RESEARCH ARTICLE OPEN ACCESS

A Survey on Clustering Techniques in Medical Diagnosis

N.S.Nithya[1], Dr.K.Duraiswamy[2], P.Gomathy[3],
Department of Computer Science and Engineering
K.S.R.College of Engineering, India.

ABSTRACT
Due to recent technology advances, large masses of medical data are obtained. These large data contain valuable information for diagnosing diseases. Data mining techniques can be used to extract useful patterns from these mass data. It provides a user- oriented approach to the novel and hidden patterns in the data. One of the major challenges in medical domain is the extraction of comprehensible knowledge from medical diagnosis data. Healthcare system becomes very important to develop an automated tool that is capable of identifying and disseminating relevant healthcare information. This paper intends to provide the survey of various clustering techniques used in medical field. The purpose of this survey is to improve the design of clustering methods for further enhancement

Keywords- Medical data mining, Hierarchical, Partitioning, Density Based, K-NN Nearest neighbor clustering techniques.

I. INTRODUCTION

The medical expert system interest for independent decisions in medical and engineering applications is growing, as data becomes easily available. In a previous, an exponential in enhancement has been witnessed in the accuracy and sensitivity of diagnostic tests, from observing external symptoms and using sophisticated lab tests and complex imaging methods increasingly that permit detailed non-enveloping internal examinations. This improved accuracy has certainly resulted in an exponential increase in the patient data available to the physician. The process of obtaining evidence to identify a probable cause of patient's key symptoms from all other possible causes of the symptom are known as establishing a medical diagnosis[1].

Data Mining Techniques applied in many application domains like e-business, Marketing, Health care and Retail have led to its application in other industries and sectors. Medical data mining has great potential for exploring the hidden patterns in the data sets of the medical field. These patterns can be utilized for clinical diagnosis. But the available raw medical data are widely in the form of distributed in nature and large. These data need to be compiled in an organized pattern. Medical diagnosis is regarded as an important yet complex task that needs to be done accurately and efficiently. The Healthcare environment is still information rich but knowledge poor [2].

Data mining technology provides a user oriented approach to the novel and hidden patterns in the data. However, there is a lack of effective analysis tools to discover hidden relationships and trends in data. Data mining is an essential step in the process of knowledge discovery in databases in which intelligent methods are used in order to extract patterns. Many techniques available in data mining such as classification, clustering, association rule, decision trees and artificial neural networks [3]. Clustering is grouped by the similarity data. Each group called clusters, the group consisting object can be similar and dissimilar from other groups. Clustering partitions a data set into several groups such that the similarity within a group is larger than that among groups. Large amount of data, we usually tend to summarize this huge number of data into a small number of groups or classes in order to further facilitate its analysis. This paper provides to analyze the different clustering techniques in diagnosis the medical disease.

II. RELATED WORK

Medical diagnosis is known to be subjective and depends not only on the available data but also on the experience of the physician and even on the psycho-physiological condition of the physician. A number of studies have demonstrated that the diagnosis of one patient can differ significantly if the patient is tested by different physicians or even by the same physician at various times. Healthcare related data mining is one of the most rewarding and challenging field of application in data mining and knowledge discovery. The reason for challenging is due to the data sets of huge, complex, diverse, hierarchical, time series and varying in quality. As the available healthcare datasets are fragmented and diffused in nature, thereby making the process of data integration is a highly challenging task [4].

Considering a small subset of medical datasets, algorithms have been formed and accurate results have been achieved by some of them [5]. Applied to a much larger generalized dataset, for every medical field, obtaining accurate results has yet been very difficult. Liad[6] is an expert system program that uses Bayesian classification to estimate the posterior probabilities of various diagnoses under consideration, given the symptoms present in a case.

DXplain is a Clinical decision support system (CDSS) available through the World Wide Web that assists clinicians by generating stratified diagnoses based on user input of patient signs and symptoms, laboratory results, and other clinical findings. DXplain generates ranked differential diagnoses using a pseudo-probabilistic algorithm. Each clinical finding entered into DXplain is assessed by determining the importance of the finding and how strongly the finding supports a given diagnosis for each disease in the knowledge base. Using this criterion, DXplain generates ranked differential diagnoses with the most likely diseases yielding the lowest rank. Using stored information regarding each disease's prevalence and significance, the system differentiates between common and rare diseases. DXplain takes the advantage of a large database of the crude probabilities of different clinical manifestations associated with different diseases, but unfortunately, it is still confined to the research laboratory or medical training setting[6].

The rest of the work is organized as Section 3 gives literature survey different clustering techniques and Section 4 describes Conclusion.

III. LITERATURE SURVEY

Clustering Techniques

Clustering is an unsupervised data mining (machine learning) technique used for grouping the data elements without advance knowledge of the group definitions. The objective of clustering is to find the intrinsic grouping in a set of unlabeled data. Transform the set of features into subsets so that features in the same subset are similar in some sense shown in Fig-1.

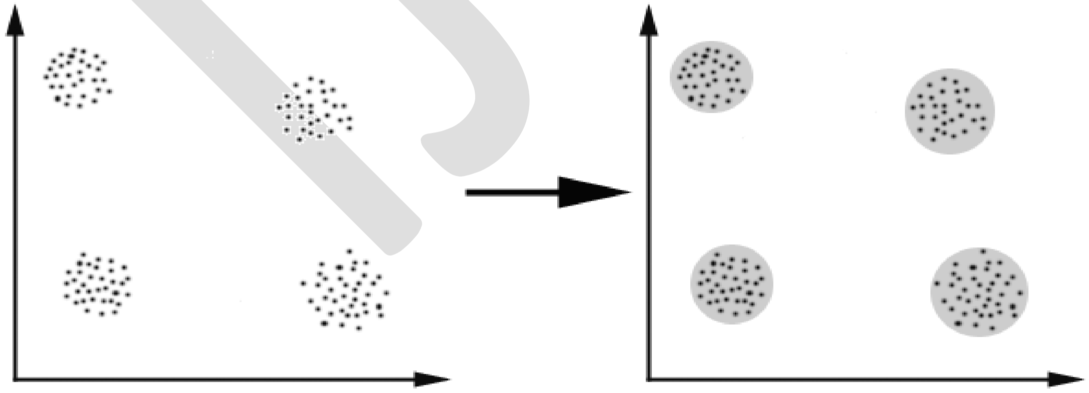

Figure 1.clustering

A. Hierarchical Clustering

Hierarchical clustering builds a cluster hierarchy is known as a dendrogram. Every cluster node contains child clusters, sibling clusters divider the points their common parent. Hierarchical clustering methods are categorized into agglomerative and divisive**.** An agglomerative clustering starts with one-point clusters and recursively merges two or more most appropriate clusters. Compute all pair wise pattern-pattern similarity coefficients. Place each of and patterns in a class of its own. Merge the two most similar clusters into replacing the two clusters into the new cluster and re-compute inter-cluster similarity scores. Merge the two most similar clusters into replacing the two clusters into the new cluster and re-compute inter-cluster similarity scores. The above step until there are *k* clusters left (*k* can be 1).

A divisive clustering starts with one cluster of all data points and recursively splits the most appropriate cluster. Start at the top with all patterns in one cluster. The cluster is split using a flat clustering algorithm. This procedure is applied recursively until each pattern is in its own singleton cluster. Chipman et al., proposed the hybrid hierarchical clustering approach for analyzing microarray data [7].The research work combines both top-down and bottom-up hierarchical clustering concepts in order to effectively utilize the strength of this clustering approach. Chen *et al.*, proposed an integrated approach for analyzing micro- array data. Belciug use the hierarchical clustering approach for grouping the patients according to their length of stay in the hospital that enhance the capability of hospital resource management [8].Figure 2 shows the grouping of the patients into two cluster using 192-gene expression profile. Liu *et al.*, predict the severity of disease in patients using gene expression profile having Rheumatoid Arthritis [9].

The advantages of hierarchical clustering are embedded flexibility regarding the level of granularity, ease of handling of any kinds of similarity or distance, consequently, applicability to any attribute types. The Disadvantage of hierarchical clustering is related to Vagueness of termination criteria [10].

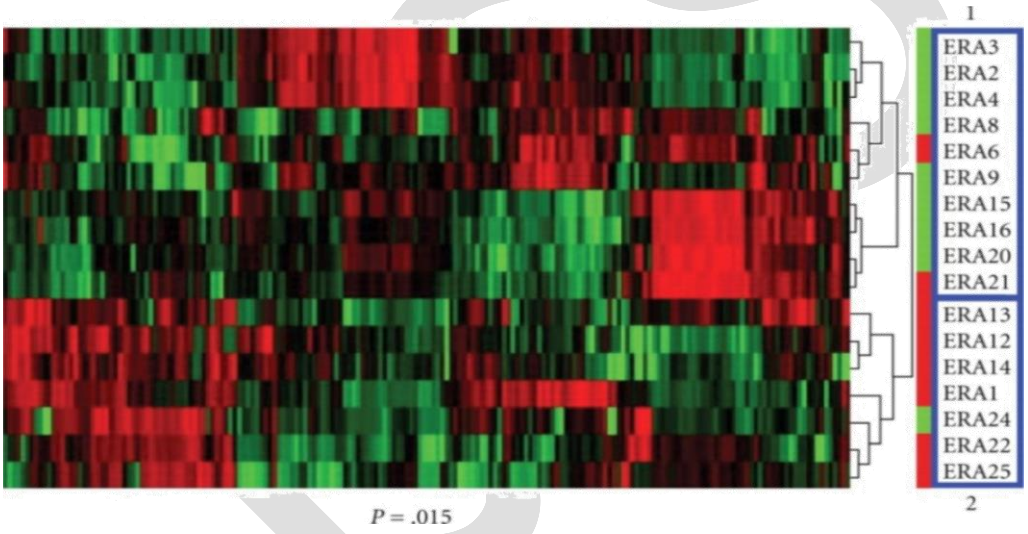

Fig-2. Hierarchical Clustering for Grouping the Gene data into Two Cluster using 192-Gene Expression Profile.

B. Partition Methods

Partition algorithms construct partitions of a database of N objects into a set of k clusters. The partitioning clustering algorithm usually adopts the Iterative Optimization. Specifically, this means different relocation schemes that iteratively reassign points between the *k* clusters. This algorithm can be categorized k-means and k-medoids. K means is very simple and easily implemented. It is a simple iterative method to recover the user specified number of cluster k which is symbolized by the centroids. It presents no limitations on attribute types the choice of methods is dictated by the position of a predominant fraction of points inside a cluster and, consequently, it is lesser sensitive to the presence of.

The grouping of person on the basis of high blood pressure and cholesterol level into high risk and low risk of having heart disease using K-means clustering. Lenert et al., utilize the application of k-means clustering in the health services of public domain [11] and Belciug et al. Detect the recurrence of breast cancer with the help of clustering techniques. The advantage is a simple clustering approach and efficient. The disadvantages require a number of clusters in advance and not discover the cluster with non-convex shape [12].

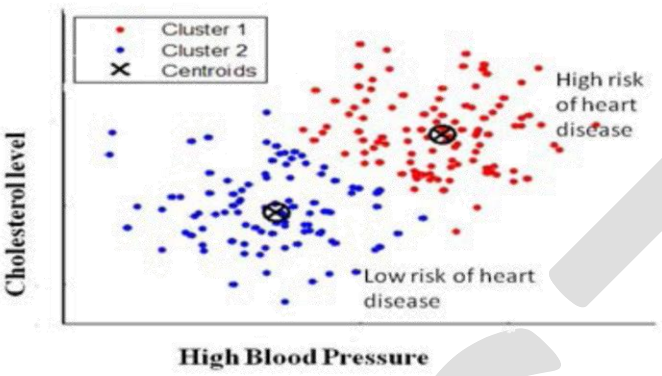

Fig- 3. K-means clustering for heart disease patients.

C. Density Based Clustering

The density based cluster is discovering the clusters of arbitrary shapes and the noise in a spatial database. It uses two parameters Epsilon and Minimum Points of each cluster and at least one point from the respective cluster. The number of neighbors is greater than or equal to minimum points, a cluster is formed. It vends all the neighbor points within distance epsilon of the starting point. The cluster is fully expanded then the algorithm proceeds to iterate through the remaining invested points in the dataset. One disadvantage is cannot cluster data sets well with large differences in densities. It maintains the set of objects in three different categories such as classified, unclassified and noise. Each classified object has an associated cluster –id. A noise object may also have an associated dummy cluster –id. Unclassified object does not have any cluster –id. This research discovers the area of homogeneous color in biomedical images. This method separates the unhealthy skin or wound from healthy skin and discovers the sub regions of varied color or spotted part inside the unhealthy skin which is again useful for classification and association task [13]. Figure-4 shows the clustering of injured skin images using DBSCAN algorithm. The advantage is no need to specify the number of clusters in advance and easily handle cluster with arbitrary shape. The disadvantage is not handling the data points with varying densities and results depend on the distance measures.

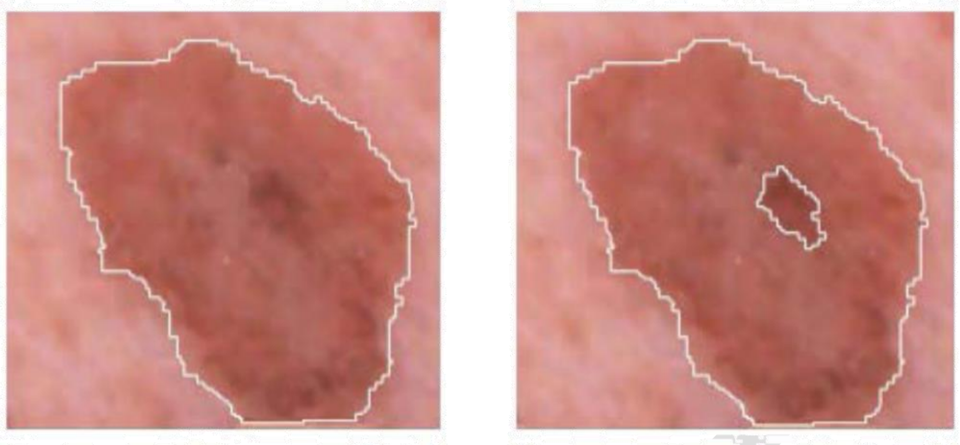

Figure- 4. Clustering of injured skin image using DBSCAN

D. K-Nearest-Neighbour (KNN)

KNN is one of the most simple and straight forward data mining techniques. It is called Memory-Based Classification as the training examples need to be in the memory at run-time. K-NN is a type of instance-based learning, or lazy learning where the mapping is only approximated locally and all computation is deferred until classification. When dealing with continuous attributes the difference between the attributes is calculated using the Euclidean distance. KNN usually deals with continuous attributes however it can also deal with discrete properties. Given a query instance x to be classified, Let x_1,x_2….x_k denote the k instances from training examples that are nearest to x_k .Return the class that represents the maximum of the k instances. Particle Swarm Optimization (PSO) was also used for specifying fuzzy strength constraint and neighbourhood size. Zuo *et al.*, also introduced an adaptive fuzzy K-NN approach for Parkinson disease [14].

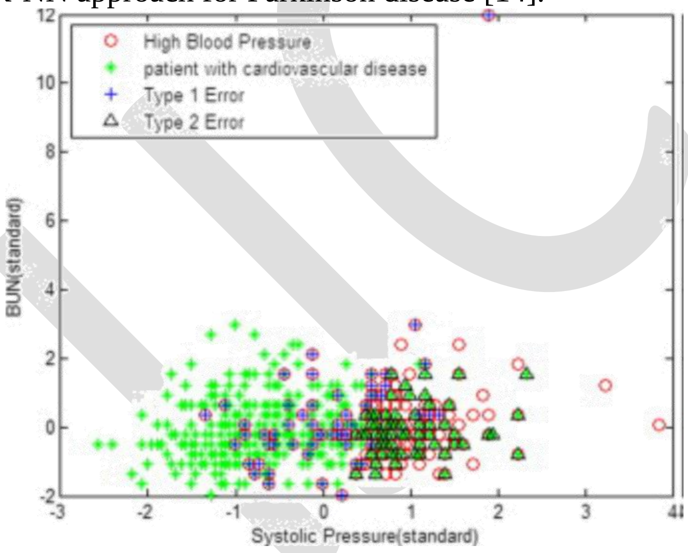

Figure 5.KNN for chronic disease

Distance usually relates to all the attributes and assumes all of them have the same effects on distance. The similarity metrics do not consider the relation of attributes which result in inaccurate distance and then impact on classification precision. KNN advantages is easy to implement and training in faster. Disadvantages is require large database, sensitive to noise, and testing is slow [15].

IV. CONCLUSION

In this paper, different clustering techniques were studied and their advantages and drawbacks have been discussed. The different methods of clustering have been used to extract the useful patterns and thus the knowledge from this variety databases. Selection of data and methods for clustering is an important task in medical diagnosis and needs the knowledge of the domain. The main focus of this survey of clustering techniques is to how the clustering techniques applied in medical field. Each clustering is suitable for some medical applications. An efficient clustering method should be selected that suits desired task.

REFERENCE

[1] E. Barati, M. Saraee, A. Mohammadi, N. Adibi and M. R. Ahamadzadeh " A Survey On Predictive Data mining Approaches for
Medical Informatics" (JSHI): March Edition, 2011.

[2] R., Zhang, Y., Katta, "Medical Data Mining, Data Mining and Knowledge Discovery", pp. 305-308, 2002.

[3] Jiawei Han and Micheline Kamber," Data Mining Concepts and Techniques". San Francisco, CA: Elsevier Inc, 2006.

[4] K.J. and G.W. Moore Cios, "Uniqueness of medical data mining ," Artificial Intelligence in Medicine, pp. 1-24, 2002.

[5] M. Berlingerio, F. B. F. Giannotti, and F. Turini, "Mining clinical data with a
temporal dimension: A case study," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, Nov.
2–4, 2007, pp. 429–436.

[6] H. R. Warner and O. Bouhaddou, "Innovation review: Iliad—A medical diagnostic support program," *Top Health Inf. Manage.*, vol. 14, no. 4, pp. 51–58, 1994.

[7] Department of Medicine Massachusetts Hospital, Boston, DXplain System. (2011). available at :http://dxplain.org/dxpdemopp/dxpdemobrief_ files/frame.html.

[8] H. Chipman and R. Tibshirani, "Hybrid hierarchical clustering with applications to microarray data", Biostatistics, vol. 7, no. 2, **(2009)**, pp. 286-301.

[9] S. Belciug, "Patients length of stay grouping using the hierarchical clustering algorithm", Annals of University of Craiova, Math. Comp. Sci. Ser., ISSN: 1223-6934, vol. 36, no. 2, **(2009)**, pp. 79-84.

[10] S. Belciug, F. Gorunescu, A. Salem and M. Gorunescu, "Clustering-based approach for detecting breast cancer recurrence", 10th International Conference on Intelligent Systems Design and Applications, **(2010)**.

[11] T. Balasubramanian and R. Umarani, "An Analysis on the Impact of Fluoride in Human Health (Dental) using Clustering Data mining Technique", Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering, **(2012)** March 21-23.

[12] J. Escudero, J. P. Zajicek and E. Ifeachor, "Early Detection and Characterization of Alzheimer's Disease in Clinical Scenarios Using Bioprofile Concepts and K-Means", 33rd Annual International Conference of the IEEE EMBS Boston, Massachusetts USA, **(2011)** August 30-September 3.

[13] Velmurugan, T. and Santhanam, T. "A survey of partition based clustering algorithms in data mining: An experimental approach", Information Technology Journal., Vol.10, pp. 478-484, 2011.

[14] László Kozma, Lkozma@cis.hut.fi" k Nearest Neighbors algorithm (kNN)",feb 2008.