DOI: http://dx.doi.org/10.26483/ijarcs.v8i7.4582

ISSN No. 0976-5697

Volume 8, No. 7, July – August 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

IMPLEMENTATION AND APPLICATIONS OF DATA MINING IN MEDICAL DECISION MAKING PREDICTIONS

Hinayat Sawhney
Department of Computer Engineering
Punjabi University
Patiala Punjab, India

Harpreet Kaur
Department of Computer Engineering
Punjabi University
Patiala Punjab, India

Abstract: Data mining is the process of discovering patterns and trends from the huge amount of data**.** This research studied the various existing techniques of data mining**.** For reviewing different mining techniques, the research progressed to analyze the advantages and limitations of each technique**.** Machine learning based on the clustering technique has been applied in the context of Medical Data Mining**.** This process involves the exploration of the Medical dataset to discover interesting patterns in the decision-making process**.** The issue in exploring the Medical dataset is that the data produced by any healthcare organization is huge and complex which increases the complexity of mining**.** This research aimed to collect the dataset of breast cancer for medical decision making that was carried out using the clustering and data mining techniques**.** The results show that c***lu***stering can be a better technique to group the patients based on their disease and other parameter requ***ir***ed to diagnose cancer.

Keywords: Machine learning, Data Mining, Clustering algorithms, Classification, Knowledge Discovery in Databases (KDD, pattern discovery.

1. INTRODUCTION

Data mining is a technique of collecting the data by finding the patterns in the information already acquired from various sources. The process is based on three main sections which are mainly machine learning, statistics, and large databases. The motive to achieve the data mining task is to collect the relevant information by filtering it from the vast amount of information which may do not belong to the topic of research. Data mining works by the extraction of the information and pattern from a large pool of knowledge data sets. [3]

2. DATA MINING PROCESSES

The process of data mining can be broadly divided into two major categories which are data preprocessing and data mining. Sometimes the data preprocessing is also termed as data preparation. Further, the process of data mining is divided into six steps for providing an ease while working on the process [4]:

The first three steps which fall under data preparation are:

- Data cleaning,
- Data integration
- Data selection and data transformation.

The next three steps are part of knowledge discovery or data mining

- Data mining
- Pattern evaluation
- Knowledge representation.

Knowledge Discovery in Databases (KDD) refers to the wide procedure of discovering knowledge in data and underlines the high-level use of specific data mining techniques. [5] It is important to experts in machine learning, databases, pattern recognition, artificial intelligence, statistics, knowledge procurement for data visualization and expert frameworks. The objective of the KDD procedure is to mine knowledge from data with regards to huge databases. The KDD process includes the following steps to extract knowledge from a large database. [4]

We will now discuss each step-in some detail to get over the idea of data mining process:

Data cleaning: Data cleaning is the process of filtering the data which is noisy and inconsistent in the raw form which cannot be used directly for any purpose. The process of data cleaning is done by using various techniques which involve computer and human effort both for filling of the missing entries and information.

Data integration: Data integration is the process of collecting the data which is required for any specific purpose, at one single place by locating the relevant data from different sources. The data integration is a process of integrating data from sources such as spreadsheets, word documents, PowerPoint files, online sources, research databases, etc. [5]

Data selection and data transformation: The data which is integrated is always in the greater quantity than its use for serving a purpose. The data selection process is done to select the relevant data which can serve the purpose appropriately and rest of the information is eliminated. While the process of data transformation is performed to get a transformed data, which is gone through processes of normalization, generalization, and aggregation. The purpose of data transformation is to get suitable data for data mining.

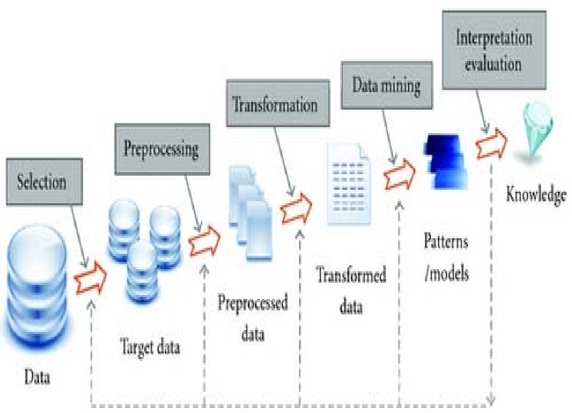

Figure 1: KDD process in Data mining [8]

Data mining: The process of data mining involves various tasks related to data clustering, classification of data, distinct types of analysis and prediction.

Pattern evaluation: The process of pattern evaluation works by finding the different patterns in the several types of data which can serve a single specific purpose. The pattern evaluation is done to increase the relevancy of data.

Knowledge representation: The knowledge representation process of data mining technique helps in turning the data into the attractive form of representation by using specific terms, numbers, or images, etc.

The need of this study is that data mining can do great in the field of healthcare and medical facilities. The use of data mining techniques in the healthcare system can help in designing treatment models which are effective and cost effective so that an improved care can be initiated by the medical industry. Overall the data mining technique in the medical service can help in the management of data, improving customer relationships, effective treatment processes and risk assessment and identification of serious health issues. [7]

3. RELATED WORK

Data mining techniques are used to check the problems in diabetic patients. The data were collected from diabetic patients using questionnaires, direct interview and observing medical records. After receiving feedback from the patient, the modifications occur in techniques. Three type of clusters used in this research is type-1, type-2, and gestational diabetes. The clusters used Simple K-means algorithm. The Random Tree, Simple Cart, and Simple Logistic algorithm are also used for an expert clinical system. [1]

Pan Deng and Feng Chen focused on data mining algorithm. Data mining algorithm is used to collect hidden information for data. The author uses various types of analysis like association analysis, time series analysis, etc. It proposes a modified data mining system. Various steps are involved in the data mining algorithm are data preparation, data mining, and data presentation. The internet of things is used to manage devices and instrument. Data mining are combined with the internet of things (IOT) for optimization. Data mining occur in three views namely knowledge view, technical view, and an application view. In knowledge view clustering, classification and analysis were included. In application view applications like industry were included, and the technical view wass used with knowledge and application view [2].

Data mining is essential in determining the patterns, the discovery of the knowledge, forecasting, etc. in the various business domains. The application of the data mining is vast in every industry in which the data is generated. So, it is referred as the essential frontiers in the information systems and database. Data mining helps in the interdisciplinary developments in the Information technology [7].

Cluster analysis was used for data compression and in the field of psychology, biology, pattern recognition, information retrieval and data mining. Various algorithm and clustering techniques are used in data mining. Clustering was used to convert a collection of pattern into clusters. The clustering method used are partitioning methods such as K-Means, Hierarchical methods such as Agglomerative Nesting, Density based method is DBSCAN(Density-Based Spatial Clustering of Applications with Noise), Grid-based method and a Model-based clustering method. Cluster validation methods are internal approaches, relative approaches, and external approaches. The clustering affects the feasibility of data analysis [6].

Clustering is the technique which put the same data into the groups. It is the essential learning technique which identifies the structure in the collection of unlabeled data. Some researchers analyzed that the k-means algorithm has the biggest benefit of clustering the large data sets and their performance increases with the increase of the clusters. But their use is restricted. So, divisive and agglomerative hierarchical algorithm was accepted for the categorical data because of their complexity for assigning the rank value. The results also showed that the k-means algorithm has the better performance as compared to the hierarchical clustering algorithm. The arbitrary size clusters were determined using the density based methods such as DBSCAN and OPTICS. On the other hand, hierarchical and partitioning methods are used in order to identify the spherically shaped clusters. Density based methods are referred as the exclusive clusters, and they can be extended from full to subspace clustering **[8].**

Data mining applications increase in health care due to the vast information in the health sector which was managed with the data mining. The healthcare organization collects and generate the large information regularly. Data mining and knowledge in the information technology generated some patterns which help in removing the manual tasks, electronic transfer system helps in securing the medical records, helps in easy extraction of the data from the electronic records, minimizes the cost of the various medical services, save lives. It also helps in detecting the infectious diseases which were based on the advanced data collection. [9]

Data mining enables the healthcare organization to anticipate various trends in the patient's behavior and medical condition. Data mining also provide various options for analyzing several data models which are hidden or less visible in the common techniques of the analysis. The patterns of the data mining used in the healthcare organizations for diagnosis, forecast and the treatment of the patients [9].

4. METHODOLOGY USED

In this research, we collected Medical data set related to breast cancer disease. The dataset is taken from the UCI website http://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28diagnostic%29. This research implemented k-mean, k-medoids, and X-means clustering algorithms using Rapid Miner tool. These three algorithms were applied to evaluate the performance of each algorithm in terms of clustering the medical data. The dataset contains information related to the type of breast cancer as M (Malignant) and B (Benign). ***A benign tumor*** does not invade its surrounding tissue or spread around the body. ***Whereas a malignant tumor*** may invade its surrounding tissue or spread around the body.

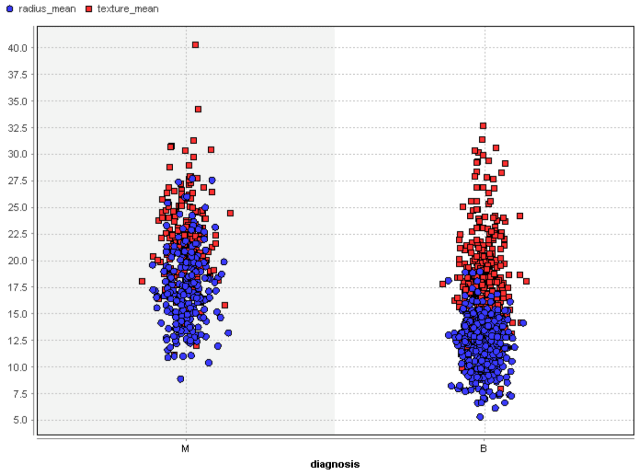

Figure 2: Radius mean and texture mean of dataset

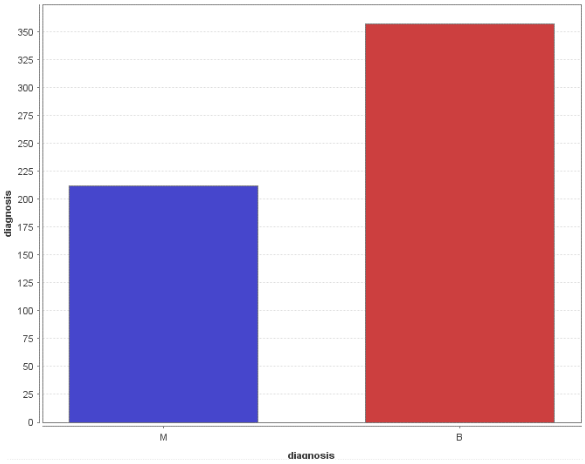

Figure 3: Frequencies of M (Malignant) and B (Benign)

Three algorithms were used to cluster the dataset. Below is the description of those three algorithms.

A. K-means algorithm

This algorithm is clustering algorithm, and its main purpose is to perform clustering which is the process of grouping objects into different clusters based on the similarity of the objects. That is objects with the same, or common attribute is put into one cluster, not with different objects. This algorithm performs clustering based on Euclidean distance. It puts one object into only one of the clusters. To make it run, we have to specify the k value that is some clusters that we want the algorithm to generate. Based on the value of the k, the k points are selected called centroids in such a way that are representations of all the values. Now we check each value one by one and assign it to clusters according to their distance and this process will continue until centroid does not move.

Advantages of K-Means Algorithm: This algorithm is simple and easy to implement. It is suitable to perform clustering of large datasets. It generates the clusters fast.

Disadvantages of K-Means Algorithm: It is difficult to estimate the value of k. Also choosing the average initial k points is difficult, and different chosen values can give different results.

Uses of K-Means Algorithm: This algorithm is used in search engines to group information of similar types. This algorithm can be used in academic departments to monitor the performance of students. It has its applications in image processing, pattern recognition, and machine learning, etc.

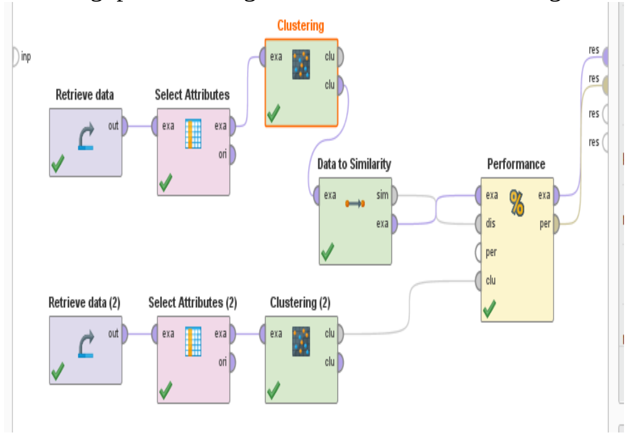

Figure 4: Clustering using K-means

B. K-medoids

This algorithm is a version of k--means clustering algorithm which is more robust to outliers and to noises. K medoids algorithm chooses the actual point in the cluster to represent it as the centroid, unlike the k-means algorithm. The point that is chosen is called media in this algorithm. The point is chosen in such a way that it is the central point which has the minimum sum of distances from other points. This algorithm works like this: it first chooses the k data objects that represent the medoids that are called k cluster. Now the data objects are put to that particular cluster which has the medoid nearest to that data object. Now new medoid is determined again, and the process of putting objects to a cluster is repeated until medoids do not change.

Advantages of K-medoids Algorithm: K-medoids is not sensitive to outliers and to noises, unlike k-means. This algorithm is suitable for small datasets and separated clusters.

Disadvantages of K-medoids Algorithm: It is more complex to implement than a k-means algorithm. This algorithm is less efficient than k-means.

Uses of K-medoids Algorithm: This algorithm is used for clustering of documents over the web. This is used for web usage mining.

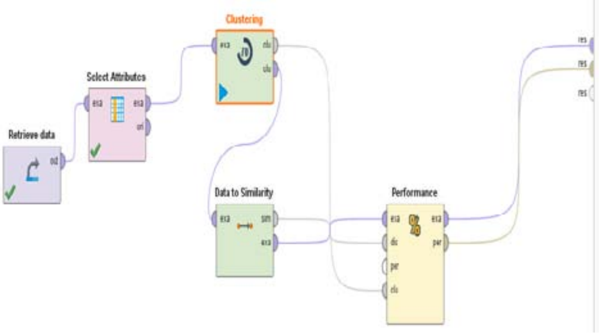

Figure 5: Clustering using K-medoids

C. New Algorithm--X means Clustering Algorithm

It determines the number of centroids based on some heuristic. It starts with a minimum collection of centroids and after that exploits if using more centroids makes sense according to the data.

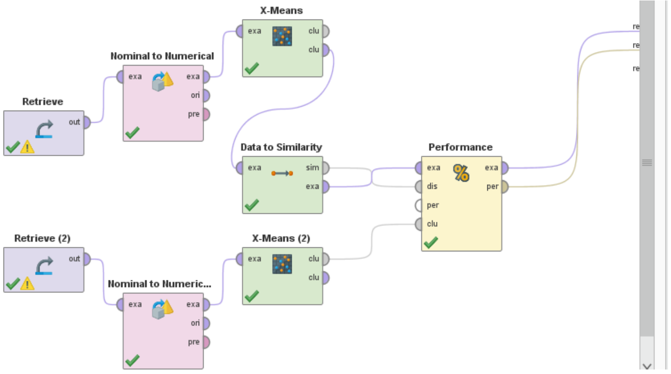

Figure 6: Clustering using X-means

5. RESULTS AND DISCUSSIONS

This section presents the results of all three algorithms used for clustering the dataset. The results are shown in the form of graphs and tables to make it easy to understand and evaluate the performance of each algorithm.

D. Results of K-means

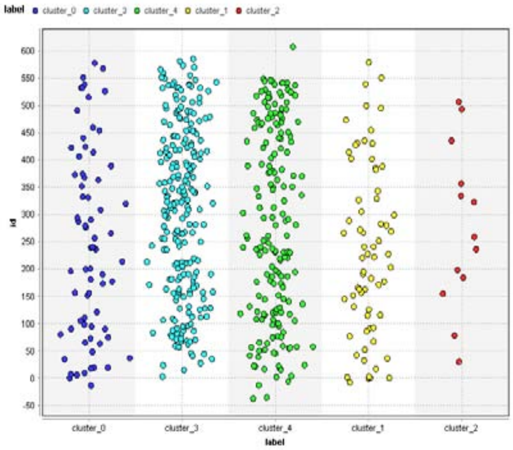

Figure 7: Clustering using K-means based on Diagnosis attribute

The above figure shows the clustering based on Diagnosis attribute. It is set as label attribute in the Rapid miner.

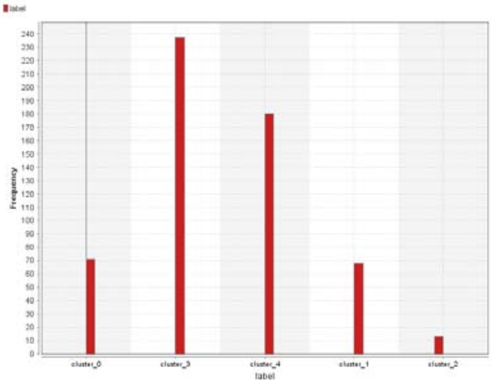

Figure 8: Number of frequencies using K-means

The histogram represents the number of frequencies in each cluster according to k-means clustering algorithm.

E. Result of K-medoids

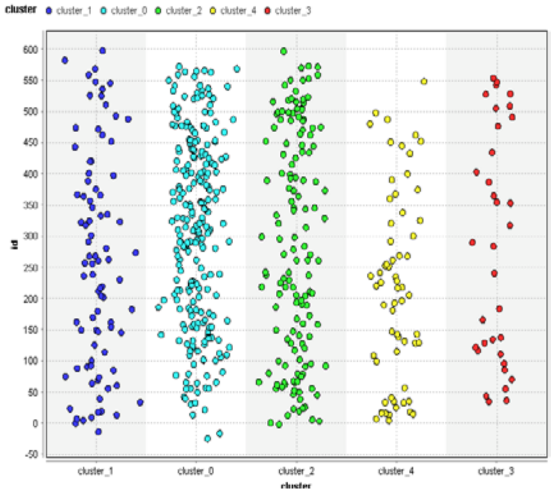

Figure 9: Clustering using K-medoids based on Diagnosis attribute

The above figure shows the clustering based on Diagnosis attribute. It is set as cluster attribute in the Rapid miner.

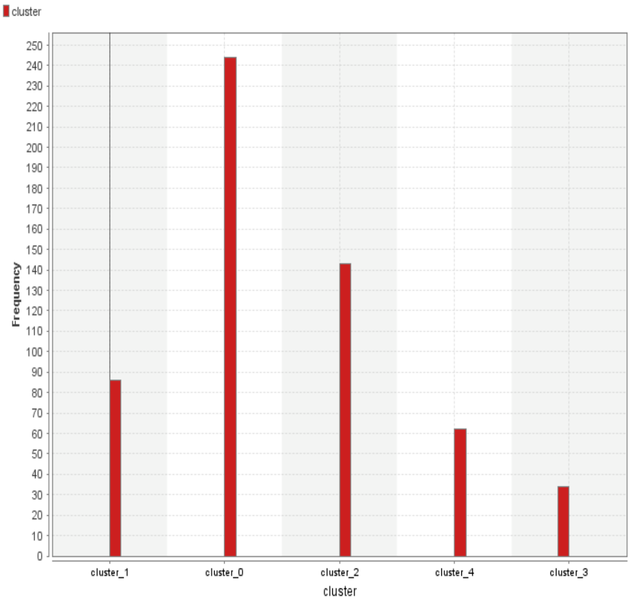

Figure 10: Number of frequencies using K-medoids

The histogram represents the number of frequencies in each cluster according to the k-medoids clustering algorithm.

F. Results of X-means

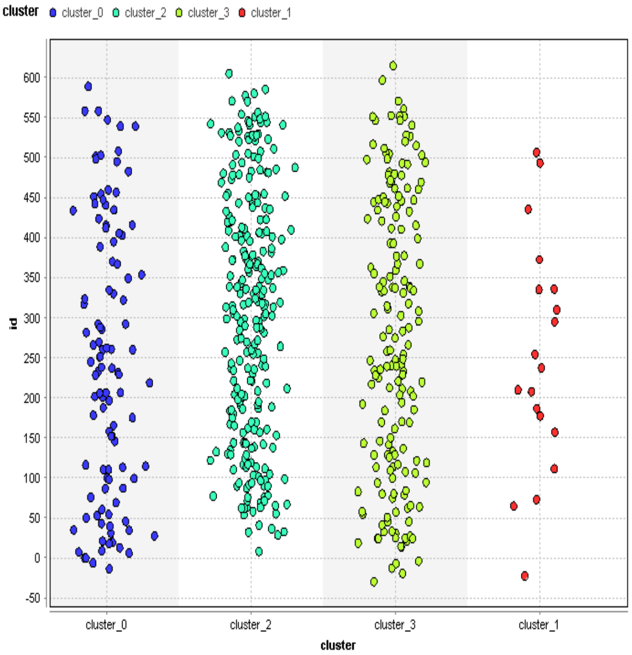

Figure 11: Clustering using X-means based on Diagnosis attribute

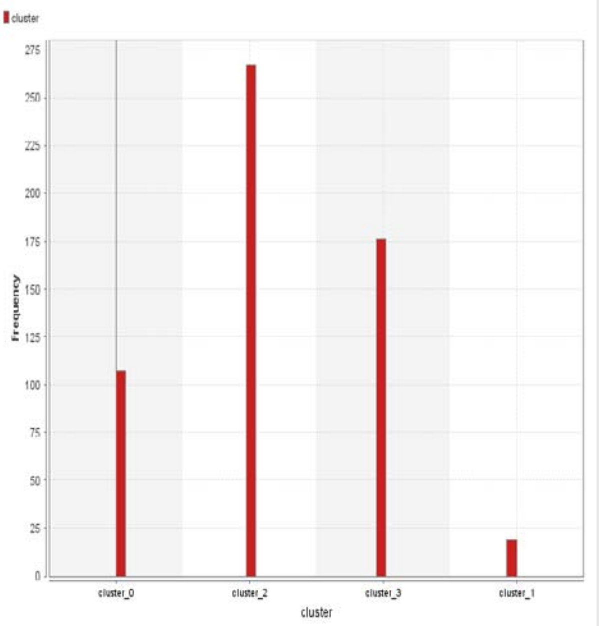

Figure 12: Number of frequencies using X-means

The histogram represents the number of frequencies in each cluster according to the X-means clustering algorithm.

G. Comparison of all algorithms

The table below shows the comparison of all three clustering algorithms. The comparison made in terms of the average distance between the clusters generated by each algorithm.

Table 1: Comparison of all algorithms

Cluster	K-means	K-medoids	X-means
0	-11562.153	-32833.76	-13971.576
1	-35164.441	-53873.488	-41691.301
2	-29445.092	-21869.375	-36330.869
3	-16113.465	-2166.247	-43494.313
4	-21806.153	-13386.599	-

Table 2: Comparison of all algorithms in terms of average distance within cluster

	K-means	K-medoids	X-means
Avg. within cluster distance	-28764.724	-29328.119	-39953.679

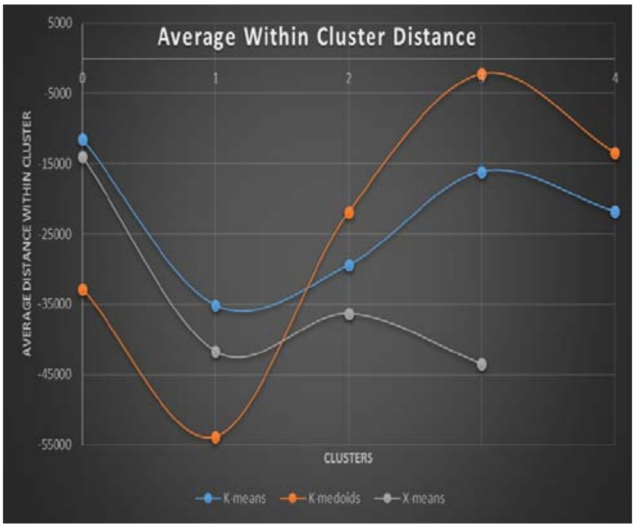

Figure 13: Comparison of all algorithms in terms of average distance within cluster

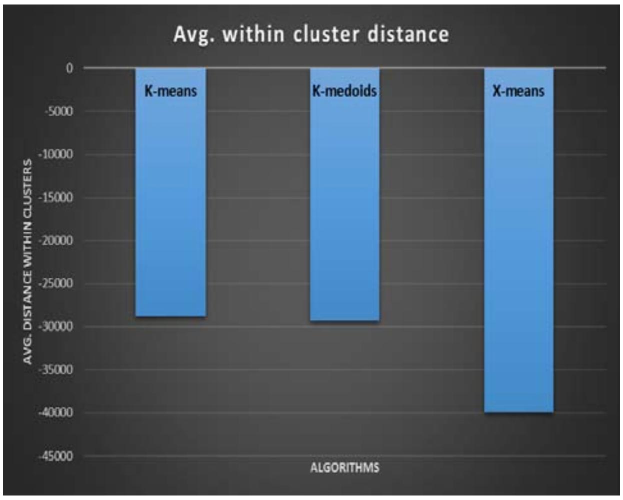

Figure 14: Histogram of all algorithms in terms of average distance within cluster

The two plots in figure 13 and 14 show the results obtained from each clustering algorithms in terms of average distance between the clusters generated by them. X-means clustering algorithm clustered the data with -28764.724 of the average distance between each cluster, whereas K-medoids clustered data with -29328.119 average distance in each cluster. From the above graphs, it can be concluded that X-means clustering algorithm has shown prominent results as compared to other two algorithms, as it clustered the data with -39953.679 average distance in each cluster.

6. CONCLUSION

With the rapid enhancement in population, there is a significant quantity of growth in the health-linked diseases. Several diseases are strongly related to symptoms that create it complicated for the doctors to forecast the precise diseases on one go. This is where 'Data mining' technique appears into the backing, which helps in predicting the disease, which is almost perfect. This research focused on various data mining techniques, their algorithms and related advantages and disadvantages. In this research, we collected Medical data set related to breast cancer disease. This research implemented k-mean, k-medoids, and X-means clustering algorithms using Rapid Miner tool. These three algorithms were applied to evaluate the performance of each algorithm in terms of clustering the medical data. The results were compared to each algorithm in terms of average distance between each cluster. It can be observed from the results that the algorithms k-means and k-medoids showed very little difference between the clusters according to average distance. Whereas the X-means algorithm divides the whole dataset into 4 clusters only based on their similarity with one another. So, it can be concluded that X-means algorithm can be used to cluster the records in fewer clusters as compared to rest of the two algorithms.

7. REFERENCES

[1]. Ashish Dutt and Saeed Aghabozrgi." Clustering algorithm applied in educational data mining" International journal of information and electronic engineering, Vol. 5, No.2, March 2015.

[2]. Berkhin, P. (n.d.), "Survey of Clustering Data Mining Techniques." Accrue Software, Inc., 56.

[3]. Erich Schubert and Alexander Koos." Framework for clustering uncertain data." An international journal, 2015.

[4]. Pan Deng and Feng Chen." Data mining for the internet of things: a literature review." International Journal of Distributed Sensor, 2015.

[5]. Ramageri, B.M. "*DATA MINING TECHNIQUES AND APPLICATIONS," Indian Journal of Computer Science and Engineering*, vol. 1, no. 4, pp. 301-305, 2011.

[6]. Sankar Rajagopal," Customer data clustering using data mining technique." International journal of data management system, Vol.3, No.4, Nov. 2011.

[7]. Srideivanai Nagarajan and R.M. Chandrasekaran," Design and implementation of an expert clinical system for diagnosing diabetes using data mining technique." Indian journal of science and technology, Vol.8, PP. 771-776, April 2015.

[8]. S.R.Pande, V.M.Thakre," Data clustering using data mining technique." International Journal of advanced research in computer and Communication Engineering(IJARCCE), Vol.1, Issue 8, Oct. 2012.

[9]. A. Shabbirl, Y. T. Ansad, A.H. Kazim, and A. EI-Hassan, "Predictive Data Mining and pattern recognition in the medical sector: Implementation and experience." Computer Applications and Information Systems, 2014.