International Journal of Mathematics and
Computer Science, **15**(2020), no. 4, 1005–1014

A Clustering Algorithm Application in Parkinson Disease based on $k-$means Method

Israa Ali Alshabeeb, Nidaa Ghalib Ali,
Saba Abdulameer Naser, Wafaa M. R. Shakir

Technical Computer System Department
Babylon Technical Institute
Al-Furat Al-Awsat Technical University
Babil, Iraq

Inb.esr@atu.edu.iq, Inb.nedaa10@atu.edu.iq,
sabaabdulameernaser@gmail.com, inb.wfa@atu.edu.iq

(Received July 21, 2020, Accepted August 17, 2020)

Abstract

Data mining methods are used to predict and compare ages of people with Parkinson's disease and so are considered a critical part in the medical community. The amount of data stored in the medical section database is increasing rapidly. There is a system used to arrange the results of ages of Parkinson disease patients by years. In this paper, a data mining technique called the $k-$means clustering algorithm for analyzing data is implemented. This technique is applied on a hospital database and analyzes the performance successfully and so is useful in giving accurate results and making an effective decision by the Ministry of Health in Iraq and related parties to find appropriate solutions.

1 Introduction

The ability to monitor the performance of a disease in a huge number of patients is becoming of increasing importance. It is crucial to know how many

Key words and phrases: Data Mining, Clustering, Parkinson disease, $k-$means clustering.
AMS (MOS) Subject Classifications: 68Q25, 68W40, 62H30
ISSN 1814-0432, 2020, http://ijmcs.future-in-tech.net

people are affected by diseases in relevance to different ages. A second popular degenerative disorder disease after Alzheimer's is Parkinson disease (PD). The most common symptoms of PD are shaking and slowness of movement. The main cause of this disease is unknown [1] [2]. In data mining, many algorithms have been used with diseases. Decision tree, factor analysis, neural net and logistic regression were implemented to examine the biomedical voice measurements to find out which measurement is more suitable to figure out the early symptoms of PD and follow early treatment [3]. Using data mining techniques such as Decision Tree Algorithm, Naive Bayes and Neural Network doctors can detect the risk rate of a heart disease [4]. Deepika and Kalaiselvi [5] gave a review of using data mining methods and comparing the results for diagnosis and prognosis of breast cancer disease, heart disease and thyroid. Using technical regression, the decision tree models and technical regression to predict diabetes using specific risk factors were discussed in [6]. While Bayesian analysis was used to estimate how PD depends on hereditary as a risk factor key and how it affected PD patients [7]. In this work, a new approach is used to study yearly comparisons of PD patients to see if the disease is growing or not. The algorithm that is proposed in this paper gives good accuracy.
Physicians and The Ministry of Health in Iraq try to know how a disease has been progressing to find solutions to treat people at an early stage. There are many clustering methods in data mining that are available to predict how many patients are affected every year. The $k-$means clustering method is used in this paper to know how PD affects people.

2 Methodology

2.1 Clustering

Clustering is a data mining technique for gathering similar objects in features or properties in one group which is called a cluster. When the distance between two objects is less than any other distance of other objects will be in the same subset and this subset should contain at least one object [8]. This technique simply puts similar data into a groups and dissimilar objects into a separate group [9]. Clustering is widely used in many different fields and applications such as pattern recognition, image processing, security, machine-learning situations, data analysis, business, web search and biology [10]. Clustering methods can be classified in many types. Those are represented by partitioning methods, density-based methods, hierarchi-

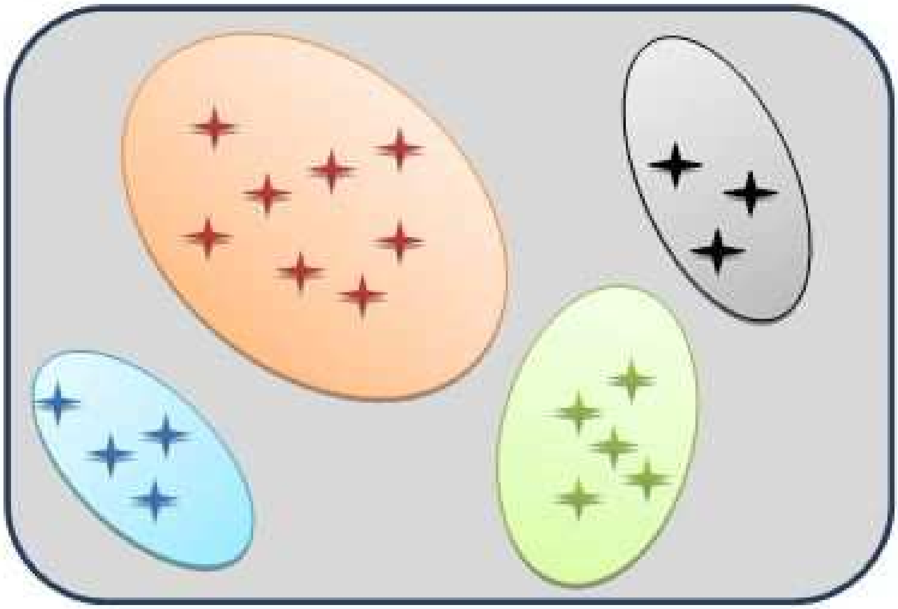

Figure 1: Partitioning clustering

cal methods, grid-based methods and model-based methods [11]. The type of data and the purpose of application are the main points to choose the best clustering method. Many researches have used clustering algorithms, like Hierarchical clustering [12], $k-$means clustering [13] [14] in applications. Partitioning methods work based on iterative reallocating data objects among subsets and determine the initial number of subsets. Partitioning algorithms are represented by the $k-$medoids and $k-$means. $k-$means algorithm is proposed in this paper. Partitioning clustering is shown in figure 1

2.2 Data mining process

In this work, data was gathered from Neuroscience hospital in Baghdad, Iraq for four years (2016-2019) and a classification method to analyze them.

2.2.1 $k-$means Method

$k-$means is one of the clustering algorithms. It is used to cluster the data into groups based on a centroid point. This algorithm is used widely in pattern recognition applications [15]. It is useful with both a big or small dataset and gives a good result. $k-$means clustering is considered as an unsupervised linear method [13]. It works in many steps. The first step is to mention the number of k clustering and the centroid point. Then we put the data that are similar to each other or have the lower distance in one group. Finally, we repeat the steps above to set all data in groups and no object is moved from cluster to another. The $k-$means flow chart is shown in figure 2 and the process of the algorithm is demonstrated in figures 3, 4 and 5.

The objective function for $k-$mean method could be written as:

$$f = \sum_{i=1}^{k} \sum_{j=1}^{n} (pj - yi)^2, \tag{2.1}$$

where k is the number of clusters, n is the number of cases, pj represents the point or the object in the cluster and yi is the centroid point for cluster j. The centroid point represents the mean value for each cluster.

In figure 3, the number of clusters was selected and the centroid point for each group was found.

In figure 4, the distance between centroid points and the object was computed and assigned the object to the nearest cluster depending on the distance value. In figure 5, we recalculated the centroid for each cluster and reassigned objects to groups until there was no change any longer.

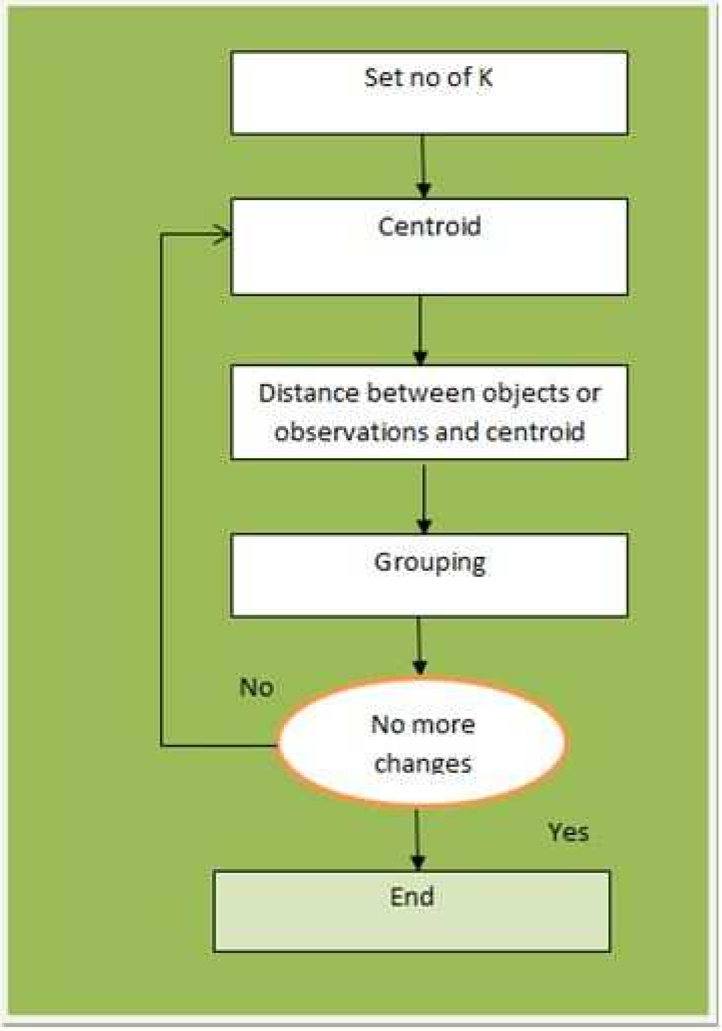

Figure 2: Flowchart of $k-$means algorithm

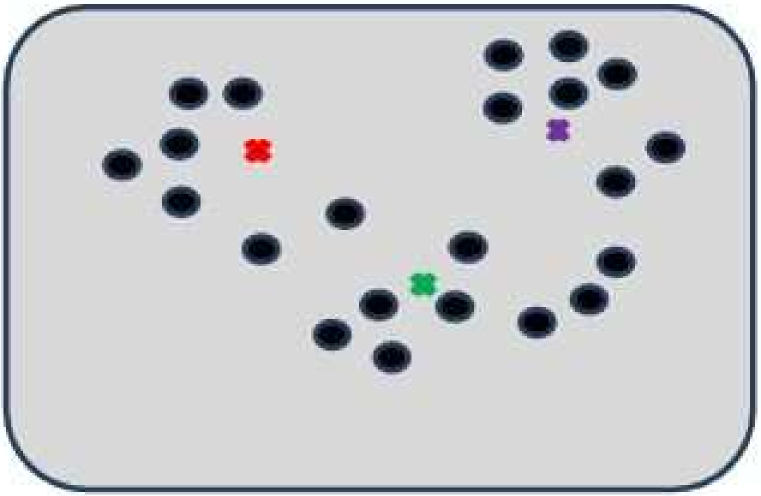

Figure 3: $k-$means at initialization

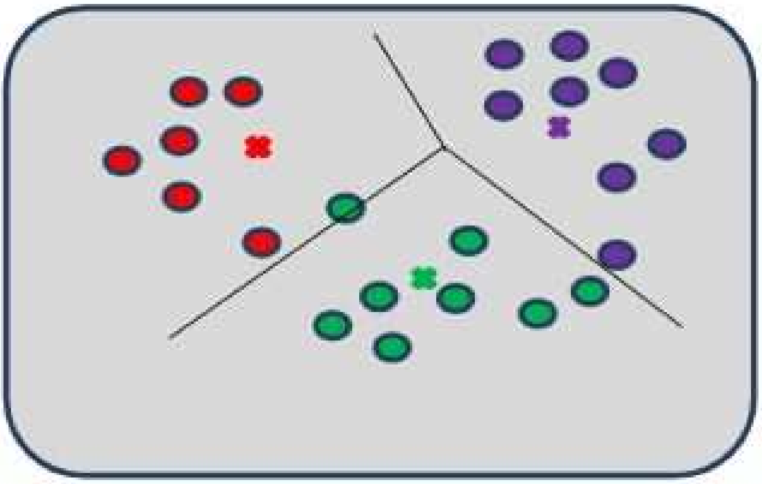

Figure 4: Centroid and grouping

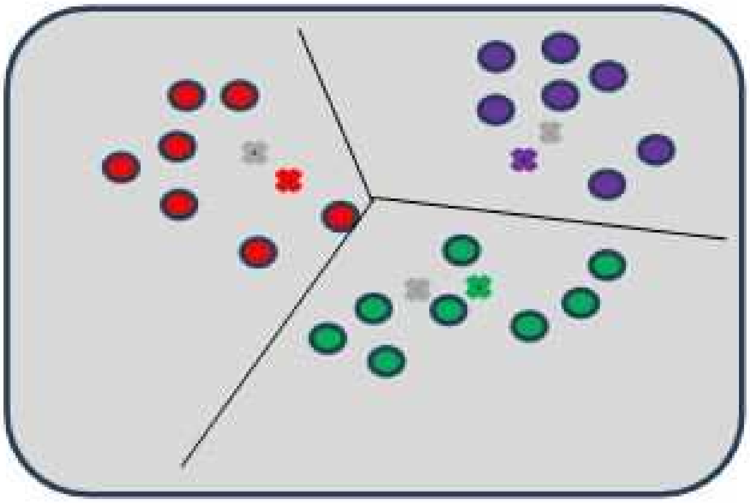

Figure 5: Recomputed centroid and regrouping

3 Result and Discussion

In this model, the data set was applied on 35 patients at the hospital each year for 4 successive years. The Xl-miner tool was used to analyze the dataset and implement the method. The generated results are shown in tables 1, 2, 3 and 4, respectively. In table 1, for the year 2016, the overall age for cluster size 14 is 75.3 while overall age of cluster size 13 is 63. That means 27 out of 35 of patients are older than 62. The rest of patients are from 44 to 50 years old. In table 2, for the year 2017, the overall age for cluster size 10 is 62.7 while overall age of cluster size 15 is 52.9. That means 25 out of 35 patients are older than 50 years and the rest are from 14 to 41 years old. For table 3, for the year 2018, the overall age for cluster size 11 and size 10 are between 47.7 and 39 years old. The rest of patients are from 15 to 28 years old. In 2019, 24 patients out of 35 between 32 and 48 are suffering from the disease and 11 patients are from 8 to 23 years old. That means PD seems to affect younger people compared to 2016 and 2017.

Table 1: Data summary for 2016

cluster	Size	age
cluster1	14	75.3
cluster2	13	63
cluster3	3	44
cluster4	5	50.6
Total	35	

Table 2: Data summary for 2017

cluster	Size	age
cluster1	10	62.7
cluster2	15	52.9
cluster3	2	14
cluster4	8	41.75
Total	35	

Table 3: Data summary for 2018

cluster	Size	age
cluster1	11	47.7
cluster2	10	39.6
cluster3	7	15.2
cluster4	7	28.2
Total	35	

Table 4: Data summary for 2019

cluster	Size	age
cluster1	13	48.3
cluster2	11	32.6
cluster3	3	8.3
cluster4	8	23.2
Total	35	

References

[1] Fan, Kuan, Pengzhi Hu, Chengyuan Song, Xiong Deng, Jie Wen, Yiming Liu, Hao Deng. "Novel Compound Heterozygous PRKN Variants in a Han-Chinese Family with Early-Onset Parkinsons Disease." Parkinsons Disease, (2019).

[2] Priyansha Raj Sinha, Amit Alexander Charan, "Parkinsons disease: A review article." The Pharma Innovation, **6,** no. 9, Part H, (2017), 511.

[3] Shianghau Wu, Jiannjong Guo, "A Data Mining Analysis of the Parkinsons Disease," iBusiness, **3,** no. 1, (2011), 71–75.

[4] J. Thomas, R. Theresa Princy, "Human heart disease prediction system using data mining techniques," In 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT), 1-5. IEEE, (2016).

[5] M. Deepika, K. Kalaiselvi, "A Empirical study on Disease Diagnosis using Data Mining Techniques," In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), IEEE, (2018), 615–620.

[6] Xue-Hui Meng, Yi-Xiang Huang, Dong-Ping Rao, Qiu Zhang, Qing Liu, "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors," The Kaohsiung journal of medical sciences, **29,** no. 2, (2013), 93–99.

[7] Abolfazl Saghafi, Chris P. Tsokos, Rebecca D. Wooten, "On Heredity Factors of Parkinsons Disease: A Parametric and Bayesian Analysis," Advances in Parkinson's Disease, **7,** no. 3, (2018), 31–42.

[8] Jiawei Han, Micheline Kamber, Jian Pei, Data mining concepts and techniques, 3rd edition, Morgan Kaufmann, 2011.

[9] Hina Gulati, P. K. Singh, "Clustering techniques in data mining: A comparison," In 2015 2nd international conference on computing for sustainable global development (INDIACom), IEEE, (2015), 410–415.

[10] J. Han, M. Kamber. Data Mining: Concepts and Techniques, 2nd Edition, Elsevier, 2006.

[11] T. Madhulatha, T. Soni, "An overview on clustering methods," arXiv preprint arXiv:1205.1117, (2012).

[12] Lin Liao, Zhen Jia, Yang Deng, "Coarse-Graining Method Based on Hierarchical Clustering on Complex Networks," Communications and Network, **11,** no. 1, (2019), 21–34.

[13] Ling-Li Jiang, Yu-Xiang Cao, Hua-Kui Yin, Kong-Shu Deng, "An improved kernel $k-$means cluster method and its application in fault diagnosis of roller bearing," (2013).

[14] Manyun Lin, Xiangang Zhao, Cunqun Fan, Lizi Xie, Lan Wei, Peng Guo, "Polarimetric Meteorological Satellite Data Processing Software Classification Based on Principal Component Analysis and Improved $k-$means Algorithm," Journal of Geoscience and Environment Protection, **5,** no. 7, (2017), 39.

[15] Siwei Wang, Miaomiao Li, Ning Hu, En Zhu, Jingtao Hu, Xinwang Liu, Jianping Yin, "$k-$means clustering with incomplete data," IEEE Access, **7,** (2019), 69162-69171.