RESEARCH REVIEW JOURNALS
Volume-04
Issue-05
May-2019

ISSN: 2455-3085 (Online)
RESEARCH REVIEW International Journal of Multidisciplinary
www.rrjournals.com[UGC Listed Journal]

Heart Diseases Prediction Using Unsupervised K-means Clustering Data Mining Technique – An Experimental Approach

[*1]Qureshi Mujtaba Ashraf, [2]Mir Irshad Ahmad & [3]Wani Anwaar Ahmad

[1]*Research Scholar, Department of Information Technology, Mewar University, Gangrar, Chittorgarh (India)*
[2]*Asssistant Professor, Department of Computer Applications, Cluster University, Srinagar (India)*
[3]*Research Scholar, Department of Computer Applications, Mewar University, Gangrar, Chittorgarh (India)*

ARTICLE DETAILS

Article History
Published Online: 25 May 2019

Keywords
Heart Disease Prediction System, K-means Clustering, Data Mining techniques, WEKA tool.

[*]**Corresponding Author**
Email: mujtaba170[at]gmail.com

ABSTRACT

Data mining is an advanced and interactive approach to explore and obtain the useful patterns of information from the large, undiscovered and distant data warehouses. There is enormous raw data available in the form of text, images, and videos in health care management systems. Data is stored in the medical databases after every microsecond. So there is an open confront to extract these hidden and undiscovered patterns to utilize for diagnosing systems in medical industry. Data mining techniques in medical science plays an imperative role to devise predictive systems for efficient diagnosis. In this work a heart diseases predictive system is devised using k-means clustering mining technique to form the high quality and efficient clusters. The primary goal of k-means algorithm is to form groups based on the identical data objects. The model which shows better performance results based on the size of clusters selected, is adopted and proposed in this work. Numbers of clusters i.e. k is varied to select the best clustering size in the proposed system. To perform experimental efforts, weka simulation tool is employed.

1. Introduction

Data mining is the process of discovering the useful patterns and optimistic supportive knowledge from the large and voluminous data warehouses in an efficient and excellent manner. A large amount of medical data related to various dreadful diseases is generated in different formats such as texts, videos and images and loaded over the back end warehouses in hospitals on daily basis. Thanks to data mining community who made possible to extract the useful knowledge from this back end medical warehouses by using various devised mining techniques. The stored data is related to various dreadful diseases; however our main concern is to utilize heart disease datasets to devise prediction model to diagnose heart diseases well in time using k-means clustering technique. In developing as well as under developing countries heart diseases are the main contributors to the death of large human population. In 2008 approximately 17 million people died all over the world due to the cardiovascular diseases. According to the estimation of WHO by 2030, approximately 24 million people will die because of heart diseases [1].

K-means clustering technique is an unsupervised partition based data mining technique primarily focus to increase intra class resemblance and decreases inter class resemblance. The major purpose of K-means clustering is to divide n observations into k clusters. 'K' in k-means clustering refers to the number of clusters to be formed while using dataset/datasets such as k=2 refers to two clusters, k=4 refers to four clusters to be formed. The number of clusters to be formed plays an imperative role to devise the heart disease prediction model. In our experimental work different number of clusters to obtain the best clustering size is formed. We applied processed heart diseases dataset obtained from Jawaharlal Nehru Hospital Srinagar, J&K**.** We used 250 instances of data with 12 attributes. These instances were divided into clusters to form the best cluster size to optimize the output result. In this work we kept on changing the number/size of clusters such as k=2, k=4, k=8, and k=14 to obtain the acceptable predictive results using appropriate cluster size. The high quality clusters are formed refers to more efficient prediction having least ambiguity with more certainty. The primary goal of k-means algorithm is to form groups based on the identical data objects. Numbers of clusters i.e. k is varied to select the best clustering size in the proposed system. Various models are developed during experimental stage. Comparison method is adopted between the performance results shown to obtain the suitable model among the various devised models. However, the model which shows better performance results based on the size of clusters selected, is adopted and proposed in this work. Both merits and demerits are associated as number of cluster size goes on increasing approach. To perform this experimental work weka simulation tool is used to develop models and to select the appropriate heart disease prediction system. The diagram 1.1 shows the flowchart for the proposed experimental work used to devise and diagnosis model for heat diseases.

RRIJM 2015, All Rights Reserved

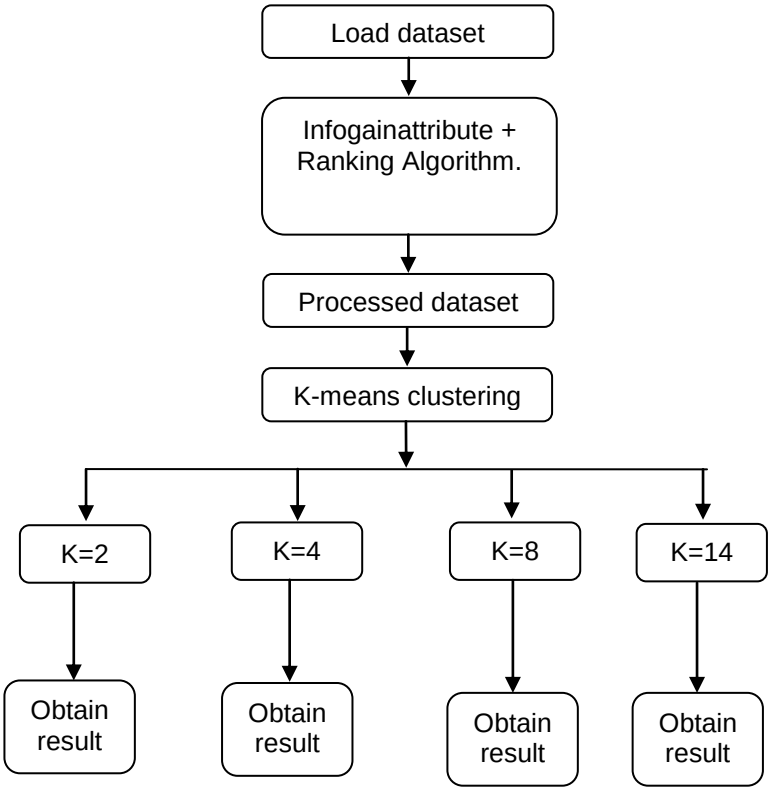

Diagram 1.1: Flowchart for the Proposed Model using K-means algorithm.

2. Literature Review

A model for the prediction of cardiovascular diseases is developed by means of NN, K-means clustering and FIS generation techniques *[2].* Different researchers used different number of attributes to perform their work. Here 13 attributes are selected and used to devise the model to diagnose heart diseases.

Using datasets of CVDs the act of clustering algorithm is proposed by Pandey et.al. Various clustering algorithms are evaluated to find the accuracy and prediction power. The system developed i.e. Make Density Based Cluster with the 85.8086% accuracy, as the most adaptable algorithm to predict heart diseases *[3].*

Singh, et al. *[4]* employed two techniques of clustering namely K-means and hierarchical clustering. K-means clustering show more efficiency and come tighter quickly upon the application of large data sets. Hierarchical clustering develops clusters either by combing smaller ones into larger or opening up of large cluster into smaller clusters. Weka simulation tool is used to compare the accuracy and run time between the two employed clustering techniques.

Shekar et al develops an algorithm to extract association rules from the database of hospitals by using digital sequence mechanism. Clustering is used to divide large databases and group the data points into separate clusters to predict heart diseases. The load problem over the main memory diminishes very much as the only small clusters are used which efficient and scalable *[5].*

K-means clustering technique is used to extract the useful patterns of data from the medical databases to predict heart diseases. MAFIA stands for Maximal Frequent Item set Algorithm plays a vital role to calculate the weightage of the useful patterns for the diagnosis of heart diseases *[6].*

In *[7]* researcher used the clustering and association rules to predict and diagnose heart diseases. This process was implemented using C programming language and also reduced the main memory requirement by forming small clusters of data.

In *[8]* heart disease prediction model is developed using K-means clustering with decision tree. They proposed several methods of the selection of centriods to enhance efficiency and accuracy. In addition they integrate decision tree with K-means clustering to achieve better results to diagnose heart diseases. By the selection of two clusters 83.9% accuracy is achieved by the application of enabler method.

3. Methodology

3.1 Material Used

The heart disease datasets required to perform experimental work were collected from Jawaharlal Nehru Hospital Srinagar, J&K. The collected dataset consists of 1600 instances with 14 attributes. The table 1.0 contains the collected attribute details of dataset. The collected datasets belongs to both genders.

Table 1.0: Attributes Used in Experimental Approach.

Attributes	Description
Patient Id	Dummy values are used.
Gender	Defines gender of a patient, i.e; 1= male or 0= female.

RRIJM 2015, All Rights Reserved

Age	Child, young old (in years)
Exang.	Exercise induced angina.
Old peak	Depression induced angina.
Heart rate	To know the frequency of heart rate(max. achieved)
Chol.	To know the levels of various types of cholesterol like LDL,HDL
fbs	To check whether fasting blood sugar crossed the risk level, i.e.; if >120 mg/dl increases risk.
rbs	Resting blood sugar
trestbps	Blood pressure, To check the range of systolic.
Obesity	Whether a person is obese or not calculated using BMI.
restecg	Normal - ST_T wave Abnormality, Left Ventricular Hypertrophy.
smoking	Yes or No.
Thal.	Blood flow in the heart.

3.2 Pre-processing

In this work, primary focus is made to achieve the real and pure datasets to devise efficient heart diseases prediction model. K-means algorithm is very susceptible to outliers. Outliers are located very distant from the center of cluster and thus affect significantly the mean value of the cluster. Preprocessing lays down a well-organized foundation stone to acquire successful results in the process of data mining. The proper and efficient selection of attributes enhances more chances to obtain a flourishing diagnosis model and lessens the misclassification error rate in both the cases. In this paper we employed the information gain (infogainattribute) method with ranker search to estimate the value of an attribute by measuring the information gain with respect to class and ranks attributes by their individual evaluations respectively. By the process of pre-processing we detached the recurring, missing and error prone instances to augment the quality of results and achieved 200 instances with 12 attributes to perform the experimental work.

3.3 K-means Clustering Algorithm

Clustering refers that objects of the same cluster resembles with each other. K means Clustering algorithm is developed by Macqueen in 1967. "K-means algorithm *[9] [10]* is well-known clustering algorithm extensively used in data mining technology." This technique used falls under the partition clustering algorithm. K-means algorithm is an unsupervised learning algorithm used for unlabeled data. Primary goal of clustering is to divide data into identical groups and show more resemblance with data objects within the same group when comparison is made. Each group formed in this algorithm is represented by the K variable. K-means clustering algorithm results are;

- To find centroids of the formed K clusters used to tag new data items (k refers to number of clusters).
- Every data point present in employed datasets is assigned to related cluster.

The resulting groups are defined by the characteristics of each centroid. To observe the features of centroid weights, qualitative interpretations of the formed groups are obtained.

Consider,

X = {x1, x2, x3,… xn} data points set, and

"V = {v1, v2, v3… vc} set of centers of clusters."

The equation in K-means algorithm used to compute new cluster centers as

$$Vi = (1/Ci) \sum_{j=1}^{Ci} \text{xi}$$

Where, 'ci' symbolize the number of points of data in *ith* cluster

The Algorithm used for k-means clustering in every iteration i.e. as the value of k changes, is presented below.

Algorithmic steps for proposed model using K-means clustering.

Step 1: Select dataset.
Step 2: Apply preprocessing methods to obtain processed dataset.
Step 3: Cluster centers are chosen by K or c points randomly (*Input cluster number to be formed).*
Step 4: Apply Euclidean distance function to gauge the distance between center of cluster and data point
Step 5: Assign data point to the cluster whose distance is minimum from centre of cluster to data point.
Step 5: Using mean value of object to gauge new centers.
Step 6: Execute 2, 3, 4 and 5 step in rounded fashion until same factors are assigned to the each correlated clusters.
Step 7: Repeat from Step 1 to Step 6 for every new cluster size formation.

In conclusion, this algorithm aims at minimizing squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\left\| x_i - v_j \right\| \right)^2$$

Where,

'||xi - vj||' is the Euclidean distance function amid xi and vj.
'ci' is the data points number in ith cluster.
'c' is the figure of cluster centers.

At the very beginning the collected data is loaded in to the WEKA simulation tool. As we know K-means clustering algorithm is affected very much due to the presence of outliers which upset our expected results heavily. So it becomes indispensable that loaded dataset is encountered using feature selection and ranking algorithms to enhance the clarity and efficiency of the loaded dataset. The clean and non-ambiguity obtained dataset is used to develop various models using K-means clustering data mining algorithm. Here various models are devised by changing the number of clusters i.e. k=2, k=4, k=8, and k=14 in k-means algorithm and the results of every model is recorded for the purpose of comparison study performance.

RRIJM 2015, All Rights Reserved

4. Experimental Results and Discussion

The experimental evaluation is conceded out on the proposed model using the JRLN hospital dataset, to authenticate the viability and legitimacy of the models for the diagnosis of the heart diseases. A practical and comparative study is performed between the results shown by the proposed models based on the different size of clusters formed. Experimental results achieved by the selection of various clusters using K-means clustering technique in separate fashion are presented in the table 1.1 and diagram 1.2.

Table 1.1: Comparative results using various cluster numbers.

Technique used	Performance Measures	K=2	K=4	K=8	K=14
K-means Clustering Technique	Squared error (in %)	27.45	23.09	14.23	3.04
	Time taken (in seconds)	0.0	0.02	0.04	0.05
	Number of iterations	03	06	07	09

The diagram 1.2 represents the performance measures, shown by the k-means clustering technique by the submission of diverse clusters, in graphical form

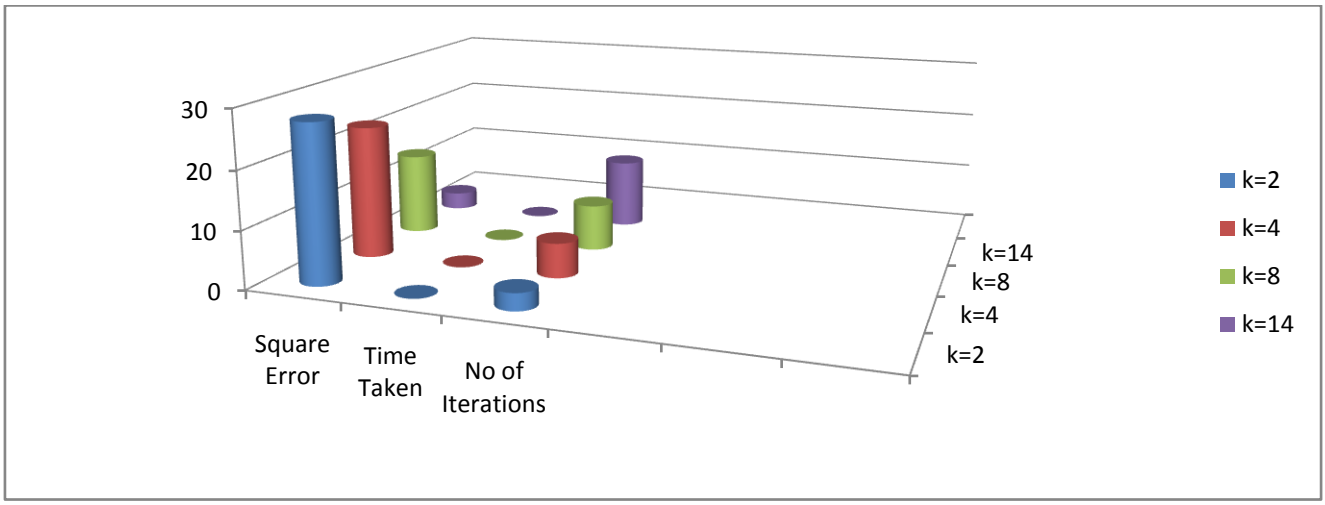

Diagram 1.2: Comparative results using various cluster numbers in graphical form.

In this experimental work we developed predictive models using clustering technique by varying the size of clusters in detached approach. The attainment of the appropriate cluster formation plays a primary role to develop a successful predictive model. When K-means clustering is employed to devise predictive models, the foremost attention in our work is given to;

- To employ processed datasets.
- Selection of appropriate number of cluster size.
- Assign data point to the cluster whose distance is minimum from centre of cluster to data point (minimizing squared error function).
- Time taken to build a model
- Number of iterations.

We shaped various cluster numbers as K=2, K=4, K=8 and K=14 to accomplish the satisfactory results on the basis of above mentioned properties using comparison approaches between the models. As we begin to devise models we constantly altered the number of cluster size as K=2, K=4, K=8 and K=14. We experienced higher mean squared error when the cluster size is set to two i.e. K=2 but zero time, very less ambiguity with least number of iterations are achieved. As the size of clusters is augmented to 4, 8 and 14, no doubt mean squared error rate begins to diminish but complexity of clusters, time taken to build a model and number of iterations begins to accelerate.

5. Conclusion

Heart disease is one of the foremost causes of death globally and in the early hours the detection of heart disease is imperative. The modern heart prediction systems help medical professionals to detect heart diseases with enhance accuracy. Researchers developed a range of heart diseases prediction models using diverse data mining techniques and also attained good success. In this research paper k-means clustering technique is employed to devise various heart disease prediction models based on the size of clusters and to select the most effective one among them. Four models are developed based the number of clusters as 2, 4, 8 and 14 i.e. k=2, k=4, k=8 and k=14. The model based on the number of clusters as 8 i.e. k=8 is selected and adopted as the most appropriate based on the performance results. The model based on the more number of clusters shows good accuracy but more number of iterations and time and very high ambiguity. To select medium number of clusters is suitable to develop a more successful heart disease prediction model. Also number of clusters to be formed depends upon the size of dataset.

Acknowledgement

I am very thankful to Assistant Professor Mr. Irshad Ahmad Mir and Research Scholar Mr Anwaar Ahmad of computer application department of Mewar University for their support and assistance to make this research paper successful.

RRIJM 2015, All Rights Reserved

References

1. Miss. Chaitrali S. Dangare, Dr. Mrs. Sulabha S. Apte, A data mining approach for prediction of heart disease using neural networks, international journal of computer engineering and technology, 2012.
2. Andrea D'Souza,"Heart Disease Prediction Using Data Mining Techniques Andrea ",International Journal of Research in Engineering and Science (IJRES) ,ISSN (Online): 2320-9364, Volume 3 Issue 3 , March. 2015 , PP.74-77
3. Shafranovich. "Common Format and MIME Type for CommaSeparated Values (CSV) Files" (http:/ / tools. ietf. org/ html/ rfc4180) Retrieved September 12, 2011
4. Sellappan Palaniappan and Rafiah Awang, "Intelligent Heart Disease Prediction System using Data Mining Techniques", International Journal of Computer Science and Network Security, Vol. 8, No. 8, pp. 1-6, 2008.
5. K.Shekar, N.Deepika and D.Sujatha,"Association rule for classification of heart-attack patients", International Journal of Advanced Engineering Sciences and Technologies, vol.11, no. 2, pp.253-257, 2011.
6. S. B. Patil and Y. S. Kumaraswamy, "Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction," International Journal of Computer Science and Network Security (IJCSNS), vol. 9, no. 2, pp. 228–235, 2009.
7. M. Jabbar, P. Chandra, and B. Deekshatulu, "CLUSTER BASED ASSOCIATION RULE MINING FOR," Journal of Theoretical & Applied Information Technology, vol. 32, no. 2, pp. 196–201, 2011.
8. Mai Shouman, Tim Turner and Rob Stocker, "Integrating Decision Tree and K-Means Clustering with Different Initial Centroid Selection Methods in the Diagnosis of Heart Disease Patients", Proceedings of the International Conference on Data Mining, 2012.
9. Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43-48.
10. Masethe, H. D., & Masethe, M. A. (2014, October). Prediction of heart disease using classification algorithms. In Proceedings of the world congress on engineering and computer science (Vol. 2, pp. 22-24).

RRIJM 2015, All Rights Reserved