MINAR International Journal of Applied Sciences and Technology

Article type : **Research Article**

Date Received : **07/04/2021**
Date Accepted : **25/04/2021**
Date published : **01/06/2021**
: **www.minarjournal.com**

http://dx.doi.org/10.47832/2717-8234.2-3.11

DATA MINING IN HEALTHCARE SECTOR

Sura I. Mohammed ALI[1] & Rafid Habib BUTI [2]

Abstract

Disease detection is one of the applications where data mining techniques achieved more accurate and useful results. The healthcare sector collects massive volumes of healthcare data that are not mine to discover hidden data for better decision-making, a field of data mining introduces more efficiently and effectively to predict different kinds of diseases.

Clustering medical data into small, meaningful chunks will help in pattern discovery by allowing for the retrieval of a large number of specific data points.

The difference in using clustering the medical data from traditional data mining techniques is in extracting many features of the dataset that have been split into small segments to enable us to discover patterns by adding the data structure. By using clustering techniques, discovered overall correlations between data attributes. Selected data processing makes the mining process more efficient.

The processed disease data are clustered using the K-means algorithm with the K values. Its ease of use and speed, which enable it to perform on a massive dataset. This paper highlights the theoretical side in using the K-Means Clustering algorithm in the context of data mining of disease detection and allowing for reliable and effective diagnosis.

Keywords: Data Mining, Healthcare Sector ,K-Means Clustering.

[1] Al-Muthanna University, Iraq, suraibraheem@mu.edu.iq, https://orcid.org/0000-0003-4265-3731
[2] Al-Muthanna University, Iraq, sci.rafid@mu.edu.iq, https://orcid.org/0000-0002-5849-0327

1. Introduction

Data mining is the method of collecting valuable information from massive databases(Nishara Banu & Gomathy, 2014). The increasing availability of information technology, as well as the simplicity with which it is available, has resulted in a pre-emptive increase in the amount of information have never seen in history, causing the problem of hug data on the internet a source of contention and when we talk about huge data, we're talking about massive amounts of information.

Here, data mining appeared as a methodology aimed at extracting information from massive volumes of data, Data mining is built on mathematical algorithms that are focused on many sciences, including logic, statistics, arithmetic, learning technologies, artificial intelligence, expert systems, and physics. Machine learning, pattern analysis, and other intelligent and non-traditional technologies are examples of this.

The medical sector is capable of making important contributions in the field of data analysis, which results in disease prediction and improved patient care. The found knowledge will be used by healthcare administrators to provide quality treatment. Association Rule Mining, Clustering, and Classification algorithms are examples of data mining techniques. To evaluate various types of heart-related problems, decision trees, and the C4.5 Algorithm are used. aggregation Data mining methods such as K-Means are used in the medical sector [2] to investigate various types of heart-related issues. Data mining methods are shown in Figure 1.

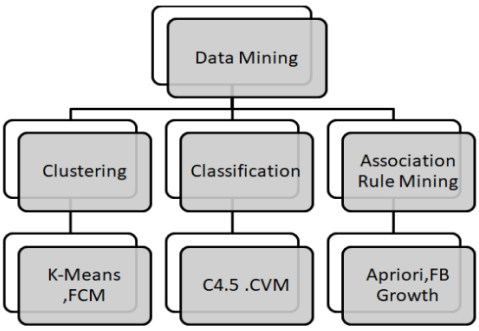

Figure 1: Data mining techniques

Proper clustering is the method of extracting information from large amounts of unstructured data. Since then, thousands of clustering algorithms have been published; K-means is the most commonly used, with a broad variety of applications. In terms of simplicity, speed, and robustness, the algorithm is advantageous. As compared to other clustering algorithms, it is relatively effective(Mittal et al., 2019).
However, unlike traditional K-means, the proposed algorithm does not measure all data points since we use intervals to identify the points that have the potential to shift cluster, this approach calculates only the inner interval points.
The rest of the paper is structured as follows: Section II provides an outline of data clustering techniques and the principles that explain them... Section III presents one of the major phases within the knowledge discovery process is data preprocessing. Section IV introduces K-mean clustering. Section V introduces the conclusion.

I. Related Works

Various proposed used data mining methods to analyze diseases such as cancer, diabetes, hepatitis, and heart disease in the data mining. (De Vos & Soens, 2008) They suggested using association rule mining to predict heart disease and created when association rules were applied to datasets, a large number of rules were produced. In(M Anbarasi, E Anupriya, 2010) authors used a genetic algorithm to minimize the real data size to select the best subset of attributed enough for heart disease prediction. Traditional clustering strategies create partitions; each pattern There is just one cluster in a partition. The k-modes algorithm (Banerjee & Ghosh, 2006) is a recent partitioning algorithm that deals with categorical attributes using the basic matching coefficient calculation. The Fuzzy C-Means (FCM) algorithm (Wagstaff et al., 2001), which is built on k-means, is one commonly used algorithm. FCM tries to locate the most defining point in each cluster, which can be treated as the cluster's "middle," and then, the membership grade for each instance in the clusters. In (Taniar, 2008) the author showed a close interaction between kernel k-means and spectral clustering.
In(Peña et al., 1999) authors suggested the most popular initialization. This approach chooses K points at random from the data set to act as centroids. The method's key advantage is its flexibility and the ability to occupy a significant portion of the solution space by several initializations of the algorithm.
Primarily focused on an initial cluster and centroids with an efficient interval, this paper suggested an adjusted k-means clustering algorithm. The standard k-means algorithm is used as input to the initial centroids and initial cluster data, which then uses the points with the ability to adjust the cluster to have better performance than other methods.

II. Materials And Method

a. Data Preprocessing

Raw data usually contains several weaknesses such as contradictions, missed values, redundancies, and/or noise. If subsequent learning algorithms are provided with low-quality results, their output may suffer. As a result, by performing appropriate preprocessing procedures, we can greatly affect the accuracy and efficiency of subsequent automated discoveries and decisions. One of the most important stages in the information discovery process is data preprocessing. Data preprocessing, while about as well as other steps such as data mining, also requires additional effort and time during the entire data analysis period (more than 50% of total effort. Preprocessing starts by selecting an attribute to choose a subset of attributes with high predictive capabilities(Nishara Banu & Gomathy, 2014), it deals with all lost qualities and explores all possibilities. Where more than 5% of an attribute's values are absent, the record should not be deleted; instead, values should be imputed where data is missing to use a suitable method.

b. Clustering

Clustering is a data mining method used to reveal patterns of importance in a given dataset. Clustering is an effective method for separating groups or groups of points Clustering should be used as a preprocessing method because it is less costly. for attribute subset collection and classification. For example, to categorize genes with identical functionalities, or to shape groups of customers based on purchasing patterns(Ahmed & Hannan, 2012). Data mining is concerned with massive datasets that place constraints on clustering. The clustering issue has been discussed in a variety of contexts and by researchers from a variety of fields and is known to be useful in a variety of medical applications. Reflecting its wide interest and utility as a phase in exploratory data analysis. Partitioning Methods is one of the types of clustering methods. Clustering medical data into tiny, meaningful chunks can help with pattern discovery by allowing extraction of different related features from each cluster, incorporating structure into the data, and facilitating the implementation of traditional data mining techniques. There are several clustering techniques available in the literature, including the well-known K-Means clustering approach. The Kmeans algorithm with K values is used to

cluster the preprocessed heart disease results. K-Means clustering generates a particular number of disjoint, flat (non-hierarchical) clusters. It is ideal for the formation of globular clusters.

III. Proposed Methodology

a. **K-Means is a computational, unsupervised, non-deterministic system**. K-Means clustering can produce more compact clusters than hierarchical clustering, particularly when the clusters are globular. (SundarV et al., 2012). Many experiments have been performed in the past to develop the K-Mean clustering strategy they attempted to improve the clustering result while still correcting the weaknesses of the already existing methods. We were motivated and study papers and wanted to make certain changes that would make things more functional and smoother. We aim to minimize calculation by locating the most practicable points of operation as well as to a marked change in the performance of the clusters.

b. Adjusted K-Means Algorithm

Input: S={P1,P2,…,Pn}// A set of P data points.
Output: Number of Clusters //A set of C Clusters

Steps:

First Stage “” Data Assignments”

1) Calculate C $\cong \sqrt{\frac{p}{2}} \ldots . (1)$, C= number of cluste,. P= number of data inputted.
2) Allocate i=1, f=0,75 // numbering of Set,
3) Locate the nearest pair of data points from S. Transfer those points to the current set Ni&
4) Locate the nearest points of pair data of Ni and move it to Ni from S.
5) Repeat step 4 until Ni is full , Ni=($\frac{P}{C}$), if we fix the number of cluster then Ni= ($\frac{P}{C} \times 0.75$)…(2)
6) If Ni is full and s$\neq 0$, then i=i+1, Repet step 4.

Second Stage”initial centroids” or ”Rlocation of means”

7) Determine the nearest centroid Cj for each data point Ni and assign each data point for each cluster, calculate a new centroid for each Cm.

8) Calculate distance d(largestdistance) between each point and cluster center.

9) Collect the data points during the interval. of dI.dI$\times \frac{4}{9}$ …(3)(Fahad & Alam, 2016)
10) Determine the nearest centroid for those points and assign them to the nearest centroid cluster.
11) Collectnew centroid for each Cm.

If, no change in any Cm then go to end,otherwise repeat step 10.

In step 1we definition of cluster number in equation 1,in the this equation, we fix the number of clusters. Next, we will evaluate the basis of input f, in steps 3,4 compute the nearest pair and hold it in Ni,and deleting it from input set P. Then in step 5 we fill Ni by measuring the closest of those pairs. If any data is sent to any Ni, it is removed from P. In step 6 when Ni is complete, we raise the value of x and repeat the loop until there are no more data points in P. The closest pair of points indicatevalue is start centroids. Based on centroids, we assign nearest points into Cm this target finds in step 7. Those are the first clusters. To find the centroids, measure the centre of gravity. When we have determined the center of gravity for each Cm . Each Cm 's centroids are as follows. In steps 8,9 we calculate inner intervial points by equation 3 dI.dI$\times \frac{4}{9}$,and assign this interval points on owner cluster then calculate new centroids, We retain a lot of points outside of the interval in this scale.Those points cluster are unchanged. It improves working of algorithm compared to conventional k-means clustering algorithms.

c. Time Complexity

Its time complexity is O(p*C*I), where p denotes the points of the interval, C denotes the number of clusters, and I represents the number of iterations needed by the algorithm to converge. Where I stopped is where the algorithm decides whether or not to begin clustering. Since C is usually fixed in progress, the algorithm has linear time complexity in the size of the data collection.

The total cost of the algorithm is O(P) since C is much less than P. For this feature, the algorithm saves a lot of time. It minimizes a lot of calculation, because of a lot of points outside of the interval.

d. Space Complexity

It has an O(C + p) space complexity. Where the points of the p interval are. Just the points of such intervals were used in this algorithm. Other points do not normally shift clusters.

Conclusions

The maximum accuracy of disease prediction is often needed for health diagnosis. Many experiments have been conducted in disease prediction using various techniques. In this paper, the Clustering algorithm was adjusted based on an adjusted K-Means algorithm. This paper focuses on interval points or objects and Connectedness to achieve an accurate outcome. Used interval points, which uses Euclidean distance to minimize the number of points that do have a chance in change current cluster, which conducts the closest measurements in the random method and then defines the mean and the centroid value of the points. This study will be followed by the implementation of the k-mean algorithm to enhance clustering accuracy, we will be able to evaluate our algorithm using a randomly generated dataset and compared it to different proposed methods such as the traditional K-means Algorithm, and we will be able to implement our algorithm in the C++ programing language.

References:

Ahmed, A., & Hannan, S. A. (2012). Data Mining Techniques to Find Out Heart Diseases : An Overview. (Sem Qualis) International Journal of Innovative Technology and Exploring Engineering (IJITEE), 1(4), 18–23.

Banerjee, A., & Ghosh, J. (2006). Scalable clustering algorithms with balancing constraints. Data Mining and Knowledge Discovery, 13(3), 365–395. https://doi.org/10.1007/s10618-006-0040-z

De Vos, A., & Soens, N. (2008). The power of career counseling for enhanced talent and knowledge management. Smart Talent Management: Building Knowledge Assets for Competitive Advantage, 17(8), 119–138. https://doi.org/10.4337/9781848442986.00014

Fahad, S. K. A., & Alam, M. (2016). A Modified K-Means Algorithm for Big Data Clustering. 6(4), 129–132.

M Anbarasi, E Anupriya, N. C. S. N. I. (2010). Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm. International Journal of Engineering Science and Technology, 2(10), 5370–5376.

Mittal, K., Aggarwal, G., & Mahajan, P. (2019). Performance study of K-nearest neighbor classifier and K-means clustering for predicting the diagnostic accuracy. International Journal of Information Technology (Singapore), 11(3), 535–540. https://doi.org/10.1007/s41870-018-0233-x

Nishara Banu, M. A., & Gomathy, B. (2014). Disease forecasting system using data mining methods. Proceedings - 2014 International Conference on Intelligent Computing Applications, ICICA 2014, 130–133. https://doi.org/10.1109/ICICA.2014.36

Peña, J. M., Lozano, J. A., & Larrañaga, P. (1999). An empirical comparison of four initialization methods for the K-Means algorithm. Pattern Recognition Letters, 20(10), 1027–1040. https://doi.org/10.1016/S0167-8655(99)00069-0

SundarV, B., Devi, T., & Saravanan, N. (2012). Development of a Data Clustering Algorithm for Predicting Heart. International Journal of Computer Applications, 48(7), 8–13. https://doi.org/10.5120/7358-0095

Taniar, D. (2008). Data mining and knowledge discovery technologies. Data Mining and Knowledge Discovery Technologies, 1–369. https://doi.org/10.4018/978-1-59904-960-1

Wagstaff, K., Cardie, C., Rogers, S., & Schrödl, S. (2001). Constrained K-means Clustering with Background Knowledge. International Conference on Machine Learning ICML, pages, 577–584. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.90.4624&rep=rep1&type=pdf