business Analytics: Essentials for Developing a Competitive Advantage, Springer, 2016

[9] O. Maimon, L. Rokach, Data Mining and Knowledge Discovery Handbook, Springer Science & Business Media, 2010

[10] K. R. Venugopal, K.G. Srinivasa, L. M. Patnaik, Soft Computing for Data Mining Applications, Springer, 2009

[11] http://nlp.stanford.edu/IR-book/html/htmledition/flat-clustering-1.html

[12] H. J. Miller, J. Han, Geographic Data Mining, and Knowledge Discovery, Second Edition, CRC Press, 2009

[13] J. Han, J. Pei, M. Kamber, Data Mining: Concepts and Techniques, Elsevier, 2011

[14] N. Y. Saiyad, H. B. Prajapati and V. K. Dabhi, "A survey of document clustering using semantic approach," *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, Chennai, 2016, pp. 2555-2562

[15] A. Fahad, N. Alshatri, Z. Tari," A Survey of Clustering Algorithms for Big Data: Taxonomy & Empirical Analysis", IEEE Transactions on Emerging Topics in Computing, 2014

[16] J. Li; H. W. Lewis, "Fuzzy Clustering Algorithms – Review of theApplications", IEEE International Conference on Smart Cloud, 2016

[17] H. Al-Dmour and A. Al-Ani, "MR Brain Image Segmentation Based on Unsupervised and Semi-Supervised Fuzzy Clustering Methods," *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, Gold Coast, QLD, 2016, pp. 1-7.

[18] A. Proietti, L. Liparulo and M. Panella, "2D hierarchical fuzzy clustering using kernel-based membership functions," in *Electronics Letters*, vol. 52, no. 3, pp. 193-195, 2 4 2016.

[19] M. Al-Qizwini, C. Dang, M. Aghagolzadeh and H. Radha, "Image super-resolution via Dual-Manifold Clustering and Subspace Similarity," *2016 IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, 2016, pp. 1429-1433.

[20] J. Li, S. Song, Y. Zhang, and Z. Zhou, "Robust K-Median and K-Means Clustering Algorithms for Incomplete Data, Hindawi Publishing Corporation Mathematical Problems in Engineering, 2016

[21] A. Ahmad, "Evaluation of Modified Categorical Data Fuzzy Clustering Algorithm on the Wisconsin Breast Cancer Dataset", Hindawi Publishing Corporation Scientifica, 2016

[22] T. Zhang, Z. Yi, J. Zheng, D. C. Liu, W-M Pang, "A Clustering-Based Automatic Transfer Function Design for Volume Visualization", Hindawi Publishing Corporation, Mathematical Problems in Engineering, 2016

[23] S. E. El-Khamy, R. A. Sadek and M. A. El-Khoreby, "An efficient brain mass detection with adaptive clustered based fuzzy C-mean and thresholding," *2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, Kuala Lumpur, 2015, pp. 429-433.

[24] T. Kitrungrotsakul, X. H. Han and Y. W. Chen, "Liver segmentation using superpixel-based graph cuts and restricted regions of shape constrains," *2015 IEEE International Conference on Image Processing (ICIP)*, Quebec City, QC, 2015, pp. 3368-3371.

[25] P. Shabanzadeh and R. Yusof1, "An Efficient Optimization Method for Solving Unsupervised Data Classification Problems", Hindawi Publishing Corporation, Computational and Mathematical Methods in Medicine, 2015

[26] R. A. Haraty, M. Dimishkieh, and M. Masud, "An Enhanced k-Means Clustering Algorithm for Pattern Discovery in Healthcare Data", Hindawi Publishing Corporation, International Journal of Distributed Sensor Networks, 2015

[27] T. Kitrungrotsakul, X. H. Han and Y. W. Chen, "Liver segmentation using superpixel-based graph cuts and restricted regions of shape constrains," *2015 IEEE International Conference on Image Processing (ICIP)*, Quebec City, QC, 2015, pp. 3368-3371

[28] X. Wen, L. Qiao, S. Ma, W. Liu and H. Cheng, "Sparse Subspace Clustering for Incomplete Images," *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, Santiago, 2015, pp. 859-867.

[29] Y. Zhan, H. Pan, Q. Han, Xiaoqin Xie, Zhiqiang Zhang and Xiao Zhai, "Medical image clustering algorithm based on graph entropy," *2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, Zhangjiajie, 2015, pp. 1151-1157.

[30] I. M. Ziko, E. Fromont, D. Muselet and M. Sebban, "Supervised spectral subspace clustering for visual dictionary creation in the context of image classification," *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, Kuala Lumpur, 2015, pp. 356-360.

[31] S. Aghabozorgi, T. Y. Wah, T. Herawan, H. A. Jalab, "A Hybrid Algorithm for Clustering of Time Series Data Based on Affinity Search Technique", Hindawi Publishing Corporation The Scientific World Journal, 2014

[32] N-E El Harchaoui, M. A. Kerroum, A. Hammouch, "Unsupervised Approach Data Analysis Based on Fuzzy Possibilistic Clustering: Application to Medical Image MRI", Hindawi Publishing Corporation, Computational Intelligence, and Neuroscience, 2013

[33] T. Schultz and G. L. Kindlmann, "Open-Box Spectral Clustering: Applications to Medical Image Analysis," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2100-2108, Dec. 2013.

[34] A. Boulemnadjel and F. Hachouf, "A new method for finding clusters embedded in subspaces applied to medical tomography scan image," *2012 3rd International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Istanbul, 2012, pp. 383-390.

[35] R. Paul and A. S. M. L. Hoque, "Clustering medical data to predict the likelihood of diseases," *2010 Fifth International Conference on Digital Information Management (ICDIM)*, Thunder Bay, ON, 2010, pp. 44-49.

[36] S. N. Sulaiman and N. A. Mat Isa, "Adaptive fuzzy-K-means clustering algorithm for image segmentation," in *IEEE Transactions on Consumer Electronics*, vol. 56, no. 4, pp. 2661-2668, November 2010.