Heart Disease Prediction with Data Mining Clustering Algorithms

Mirpouya Mirmozaffari[1], Alireza Alinezhad[2], and Azadeh Gilanpour[3]

***Abstract*—**Vast number of people annually suffer from heart malfunction worldwide. Various symptoms result in heart disease which in many cases is hard to diagnose a patient as a heart patient. Data mining, as a solution to extract hidden pattern from the clinical dataset are applied to a database in this research. The database consists of 209 instances and 8 attributes. All available algorithms in clustering technique, are compared to achieve the highest accuracy. To further increase the accuracy of the solution, the dataset is preprocessed by different supervised and unsupervised algorithms. The system was implemented in WEKA and prediction accuracy for 5 stages, and 40 approaches, are compared. Three clusters with an accuracy of 100% are introduced as the highest performance algorithms.

***Keywords*—** Data mining, Clustering, WEKA.

I. INTRODUCTION

AN estimated 17.3 million people annually die from heart disease worldwide [1]. Medical practitioners conduct different surveys on heart diseases and gather information of heart patients, their symptoms and disease progression. Increasingly have been reported patients who suffers from common symptoms. Accordingly, there is valuable information hidden in their dataset to be extracted.

Data mining is the technique of extracting hidden information from a large set of database [2]. It helps researchers to gain novel and profound insights into large medical datasets. The principal goals of data mining are prediction and description of diseases. It is achieved through processing of a set of variables (attributes) in the dataset and finding the future states of remainder variables.

To find the unknown trends in heart disease, all the available clustering algorithms are applied to a unique dataset and their accuracy are compared. A dataset of 209 instances and 8 attributes (7 inputs and 1 output) are used to test and justify the differences between algorithms. To further enhance accuracy and achieve more reliable variables, the dataset is purified by supervised and unsupervised filters. Finally, the optimum algorithms in clustering technique with the highest accuracy are introduced.

Mirpouya Mirmozaffari[1], Msc. student, Faculty of Industrial and Mechanical Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran

Alireza Alinezhad[2], Associate Professor, Faculty of Industrial and Mechanical Engineering, Qazvin Branch, Islamic Azad University, Qazvin,Iran

Azadeh Gilanpour[3], Islamic Azad University (IAU).

II. BACKGROUND AND LITERATURE REVIEW

An increasing number of heart patients worldwide have motivated researchers to do comprehensive research to reveal hidden patterns in clinical datasets. An overview of reported computational studies on pattern recognition in heart disease is covered in this section. Not only are different techniques addressed, but also various heart disease datasets are provided. Finally, the gap in existing literature, which was the main motivation of this study is also mentioned. Some of the key studies are as follows:

- Pandey et al. proposed the performance of clustering algorithm using heart disease dataset. They evaluated the performance and prediction accuracy of some clustering algorithms. The performance of clusters will be calculated using the mode of classes to clusters evaluation. Finally, they proposed Make Density Based Cluster with the prediction accuracy of 85.8086%, as the most versatile algorithm for heart disease diagnosis [3].
- Das et al. introduced a neural network classifier for diagnosing of the valvular heart disease. The ensemble-based methods create new models by combining the posterior probabilities or the predicted values from multiple predecessor models. An effective model has been created and experimentally tested. A classification accuracy of 97.4% from the experiment on a dataset containing 215 samples is achieved [4].
- Karaolis et al. developed a data mining system using association analysis based on the Apriori algorithm for the assessment of heart-related risk factors with WEKA tools. A total of 369 cases were collected from the Paphos CHD Survey, most of them with more than one event. Selected rules were evaluated according to the importance of each rule. Each extracted rule was further evaluated by inspection of the number of cases within the database [5].

Therefore, pattern recognition in heart disease can be addressed through different computational techniques. In regard to clustering algorithms, other respected works, focused on diverse aspects of heart disease on different datasets can be mentioned: Shilna et al., 2016 [6]; Lakshmi et al., 2013 [7]; Solanki et al., 2016 [8]. Also, different computational techniques have been applied to other health care issues in the literature [9-11].

It is observed various clusters are frequently utilized in different studies to predict heart disease. Therefore, a comprehensive comparison of clustering algorithms practically provides an insight into clustering performances. This comparison is of great importance to medical practitioners who desire to predict heart failure at a proper step of its progression. Furthermore, except for Ref. [12], which has evaluated 4 classification techniques, there is not any other study on the current dataset. Finally, a unique multilayer filtering in preprocessing step is applied which eventually results in increased accuracy within most of the clustering algorithms, covered in this study.

III. Dataset Description

The standard dataset, compiled in this study contains 209 records, which is collected from a hospital in Iran, under the supervision of National Health Ministry. Data is gathered from a single resource, so it precludes any integration operations. Eight attributes are utilized, from them, 7 are considered as inputs which predict the future state of the attribute "Diagnosis". All the attributes, along with their values and data types are discussed in Table I.

TABLE I
THE ARRANGEMENT OF CHANNELS

Attributes	Descriptions	Encoding\Values	Feature
Age	Age in years	28-66	Numeric
Chest Pain Type	It signals heart attack and has four different conditions: Asymptotic, Atypical Angina, Typical Angina, and without Angina.	Asymptotic = 1 Atypical Angina = 2 Typical Angina = 3 Non-Angina = 4	Nominal
Rest Blood Pressure	Patient's resting blood pressure in mm Hg at the time of admission to the hospital	94-200	Numeric
Blood Sugar	Below 120 mm Hg- Normal Above 120 mm Hg- High	High = 1 Normal = 0	Nominal Binary
Rest Electrocardiographic	Normal, Left Ventricular Hypertrophy (LVH) ST_T wave abnormality	Normal=1 Left Vent Hyper = 2 ST_T wave abnormality = 3	Nominal
Maximum Heart Rate	maximum heart rate attained in sport test	82-188	Numeric
Exercise Angina	It includes two conditions of positive and negative	Positive = 1 Negative = 0	Nominal Binary
Diagnosis	It includes two conditions of positive and negative	Positive = 1 Negative = 0	Nominal Binary

IV. Research Methodology

The objective of this study is to effectively predict possible heart attacks from the patient dataset. Applying a prediction methodology, a model was developed to determine the characteristics of heart disease in terms of some attributes. Data mining in this research is utilized to build models for prediction of the class based on selected attributes. Waikato Environment for knowledge Analysis (WEKA) has been used for prediction due to its proficiency in discovering, analysis and predicting of patterns [13]. Generally, the whole process can be split into two steps as follows:

A. Multilayer filtering preprocess

The data in the real world is highly susceptible to noise, missing, and inconsistency. Therefore, pre-processing of data is very important. We apply a filter on datasets and purify them from dirty and redundant data present in the dataset. Both attribute (attribute manipulation), and instance (instance manipulation) filters in either case of supervised or unsupervised, can be applied in WEKA 2016 (version 3.9.0). In this study, a multilayer filtering process is applied to the dataset to make imbalanced data balanced. This process is implemented in three steps as follows:

- Step A: "Discretization" which is unsupervised attribute filter changes numeric data into nominal.
- Step B: The output of step A is applied to "Attribute selection" which is a supervised attribute filter. "Attribute selection" consists of "Common Features Subset evaluator (Cfs)" and "BestFirst search method".
- Step C: The output of step B is applied to a "Stratified Remove Folds" supervised instance filter.
- Step D: The output of step C is applied to a "Resample" unsupervised instance filter.
- Step E: The output of step D is applied to a "Resample" supervised instance filter.

B. Evaluation in clustering

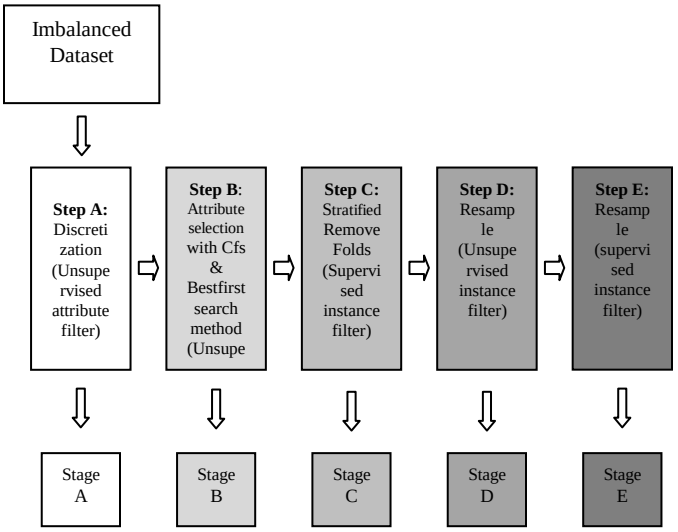

Fig 1: Implementation of clustering Algorithms for accuracy analysis

The "classes to clusters evaluation" is applied as the main evaluator to each output of aforementioned steps. It is used since it is the only clustering evaluator which produces numeric accuracy as a criterion of comparison within different algorithms. Figure 1 elaborates the proposed model and different steps. The combination of each filtering step and each evaluation method results in a different stage. By applying 5

stages to 8 clusters, 40 different approaches are yielded. The accuracy and average accuracy in each stage, are compared in Table II.

TABLE II
ACCURACY COMPARISON WITHIN CLUSTERING ALGORITHMS (ALL NUMBERS ARE IN PERCENT)

Clusters	Stage A	Stage B	Stage C	Stage D	Stage E
1. Canopy	25.3589	47.8469	57.1429	66.6667	71.4286
2. Cobweb	6.6986	47.8469	57.1429	66.6667	71.4286
3.EM (Expectation Maximization)	**78.4689**	78.9474	52.381	90.4762	**100**
4. Farthest First	67.4641	**79.9043**	**90.4762**	90.4762	**100**
5.Filtered Clusterer	65.5502	69.378	80.9524	**95.2381**	**100**
6.Hierarchical Clusterer	55.5024	76.0766	**90.4762**	71.4286	71.4286
7.Make Density Based Clusterer	70.3349	69.378	80.9524	**95.2381**	**100**
8.Simple KMean	65.5502	69.378	80.9524	**95.2381**	**100**
Average of 8 clusters	54.3660	67.3445	73.8095	83.9285	89.2857

V. RESULT AND DISCUSSION

It can be inferred from table II, as the layers of filtering increase:

- The maximum of accuracy within five evaluation methods is increased.
- The average accuracy within 8 clusters, corresponds to each filtering step is increased.

It also should be noted, in each filtering step, from stage A to stage E, the accuracy of most of the clusters are increased. Therefore, to narrow down our study to the most accurate stage, a further comparison on other evaluators of the most accurate algorithms in stage E are provided. Table III compares the time taken to build the models of the five most accurate algorithms in this study with 100% accuracy.

TABLE III
EVALUATION OF THE BEST CLASSIFIERS IN STAGE C3

Clusters	Time taken to build the model (Seconds)
1. EM	0.06
2. Farthest First	0.01
3. Filtered Clusterer	0
4. Make Density Based Clusterer	0
5.Simple KMean	0

Finally, a further comparison on other evaluators of the most accurate algorithms in stage E are provided. Sum of squared errors (SSE) which is 6 and Number of iteration which is 2 are equal in three most accurate algorithms.

VI. CONCLUSION

Various clustering algorithms in data mining were compared to find the most accurate one in heart disease prediction. A unique model consisting of different filters is evolved. Multilayer filtering preprocess, as well as a numeric evaluation method, are applied to find the superior algorithm. Clusters are compared regarding their accuracies, error functions and building times. According to results achieved, Filtered Clusterer, Make Density Based Clusterer, and Simple K-Means with the highest accuracy (100%), the lowest time taken to build the models (0 seconds), equal SSE (6) and equal Number of iteration (2) are the best algorithms. The experiment can serve as a practical tool for physicians to effectively predict dangerous cases and advise accordingly.

REFERENCES

[1] A. K. Sen, S. B. Patel, and D. P. Shukla, "A Data Mining Technique for Prediction of Coronary Heart Disease Using Neuro-Fuzzy Integrated Approach Two Level," International Journal of Engineering and Computer Science, Vol. 2, No. 9, pp. 1663–1671, 2013.

[2] G. Karraz, G. Magenes, "Automatic Classification of Heart beats using Neural Network Classifier based on a Bayesian Frame Work," IEEE, Vol 1, 2006.

[3] A. K. Pandey, P. Pandey, K. L. Jaiswal, and A. K. Sen, "Data Mining Clustering Techniques in the Prediction of Heart Disease using Attribute Selection Method," International Journal of Science, Engineering and Technology Research (IJSETR), ISSN: 2277798, Vol 2, Issue10, October 2013.

[4] R. Das, I. Turkoglu, and A. Sengur, "Diagnosis of valvular heart disease through neural networks ensembles," Elsevier, 2009.

[5] M. Karaolis, J. A. Moutiris, and C. S. Pattichis, "Association rule analysis for the assessment of the risk of coronary heart events," Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2009.

https://doi.org/10.1109/iembs.2009.5334656

[6] S. Shilna and E. Navya, " Heart disease forecasting system using k-means clustering algorithm with PSO and other data mining method," International Journal On Engineering Technology and Sciences (IJETS™), ISSN(P): 2349-3968, ISSN (O): 2349-3976, Vol. 3, Issue 4, April 2016.

[7] K.R. Lakshmi, M. V. Krishna and S. P. Kumar, "Performance Comparison of Data Mining Techniques for Predicting of Heart Disease Survivability," International Journal of Scientific and Research Publications, ISSN 2250-3153, Vol.3, Issue.6, June 2013.

[8] K. Solanki, P. Berwal and S. Dalal, "Analysis of application of data mining techniques in healthcare," International Journal of Computer Applications, Vol. 148, No.2, August 2016.

https://doi.org/10.5120/ijca2016911011

[9] M. Verma, M. Srivastava, N. Chack, A. K. Diswar and Nidhi Gupta," A Comparative Study of Various Clustering Algorithms in Data Mining, "International Journal of Engineering Research and Applications (IJERA), Vol. 2, Issue 3, pp. 1379-1384, 2012.

[10] S. Revathi and T. NalinI, "Performance Comparison of Various Clustering Algorithm," Vol. 3, Issue 2, ISSN: 2277 128X, February 2013.

[11] R. Chauhan, H. Kaur and A. Alam, "Data Clustering Method for Discovering Clusters in Spatial Cancer Databases," International Journal of Computer Applications, Vol. 10, No.6, November 2010.

https://doi.org/10.5120/1487-2004

[12] B. Bahrami, and M. H. Shirvani, "Prediction and Diagnosis of Heart Disease by Data Mining Techniques," Journal of Multidisciplinary Engineering Science and Technology (JMEST), ISSN: 3159-0040, Vol. 2 Issue 2, February 2015.

[13] I. H. Witten and E. Frank, "Data Mining Practical Machine Learning Tools and Techniques," Morgan Kaufman Publishers, 2005.