CLUSTERING ALGORITHM FOR A HEALTHCARE DATASET USING SILHOUETTE SCORE VALUE

Godwin Ogbuabor[1] and Ugwoke, F. N[2]

[1]School of Computer Science, University of Lincoln, United Kingdom
[2]Department of Computer Science, Michael Okpara University of Agriculture Umudike, Abia State, Nigeria

ABSTRACT

The huge amount of healthcare data, coupled with the need for data analysis tools has made data mining interesting research areas. Data mining tools and techniques help to discover and understand hidden patterns in a dataset which may not be possible by mainly visualization of the data. Selecting appropriate clustering method and optimal number of clusters in healthcare data can be confusing and difficult most times. Presently, a large number of clustering algorithms are available for clustering healthcare data, but it is very difficult for people with little knowledge of data mining to choose suitable clustering algorithms. This paper aims to analyze clustering techniques using healthcare dataset, in order to determine suitable algorithms which can bring the optimized group clusters. Performances of two clustering algorithms (K-means and DBSCAN) were compared using Silhouette score values. Firstly, we analyzed K-means algorithm using different number of clusters (K) and different distance metrics. Secondly, we analyzed DBSCAN algorithm using different minimum number of points required to form a cluster (minPts) and different distance metrics. The experimental result indicates that both K-means and DBSCAN algorithms have strong intra-cluster cohesion and inter-cluster separation. Based on the analysis, K-means algorithm performed better compare to DBSCAN algorithm in terms of clustering accuracy and execution time.

KEYWORDS

Dataset, Clustering, Healthcare data, Silhouette score value, K-means, DBSCAN

1. INTRODUCTION

Data mining is becoming one of the most important and motivating areas of research with the aim of discovering useful information from large amount of datasets [1]. Data mining is the use of automated data analysis methods to unveil unrecognized relationship among datasets [13]. It offers several benefits such as early detection of diseases, availability of medical solutions to the patients at lower cost, identification of medical treatment methods and detection of fraud in healthcare. It also assists the healthcare researchers in developing drug recommendation systems and making adequate healthcare policies [24]. Most of the data mining methods fall into one of two learning tasks: supervised and unsupervised learning. The major aim of supervised learning is to develop a model to predict a situation based on the class label while in unsupervised learning there is no class label; the goal is to model the distribution in the data in order to learn more about the data.

Clustering is one of the unsupervised data mining techniques which deal with finding groups in a set of unlabelled data. It is used to partition a set of data into different groups so that objects in the same group cluster have high similarity to each other and dissimilar to objects in another cluster [10]. Clustering analysis has assisted widely in numerous applications such as pattern recognition, image processing and customer purchase pattern analysis [10]. It is most essential during exploratory and evaluation of data analysis, where researchers try to uncover underlying features that exist without previous knowledge of the data [8]. However, choosing appropriate clustering techniques and algorithm is determined by an understanding of the structure of the data, the kind of analysis to be carried out, and the size of the dataset [8]. There are several clustering algorithms such as fuzzy clustering, density based clustering, Hierarchical clustering and partition clustering. These clustering algorithms are grouped according to the creation of clusters of objects [20]. In partition clustering, the dataset having 'N' data points are grouped into 'K' groups or clusters. Each cluster has at least one data point and each data point must belong to only one cluster [24]. Unlike partitioned clustering, specifying the number of clusters is not necessary in hierarchical clustering. This clustering decomposes the data points in hierarchical way. Density based clustering discover clusters on the basis of density connectivity while fuzzy clustering is a form of clustering in which each data point can belong to more than one cluster.

Choosing appropriate clustering method and optimal number of clusters in healthcare data can be confusing and difficult most times. Currently, a large number of clustering algorithms are available for clustering healthcare data, but very difficult for people with little knowledge of data mining to choose appropriate clustering algorithm [23]. To address this challenge, this paper aims to review existing methods of clustering in healthcare and evaluate the performance of K-means and DBSCAN clustering algorithms using movement activity dataset; in order to determine suitable algorithm which can bring the optimized groups cluster based on level of activities. Silhouette analysis, which is used to study the separation distance between the resulting clusters will be used to determine the appropriate algorithm to cluster the dataset. K-means and DBSCAN methods are selected for the analysis due to their differences in selecting number of clusters; K-means requires prior specification of the number of clusters while DBSCAN does not. Instead clusters are given dataset based on number of clusters discovered by the algorithm.

The rest of the work is organized as follows: in section 2, we review some research carried out by other researchers in clustering healthcare data; section 3 describes method used in the analysis; section 4 discusses clustering algorithms; section 5 presents the performance evaluation; section 6 shows the result of the analysis while section 7 concludes the work and presents future direction.

2. RELATED WORK

Clustering techniques have been massively used in the healthcare industry for easy diagnosis and prediction of diseases, thereby providing fast, adequate, reliable and less costly healthcare delivery to patients. Jabel and Srividhya [12], compared the performance of three clustering algorithms using heart dataset. They used Silhouette width measure to evaluate the performance of the algorithms, from their experimental results, CLARA clustering shows better performance compared to K-means, and PAM. The experiment was however limited to only partitioning clustering algorithms, ignoring other clustering algorithms such as Hierarchical and density-based clustering algorithms. In partitioning clustering, the number of clusters has to be specified by the developer, which can lead to incorrect clustering of the given dataset, while Hierarchical and Density-based clustering chooses the number of clusters by themself. Nithya et al. [20] extended

their research to include other clustering techniques. They compared the performances of three clustering algorithms- Hierarchical clustering, Density based clustering and k-Means clustering algorithms. Diabetes dataset was used to compare the performance of the algorithms based on their execution time and the number of clustered instances. The diabetic dataset was collected from UCI repository and it contains 769 instances and 9 attributes. They argued that using a training set parameter, k-Means algorithm gave better clustered instances compared to other clustering algorithms. They equally recommended other parameters such as cross validation, percentage split, and supplied test set. An exploratory data mining approach based on a density-based clustering algorithm was also presented by [7]. For patients' clustering, they proposed a novel combine distance measure. Their aim was to discover cohesive and well-separated groups of diabetes patients with the same profile (i.e. age and gender) and examination history. They used diabetes dataset from an Italian Local Health Centre. Their experiment shows that their approach was effective in discovering groups of patients with the same examination history. Paul et al. [18], proposed K-means-Mode clustering algorithm using medical data. They pointed out that background knowledge of the medical domain in clustering process will increase the performance of the algorithm. They argued that their algorithm can handle both continuous and discrete data. Balasubbramanian and Umarari [2], analysed the effect of ground water on human health using clustering method. They used K-Means clustering algorithm to find out the risk factors related to the level of fluoride content in water. With the help of this analysis, they were able to discover useful hidden patterns which can help in the making of reasonable decisions in our society. Banu and Jamala [3], developed an algorithm for heart attack prediction using Fuzzy C means which is an unsupervised clustering algorithm that permits one data object to belong to more than one cluster. He argued that the proposed system will provide an aid for physicians to diagnose heart attack in a more efficient manner. Escudero et al. [9], in their work, used the concept of bioprofile and K-means clustering algorithm for early detection of Alzheimer disease and classification of Alzheimer disease into pathologic and non-pathologic groups. Zheng et al. [25] developed weight based K-means algorithm to identify the leukaemia, inflammatory, bacteria or viral infection, HIV infection and pernicious anaemia disease from hemogram blood test samples collected from Kovai Scan Centre. Their aim was to predict the diseases and also to find the efficiency of the algorithm. The performance of their algorithm was evaluated based on the clustering accuracy, execution time and error rate. They argued that their algorithm performed better than other clustering algorithms such as K-means and fuzzy c means. Belciug [5] used agglomerative hierarchical clustering algorithm to group patients according to their length of stay. This can help management in planning and decision making process. Alsayat and El-Sayed [1] presented an efficient K-means clustering algorithm that uses Self Organizing Map (SOM) method to overcome the challenge of finding number of centroids in ordinary k-means. They carried out performance evaluation of the algorithm using two healthcare datasets-Liver disease and heart disease datasets. They claimed that their proposed method shows better clustering performance. Belcliug et al. [6] assessed the effectiveness of three clustering algorithms using Wisconsin Recurrence Breast Cancer dataset. They compared the performance of K-means algorithm with Self-Organizing Map and a cluster network implemented on a real-time decision support system for breast cancer recurrence detection. From their experiment, the result shows that cluster network had the highest performance. They argued that their clustering model showed a better diagnosing performance compared to the standard medical experience and it was also faster and cheaper. They pointed out that clustering method using imaging technique instead of database will likely perform better and will likely be more reliable.

To bridge the gap between supervised and unsupervised learning, some researchers have combined clustering algorithms and classifier model for effective prediction and diagnosis.

Wisconsin Diagnostic Breast Cancer dataset from the University of California was also used by [26] to model algorithms for breast cancer diagnosis. They applied hybrid method of data mining using K-means and Support Vector Machine. K-means algorithm was used to uncover the hidden patterns of the Malignant and Benign tumours separately, while support vector machine was used to design a new classifier to detect breast cancer. Hybrid method of data mining was used by [4] to predict heart attack. They used K-Means to cluster the data after pre-processing. Maximal Frequent Itemset Algorithm (MAFIA) was used for mining maximal frequent patterns in the heart disease database. MAFIA is association rule mining techniques which are mainly efficient when the item set in the database is too large. The frequent pattern from the heart disease dataset was classified using C4.5 algorithm as the training algorithm. Applying the concept of information entropy, they argued that the designed prediction system is capable of predicting heart attack successfully. Ibrahim et al. [11] also applied hybrid the method of data mining. They compared the performance accuracy of Decision Tree Classifier with Agglomerative Hierarchical clustering to standard Decision Tree Classifier using diabetes dataset. The Agglomerative Hierarchical clustering algorithm was used to cluster the dataset, and then the resulting clusters were fed into the Decision Tree algorithm. They reported that the hybrid model achieved higher accuracy compared to the standard model.

3. METHOD

To evaluate the performance of the developed clustering algorithms, movement activity dataset from 'MyHealthAvatar' domain were used. The dataset is made up of movement activity attributes generated from mobile phone sensors over a period of time. The attributes include activities performed by participants (walking and transporting, running), steps taken by each participant, distance covered during the activity, duration of the activity and date of the activity. Steps and duration covered by the participants were selected as features to be clustered. These features were selected due to high correlation strength discovered in them.

Silhouette score value was used to determine appropriate clustering algorithm. Silhouette helps to evaluate the correctness of the assignment of a data object in a particular cluster instead of another cluster by measuring both inter-cluster separation and intra-cluster cohesion [7]. Negative Silhouette values show incorrect placement of objects, while a positive value represents better objects placement [7]. The method used for the analysis is summarized in Figure 1 as shown below.

4. CLUSTERING ALGORITHMS

Clustering is a way of grouping a set of data points into different groups or clusters in such a way that objects within a group have high similarity compared to objects in another group [11]. Clustering is an unsupervised learning which assists professionals in discovering hidden patterns in a dataset. It helps in a situation where there is no class label in a particular data. No clustering method is universally applicable, different methods are used for different clustering purposes. Therefore, common understanding of both the clustering problem and the method is necessary to apply proper method to a particular task. There are different categories of clustering methods such as Partitioning clustering, Hierarchical clustering, Density Based Clustering, Gridbased clustering, Model based Clustering and Fuzzy clustering. In this work, two major clustering Methods (Partitioning clustering and Density Based Clustering) have been described and compared.

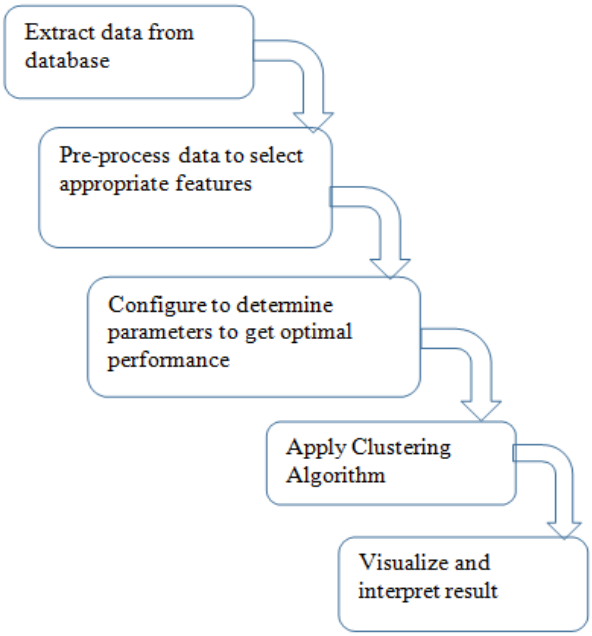

Fig 1: Steps for performing clustering analysis

4.1 Partitioning Clustering Method

Partitioning clustering is the most fundamental and simplest method of cluster analysis that arranges the objects of a dataset into different exclusive clusters [11]. K-means clustering is one of the most widely used partitioning clustering algorithms. However, one of its challenges is that it requires the number of clusters to be pre-specified before the algorithm is applied [21].

K-means algorithm is made up of two different phases. In the first phase, the K centres (centroids) are selected randomly, with a fixed value of K, while the second phase is to allocate each data point to the closest centre [17]. Euclidean distance is mostly used to measure the distance between cluster centres (centroids) and each data point. The Euclidean distance between points a and b is the length of the line segment connecting point a and point b.

$$Dist\ (a,b) = Dist(\ b,a) = \sqrt{\sum_{i=1}^{n} (bi - ai)^2}$$

4.1.1K-means clustering processes

This is performed as follows

- Randomly choose K data points from the dataset as the initial centroids based on the specified number of clusters.
- Then assign each data point to the nearest centroid by calculating the minimum Euclidean distance of each data point to each centroid.
- Set the position of each cluster to the mean of all data points belonging to that cluster.
- Repeat steps b and c above until convergence is observed.

4.2 Density Based Method

Density based method is used to discover clusters of non-spherical shape. In order to find clusters of arbitrary shape, clusters are modelled as dense region in the data space, separated by sparse regions [11]. One of the most widely used density based methods is Density-Based Spatial Clustering of Applications with Noise (DBSCAN) which can discover any shape of clusters and has the ability to identify noise points [14]. DBSCAN requires two parameter Eps and MinPts; Eps means the maximum radius of a neighbourhood from the observing point and MinPts represent the minimum number of points needed to form a cluster. There are three categories of points in DBSCAN. These are core point, border point and noisy point.

Core point is formed if the number of points which are directly density reachable from the point is more than MinPts the Eps- neighbourhood of a point. Border point is formed when the number of points in the Eps- neighbourhood is not more than MinPts, and the point is directly density reachable from a core point. A noise point is regarded as any other point that is neither a border point nor a core point. These points are discarded. A point ‘a’ is said to be density reachable from a point ‘b’ if point ‘a’ is within Eps distance from point ‘b’ and ‘b’ has enough number of points which are within distance Eps in its neighbourhood.

4.2.1The DBSCAN clustering process

This is achievable using the following steps

- Choose a point ‘a’.
- Retrieve all the points that are density-reachable from ‘a’ with respect to Eps and MinPts.
- A cluster is formed if ‘a’ is a core point.
- If no point is density-reachable from point a point ‘a’, it is a border point, then it visits the next point of the dataset. If a point is not in any particular Eps-neighbourhood, it will be regarded as noisy point, which is known as the outlier
- Continue the process until all the points are processed.

5. PERFORMANCE ANALYSIS AND RESULTS

5.1 Silhouette Analysis

Silhouette analysis is used to study and understand the separation distance between the resulting clusters. This analysis is used to measure how close each object in one cluster is close to another objects in another cluster. Silhouette score values lie between -1 to +1. The value of +1 indicates correct clustering of objects while the value of -1 show that objects are not properly clustered.

The following steps show how to calculate the silhouette value for each object.

a. Given jth object, compute its average distance to other objects in its cluster. Let this value be xj.
b. For the jth object, compute the object’s average distance to all the objects in other clusters. Calculate the minimum value with respect to all clusters. Let this value be yj.
c. For the jth object, the silhouette value is sj = (yj-xj) / max(xj,bj).

5.2 K-means Analysis

K-means clustering method is one of the most adopted methods of clustering in healthcare industry due to its simplicity and performance in other fields of research. The name was derived from representing each cluster with the mean (average weight) of the points in each cluster known as centroid. To determine the appropriate number of K value in K-means clustering is a common challenge in the clustering process. To handle this challenge, we used the silhouette score value of different range of clusters to determine appropriate number of K value for the dataset. Euclidean distance metric was used to determine both inter-cluster separation and intra-cluster cohesion. Based on the analysis, K value of two(2) gave the highest silhouette score which shows that to get optimized cluster of the dataset using K-means algorithm, the appropriate value of K is two(2). Figure 2 shows different values of K and their corresponding silhouette score value.

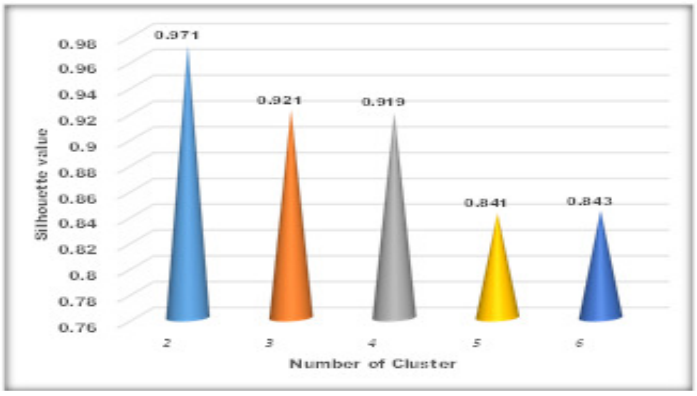

Figure 2: Comparison of K-means clustering algorithm using different range of clusters.

From Figure 2, the algorithm had optimal clusters when the value of K (number of clusters) is 2 and less optimal cluster when the value of K is 5. In order to select appropriate distance metric, we decided to compare four distance metrics and found out that using Euclidean and Minkowski gave highest silhouette score value of 0.975. However, because Euclidean distance is popularly known and used by people, we adopted it in this analysis. Figure 3 below shows result of Silhouette values for different distance metrics.

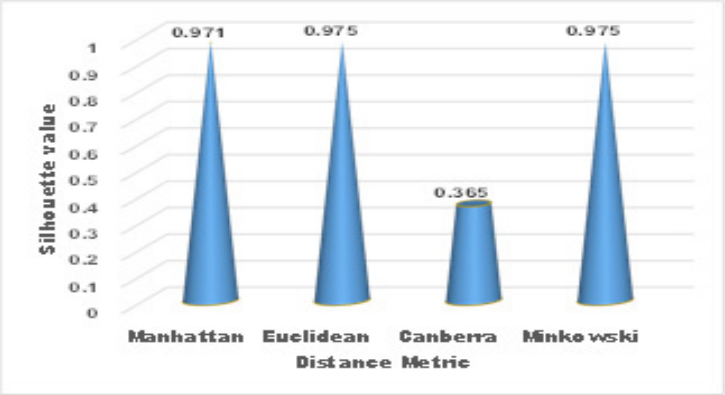

Figure 3: Comparison of different distance metrics in Silhouette analysis using K-means algorithm

5.3 DBSCAN Analysis

Density Based Spatial Clustering of Application with Noise (DBSCAN) is an example of density-based method which separates data points into three parts [8]. The three parts are core points (points that are within the cluster), Border point (points that are within the neighbourhood of the core point) and Noise points (neither core nor border points). It makes use of the specified minimum radius (Eps) and minimum number of points required to form a cluster (minPts). Though, this algorithm does not require prior specification of the number of clusters, it might not perform well in high dimensional dataset.

After pre-processing of the dataset, DBSCAN clustering algorithm was used to cluster the dataset. Silhouette score value of the algorithm was used to evaluate the performance of the algorithm. Silhouette helps to evaluate the appropriateness of the allocation of data point to a cluster rather than another cluster by calculating both intra-cluster cohesion and inter-cluster separation. Clusters within the range of 51% to 70% and 71% to 100% respectively indicate that a reasonable and a strong intra-cluster cohesion and inter-cluster separation are found [7]. We discovered that adjusting the MinPts parameter value has significant effect on determining the silhouette value; MinPts represent the minimum number of points needed to form a cluster. Therefore we decided to find appropriate value of MinPts using a range of numbers (2, 4, 6, 8, 10, and 12). Figure 4 shows different MinPts and their corresponding Silhouette score values. It indicates that the algorithm had optimal cluster when the value of MinPts is six (6) and less when it is two (2).

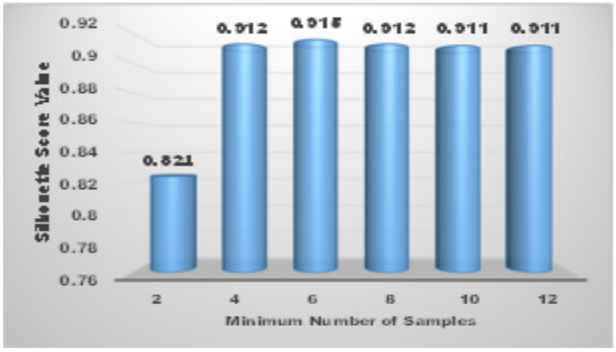

Figure. 4 Comparison of different minimum number of samples in DBSCAN using Silhouette score value

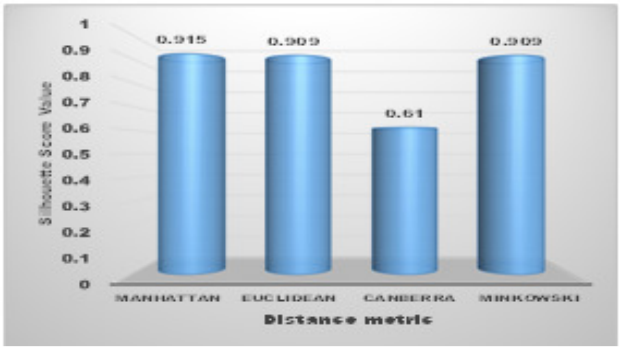

Figure 5: Comparison of different distance metric in Silhouette analysis using DBSCAN algorithm

In order to select appropriate distance metric, we decided to compare four different distance metrics as shown in Figure. 5. The graph shows that using Manhattan distance metric, the silhouette score value was greater than using Euclidean, Canberra, and Murkowski distance metrics.

6. RESULTS

After pre-processing of the data and carefully selecting appropriate parameters,we observed that both K-means and DBSCAN algorithms have strong intra-cluster cohesion and inter-cluster separation. This shows that any of the algorithms can be used to cluster the dataset. However, in terms of accuracy, K-means clustering algorithm performed better than DBSCAN algorithm. K-means algorithm performed better with 0.97 Silhouette score compare to DBSCAN algorithm with 0.91 Silhouette score. Also, execution time of K-means algorithm is 0.013seconds while execution time of DBSCAN algorithm is 0.156seconds. Figures 6 and 7 show the final results we obtained in terms of clustering accuracy and execution time respectively.

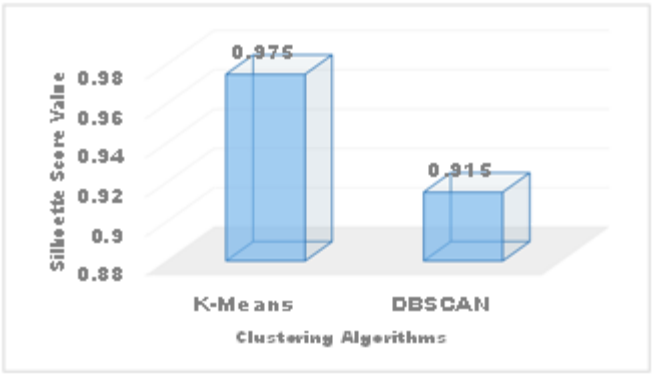

Figure 6: Comparison of K-means and DBSCAN in terms of clustering accuracy

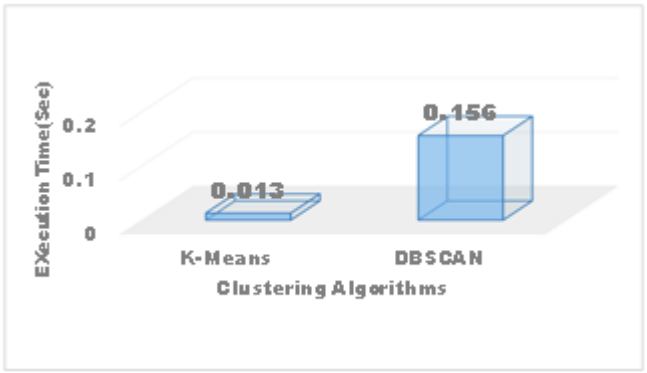

Figure 7: Comparison of K-means and DBSCAN in terms of execution time

7. CONCLUSION

In this paper, we analysed two different clustering algorithms (K-means and DBSCAN) using Silhouette analysis in order to obtain optimal clusters for a movement activity dataset. The

analysis shows that obtaining optimal cluster using the specified algorithms depend on the value of parameters chosen. Based on the analysis, both K-means and DBSCAN algorithms have strong intra-cluster cohesion and inter-cluster separation. K-means algorithm performed better compare to DBSCAN algorithm in terms of clustering accuracy and execution time. In the future, we intend to apply hybrid method (Clustering and Classification) to predict activity of individuals using the healthcare data

REFERENCES

[1] Alsayat, A., & El-Sayed, H. (2016). Efficient genetic K-Means clustering for health care knowledge discovery. In Software Engineering Research, Management and Applications (SERA), 2016 IEEE 14th International Conference on (pp. 45-52). IEEE.

[2] Balasubramanian, T., & Umarani, R. (2012, March). An analysis on the impact of fluoride in human health (dental) using clustering data mining technique. In Pattern Recognition, Informatics and Medical Engineering (PRIME), 2012 International Conference on (pp. 370-375). IEEE.

[3] Banu G. Rasitha & Jamala J.H.Bousal (2015). Perdicting Heart Attack using Fuzzy C Means Clustering Algorithm. International Journal of Latest Trends in Engineering and Technology (IJLTET).

[4] Banu, M. N., & Gomathy, B. (2014). Disease forecasting system using data mining methods. In Intelligent Computing Applications (ICICA), 2014 International Conference on (pp. 130-133). IEEE.

[5] Belciug, S. (2009). Patients length of stay grouping using the hierarchical clustering algorithm. Annals of the University of Craiova-Mathematics and Computer Science Series, 36(2), 79-84.

[6] Belciug, S., Salem, A. B., Gorunescu, F., & Gorunescu, M. (2010, November). Clustering-based approach for detecting breast cancer recurrence. In Intelligent Systems Design and Applications (ISDA), 2010 10th International Conference on (pp. 533-538). IEEE.

[7] Bruno, G., Cerquitelli, T., Chiusano, S., & Xiao, X. (2014). A clustering-based approach to analyse examinations for diabetic patients. In Healthcare Informatics (ICHI), 2014 IEEE International Conference on (pp. 45-50). IEEE.

[8] DeFreitas, K., & Bernard, M. (2015). Comparative performance analysis of clustering techniques in educational data mining. IADIS International Journal on Computer Science & Information Systems, 10(2).

[9] Escudero, J., Zajicek, J. P., & Ifeachor, E. (2011). Early detection and characterization of Alzheimer's disease in clinical scenarios using Bioprofile concepts and K-means. In Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE (pp. 6470-6473). IEEE.

[10] Han, J., Kamber, M., & Pei, J. (2012). Cluster Analysis-10: Basic Concepts and Methods.

[11] Ibrahim, N. H., Mustapha, A., Rosli, R., & Helmee, N. H. (2013). A hybrid model of hierarchical clustering and decision tree for rule-based classification of diabetic patients. International Journal of Engineering and Technology (IJET), 5(5), 3986-91.

[12] Jabel K. Merlin & Srividhya (2016). Performance analysis of clustering algorithms on heart dataset. International Journal of Modern Computer Science, 5(4), 113-117.

[13] Kar Amit Kumar, Shailesh Kumar Patel & Rajkishor Yadav (2016). A Comparative Study & Performance Evaluation of Different Clustering Techniques in Data Mining. ACEIT Conference Proceeding.

[14] Lv, Y., Ma, T., Tang, M., Cao, J., Tian, Y., Al-Dhelaan, A., & Al-Rodhaan, M. (2016). An efficient and scalable density-based clustering algorithm for datasets with complex structures. Neurocomputing, 171, 9-22.

[15] Malli, S., Nagesh, H. R., & Joshi, H. G. (2014). A Study on Rural Health care Data sets using Clustering Algorithms. International Journal of Engineering Research and Applications, 3(8), 517-520.

[16] Maulik, U., & Bandyopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(12), 1650-1654.

[17] Na, S., Xumin, L., & Yong, G. (2010, April). Research on k-means clustering algorithm: An improved k-means clustering algorithm. In Intelligent Information Technology and Security Informatics (IITSI), 2010 Third International Symposium on (pp. 63-67). IEEE.

[18] Paul, R., & Hoque, A. S. M. L. (2010, July). Clustering medical data to predict the likelihood of diseases. In Digital Information Management (ICDIM), 2010 Fifth International Conference on (pp. 44-49). IEEE.

[19] Pham, D. T., Dimov, S. S., & Nguyen, C. D. (2005). Selection of K in K-means clustering. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 219(1), 103-119.

[20] R.Nithya & P.Manikandan & D.Ramyachitra (2015); Analysis of clustering technique for the diabetes dataset using the training set parameter. International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 9.

[21] Sagar, H. K., & Sharma, V. (2014). Error Evaluation on K-Means and Hierarchical Clustering with Effect of Distance Functions for Iris Dataset. International Journal of Computer Applications, 86(16).

[22] Shah, G. H., Bhensdadia, C. K., & Ganatra, A. P. (2012). An empirical evaluation of density-based clustering techniques. International Journal of Soft Computing and Engineering (IJSCE) ISSN, 22312307, 216-223.

[23] Tan, P. N., Steinbach, M., & Kumar, V. (2013). Data mining cluster analysis: basic concepts and algorithms. Introduction to data mining.

[24] Tomar, D., & Agarwal, S. (2013). A survey on Data Mining approaches for Healthcare. International Journal of Bio-Science and Bio-Technology, 5(5), 241-266.

[25] Vijayarani, S., & Sudha, S. (2015). An efficient clustering algorithm for predicting diseases from hemogram blood test samples. Indian Journal of Science and Technology, 8(17).

[26] Zheng, B., Yoon, S. W., & Lam, S. S. (2014). Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Systems with Applications, 41(4), 1476-1482.