ORIGINAL ARTICLE

INTERNATIONAL JOURNAL OF RESEARCH IN PHARMACEUTICAL SCIENCES

Published by JK Welfare & Pharmascope Foundation Journal Home Page: https://ijrps.com

A Comparative approach to cluster maternal data using k-means and k-medoid

Bipin Nair B J*, Adarsh C K, Nishin S, Nihar K P, Shivaranjith P

Department of Computer Science, Amrita School of Arts and Sciences, Amrita Vishwa Vidyapeetham, Mysuru, Karnataka, India

Article History:

Received on: 08.03.2019
Revised on: 25.06.2019
Accepted on: 29.06.2019

Keywords:

K-means,
K-medoid,
Elective LSCS,
Emergency LSCS

ABSTRACT

In our work, we cluster the data collected on pregnancy women through hospitals, into three different categories such as elective LSCS, emergency LSCS and normal delivery. We aim to identify outliers so that we can predict the complications that can occur during pregnancy. These complications can even lead to maternal death. Clustering is performed by the aid of k-means, and k-medoid procedures which are implemented in windows form application to run in Microsoft Visual Studio, a comparison is provided on its performance and efficiency between both the algorithms and also the outlier, i.e. normal delivery cases are detected and depicted. From our work, we can clearly say that K-Means works better when the dataset is between the range of 50 – 100. Suppose the data set is greater than 500 then K-medoid works efficiently. In our work, we are considering various maternity cases. Here we are detecting normal delivery cases as an outlier.

*Corresponding Author

Name: Bipin Nair B J
Phone: +91 9961451866
Email: bipin.bj.nair@gmail.com

ISSN: 0975-7538

DOI: https://doi.org/10.26452/ijrps.v10i3.1416

Production and Hosted by

IJRPS | https://ijrps.com

INTRODUCTION

The data clustering is a tedious task in different areas, such as pattern matching and bio-medical data. Clustering is a technique in data mining which accumulates the data based on its similarity. A cluster imbibes similar kind of data items. The main reason is to organize the data item into clusters, such that each cluster imbibes similar kind of data items within. In our proposed work, we took advantage of k-medoid and k-means clustering algorithm for clustering of various types of pregnancy cases based on the maternity data.

Clustering is a method of making a group of abstract objects to eloquent subclasses. In clustering, there are six types of methods. One of the methods is partitioning method K-means, and K-medoid comes under it. K-means is a simple unsupervised learning algorithm. K-means algorithm will classify the data items into various groups. The data items are grouped by reducing the figure of squares of spaces between the centroid of the cluster and the data. K-means attempts to minimize the total squared errors.

K-medoid is a clustering method for the classical partitioning of data items, clustering will be done using the information set of n items and k clusters. The value will be defined before the algorithm executes. It will minimize the sum of dissimilitude between points within the cluster and the centroid.

A Caesarean operation, also called Caesarean, is a medical approach for deliver the baby by making the incision on the abdomen and uterus of the pregnant women.

Elective LSCS

If a doctor is planning for Caesarean, he/she is responsible to you to give all the required informa-

tion regarding Caesarean to make you choose a decision. If one take a decision for Caesarean during the initial stage, it will be done only after 38 weeks of the pregnancy.

Reasons behind planned Caesarean include

1. Sometimes the natural birth (vaginally) may not happen because the placenta is over the cervix (placenta praevia).
2. If the body of the baby was immense or not in the correct position, in such cases, we may need a Caesarean.
3. We force to do Caesarean if the cervix did not open properly.

Emergency LSCS

During the situations like if the baby is not safe to wait till the vaginal birth, sometimes due to the concern about your health or your baby's health we may force to do the Caesarean as an emergency.

Literature review

(Batra, 2011) has analysed both K-Means and K-Medoid algorithms on the basis of their approach. They generated input data items through normal and uniform distribution and measured the performance of both algorithms. (Velmurugan and Dharmarajan, 2014) Have done a novel research study on k-means and k-medoid algorithms through analysing the quality of clustering of both algorithms. Initially, they implemented the algorithms using JAVA language, and they gave the randomly distributed data items as input and calculated the performance of each algorithm. They represented the experimental results in tabular as well as graphical format. (Arora *et al.*, 2016) Here they have used Keel tool for analysis, and they have taken 10 thousand random numbers and then generated the points and clustered them. By comparing the execution time, space complexity, noise reduction and the level of dissimilarity, they identified that k-medoid was much better than k-means. (Soni and Patel, 2017) Had used the UCI Machine Learning Repository as input data and compared both algorithms. Thy considered the execution time, the sensitivity of out layer data, scalability and noise reduction of both k-means and k-medoid algorithms during the time of execution and reached a result that k-medoid is more efficient than k-means. (Reynolds *et al.*, 2004) Have done a novel study on two clustering methods, and their research work leads to the generation of an all rules algorithm that combines both PAM and k-medoid algorithm. While comparing with the k-medoid, they observed that PAM results in an effective result. (Santhanam, 2016) I have used three datasets for analysis. In this paper, they have shown that k-Means is much more efficient than k-Medoid while comparing the distance between the objects they observed that k-medoid is consuming more time than k-means. Most" "time-consuming part of the k-Medoid algorithm, when the calculation of the distance between the object. (Dharmarajan and Velmurugan, 2016) This work is done in java, and their computational complexity is also analysed. The advantages of k-means computational cost are low and high computational cost in k-Medoid. Several researchers have proven that the k-Means algorithm is suitable to compare to other clustering algorithm in the medical data set. Sufcw (2001) The paper proposes an idea to profitably modify the k-means clustering algorithm. Improvements in clustering accuracy were observed when artificial constraints were applied on six data sets. The method is implemented in real-time GPS applications for detecting the road lanes from satellite data. (Pattabiraman *et al.*, 2009) Have proposed a new spatial clustering method for edge detection. This paper discusses the obstacles related to spatial clustering in detail. They took advantage of the k-medoid method for clustering the spatial data. They considered execution time and the number of clusters from various clustering algorithms for the comparison. (Shah and Singh, 2012) Proposed a modified k-means algorithm. The modified k-means, k-means and k-medoid are evaluated using real-time data. The paper concludes that the proposed modified k-means is better as it takes a small amount of time for the execution, better in terms of a number of clusters when compared with k-means and k-medoid algorithms. (Nair *et al.*, 2015) The paper proposes the identification of dense regions of a particular city or a geographical area with the help of census data. DB scan clustering algorithm was used to cluster the data. This paper shows the experience of a data mining approach using census data. (Nair *et al.*, 2018) Have done a novel research study to find the protein functional region in human embryo cell by using the ANN classifier. They had drawn a box-plot diagram for showing the experimental result and evaluated the performance of the classifier on the basis of the comparison energy level of the coding region.

Proposed work

In our proposed work, we are collecting pregnant women's data manually from Hospital after that automating the data and converting into excel format. The collected data is made in tabular format, as shown in Table 1.

Table 1: Dataset used for clustering

Name	Age	Indication	Surgery Mode	Baby Weight
Tejaswini	25	Fetal Distrus	Emergency LSCS	2.5
Ranjitha	19	Ecephdic Presentation in Labour	Emergency LSCS	2.5
Apoorva	27	Fetal Distress in Labour	Emergency LSCS	3.9
Ashwini	25	MSAF in labour	Emergency LSCS	2.7
Tanuja	25	Fetal Distress	Emergency LSCS	2.65
Chnnajamma	31	Post JVI	Emergency LSCS	2.6
Latha	25	Fetal Distress	Emergency LSCS	2.75
Sheethal	28	Failed Induction	Emergency LSCS	3.25
Thanvi	23	None	Normal Delivery	2.5
Anitha	18	Fetal Distress	Emergency LSCS	2.75
Ramya	28	Prev Cd in Labour	Emergency LSCS	2.65
Anitha	19	Complete Bleach in II stape	breech vaginal delivery	2.5
shushma	21	Prev Cd in Labour	Emergency LSCS	3
shruthi	22	MSAF in labour	Emergency LSCS	3.4
sheela	31	cpd in labour	Emergency LSCS	3
radha	20	petve in labour	Emergency LSCS	2.4
Meena	29	Prev Cd in Labour	Emergency LSCS	2.7
Aswini	32	MSAF in labour	Emergency LSCS	2.8
ramya	21	fetal distress	Emergency LSCS	2.5
jyothi	22	severe dehydration in labour	Emergency LSCS	2.5
Anjali	22	MSAF in labour	Emergency LSCS	2
Bharathi	22	Prev Cd in Labour	Emergency LSCS	2.7
Saraswathi	23	G3,P1,L1 in labour	Emergency LSCS	2.25
Varalakshmi	24	fetal distress	Emergency LSCS	2.4
Latha	26	fetal distress	Emergency LSCS	3
sheela	25	Prime labour	Emergency LSCS	2.5
Anjali k	21	Oligohydramnios	Emergency LSCS	3
Vibha	23	Prime cpd in labour	Emergency LSCS	2.5
Ningamma	30	Failed Induction	Emergency LSCS	3.25
Chaitra	26	G2,a1 growth incuction	Emergency LSCS	3.3
Vinutha	26	ovelation induction	Emergency LSCS	3.3
Geetha	21	Oligohydramnios	Emergency LSCS	2.5
Redha	21	fetal distress	Emergency LSCS	2.8
Rashmita	18	cpd in labour	Emergency LSCS	3.75
Kavana	28	Oligohydramnios	Emergency LSCS	3.74
Dakshayini	24	polyhydranion in labour	Emergency LSCS	3.7
Swetha	24	fetal distress	Emergency LSCS	2.7
Asha	23	Cure gworth	Normal Delivery	3
Manjula	34	fetal distress	Emergency LSCS	3.4
Rohini	22	preved	Emergency LSCS	3
Indira	23	preved	Emergency LSCS	2.8
Nalani	24	preved	Emergency LSCS	2.7
Jothi	23	MCDA GDM in labour	Emergency LSCS	1.8
Soubhagya	19	Oligohydramnios	Emergency LSCS	3
Manasa	26	Oligohydramnios	Emergency LSCS	3
Vinisha	19	phomsh in labour	Emergency LSCS	2.7
Manjua	25	cpd in labour	Emergency LSCS	3.7
Bhavya	29	prime at breech	Emergency LSCS	3
Poornima	28	cpd in labour	Emergency LSCS	3
Nandini	25	pracd in labour	Emergency LSCS	3.4

Applying two clustering algorithms like K-Means and K-Medoid. To group the data for various categories such as Elective LSCS, Emergency LSCS and Normal Delivery. Compare the efficiency between two clustering algorithms based on the partitioning techniques. Based on the algorithms, we are clearly detecting the outliers, such as normal delivery. The dataflow diagram for identifying outlier is given in Figure 4.

MATERIALS AND METHODS

Step 1-collecting and preparing the pregnancy data.

Step 2-Apply K-means algorithm.

Step 3-Apply K-medoid algorithm.

Step 4-detect outlier in both cases.

Step 5-compare the execution time.

K-means

INPUT

A = set of e data points

K = number of clusters

I = Iterations desired

OUTPUT

C = set of c cluster centroids

L = set of distances, l from e to the assigned centroid.

For c in C

Allocate centroid c to remain at some a.

For a in A

Calculate distance from a to all centroids c.

Assign each e to centroid c with min. distance. Store in L.

i = 0.

minDistance = Inf

While i < I:

For c in C

Calculate the average position of all an allocated to cluster c.

Reallocate centroid c to different location.

For a in A

Compute distance from a to all centroids c.

If minDistance!= l:

Allocate each a to centroid c with min. distance = L.

Else

End

Return assignments.

Distance formula for Euclidean and the term explanation

$$E_D = \sqrt{(a2 - a1)^2 + (b2 - b1)^2}$$

E_D is the distance and (a1, b1) and (a2, b2) are the points

The algorithm receipts the input bound L, the amount of clusters to be separated between a set of Z objects. A Usual K-Medoid algorithm for separating built on Medoid or central items.

Input

L: The number of clusters

T: A data set containing Z objects

Output: a set of L clusters that minimizes the sum of the dissimilarities of all the objects to their nearest medoid.

Method: Arbitrarily choose Z objects in T as the initial representative objects.

Repeat: Assign each remaining object to the cluster with the nearest medoid; randomly select a non-medoid object S random; calculate the whole points M of exchange point Sj with S random.

If M < 0 then exchange Sj with S random to form the new set of k medoid Till no variation.

RESULTS AND DISCUSSION

The pregnancy data is taken based on the age, indication, surgery mode and the baby weight.

Using k means algorithm, we got two clusters like cluster1 and cluster 2 using various kind of deliveries. Here we are plotting the x-axis value as no of clusters and y-axis as baby weight. In this clustering approach, we found that normal and breech vaginal delivery case as an outlier. The implementation results of K-means are given in Figure 1.

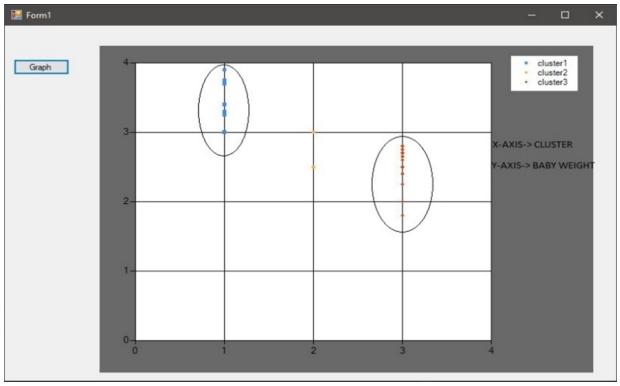

Figure 1: Clusters obtained for K-means

In K-medoid we are using the same data set like various type of pregnancy, and here we are getting three

clusters based on baby weight and surgery mode, here outlier detection is very clear as normal delivery and breech vaginal delivery. We can see the clusters formed by implementing K-medoid algorithm in Figure 2.

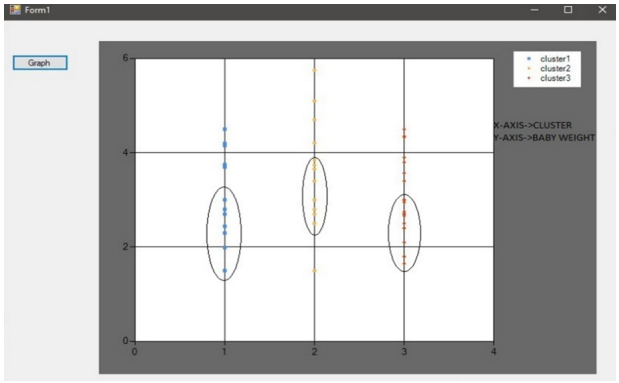

Figure 2: Clusters obtained for K-medoid

When we compare k-mean and k-medoid in terms of execution time, K-medoid is good for a 100 data set as well as outlier detection is very good and clear in K-medoid compare to K-means. The Comparative graph of K-means and K-medoid gives in Figure 3.

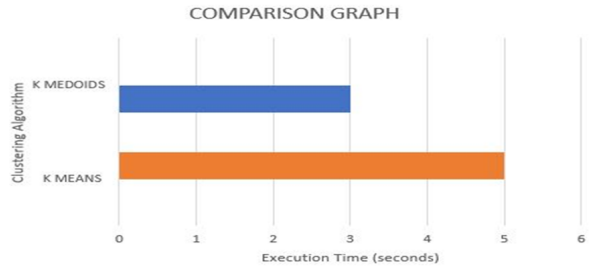

Figure 3: Comparison graph of K-means and K-medoid

CONCLUSION

From our work, we can clearly say that K-Means works better when the dataset is between the range 50 – 100. Suppose the data set is greater than 500 then K-medoid works efficiently. In our work, we are considering various maternity cases. Here we are detecting normal delivery cases as an outlier.

In future studies they can increase the dataset size accordingly partitioning algorithm will change in terms of efficiency, also instead of two groups, we can considering multiple dataset complications in maternity for clustering to increase the performance of the clustering algorithms.

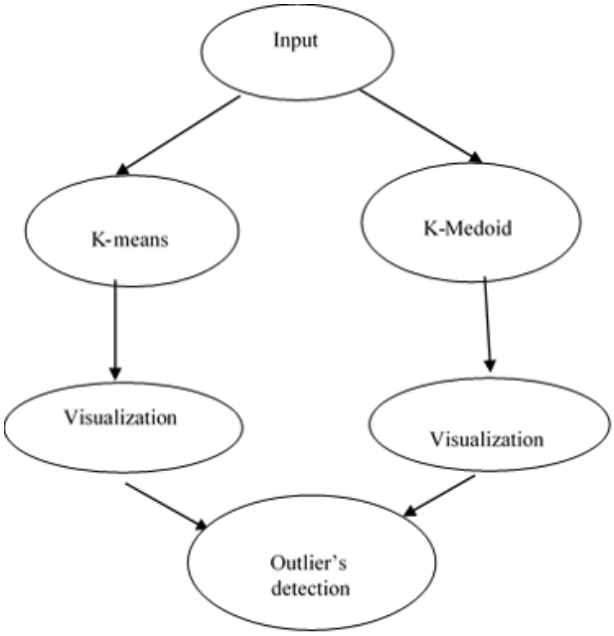

Figure 4: Data flow diagram for detecting outliers

REFERENCES

Arora, P., Deepali, Varshney, S. 2016. Analysis of K-Means and K-Medoids Algorithm For Big Data. *Procedia Computer Science*, 78:507–512.

Batra, A. 2011. *Analysis and Approach: K-Means and K-Medoids Data Mining Algorithms*, volume 274.

Dharmarajan, A., Velmurugan, T. 2016. Efficiency of k-Means and k-Medoids Clustering Algorithms using Lung Cancer Dataset. *International Journal of Data Mining Techniques and Applications*, 5(2):150–156.

Nair, B. J. B., Arjun, K., Reghunath, R. 2018. Coding and functional defect region prediction of placental protein in an embryo cell of first trimester using ANN approach. *International Journal of Engineering & Technology*, 7(1):167.

Nair, B. J. B., Arjun, K. P., Kundapur, B. P. 2015. Customization and visualization of DBSCAN algorithm for demographic analysis. *International Journal of Applied Engineering Research*, 10(55):1508–1512.

Pattabiraman, V., Parvathi, R., Neduncheziyan, R., Palaniammal, S. 2009. *A Novel Spatial Clustering with Obstacles and Facilitators Constraint Based on Edge Detection and K-Medoids*.

Reynolds, A. P., Richards, G., Rayward-Smith, V. J. 2004. The Application of K-Medoids and PAM to the Clustering of Rules.

Santhanam, T. V. T. 2016. Performance Analysis Of K-Means And K- Medoids Clustering Algorithms For A Randomly Generated Data Set,. *Int. Conf. Syst. Cybern. Informatics, no. November*, pages 578–583.

Shah, S., Singh, M. 2012. *Comparison of a Time Efficient Modified K-mean Algorithm with K-Mean and K-Medoid Algorithm*.

Soni, K. G., Patel, A. 2017. Comparative Analysis of K-means and K-medoids Algorithm on IRIS Data. *International Journal of Computational Intelligence Research ISSN*, 13(5):899–906.

Sufcw, S. 2001. *? I 4 : T I,.

Velmurugan, T., Dharmarajan, A. 2014. Clustering Lung Cancer Data by k-Means and k- Medoids Algorithms. *International Journal of Data Mining Techniques and Applications*, 3(2):95–98.