ISSN: 2278 – 7798
International Journal of Science, Engineering and Technology Research (IJSETR)
Volume 2, Issue 10, October 2013

DataMining Clustering Techniques in the Prediction of Heart Disease using Attribute Selection Method

Atul Kumar Pandey*, Prabhat Pandey, K.L. Jaiswal***, Ashish Kumar Sen******

ABSTRACT—Heart disease is the leading cause of death in the world over the past 10 years. In this paper proposes the performance of clustering algorithm using heart disease data. We are evaluating the performance of clustering algorithms of EM, Cobweb, Farthest First, Make Density Based Clusters, Simple K-Means algorithms. The performance of clusters will be calculated using the mode of classes to clusters evaluation. The selected attributes after the **Common Features Subset Evaluator** (CFs) and **Best-First Search** (BFs) are cp, restecg, thalach, exang, oldpeak, ca, thal, and num. In the final result, Make Density Based Clusters shows the high performance algorithms for heart disease data after applying the Attribute selection Method and their Prediction Accuracy is 85.80%.

KEYWORDS: - EM, Make Density-Based Clusters, Farthest First, K-Mean, and Attribute Selection.

1. INTRODUCTION

In health care the data mining is more popular and essential for all the healthcare applications [22]. It contains the many data, but these data have not been used for some useful purpose. This data will be converted in to the some useful purpose by using data mining techniques. Data mining in healthcare is an emerging field of high importance for providing prognosis and a deeper understanding of medical data. Healthcare data mining attempts to solve real world health problems in the diagnosis and treatment of diseases [1]. Researchers are using data mining techniques in the medical diagnosis of several diseases such as diabetes [2], stroke [3], cancer [4], and heart disease [5].

Heart disease is a general name for a wide variety of diseases, disorders and conditions that affect the heart and sometimes the blood vessels as well [23]. Heart disease is the number one killer of women and men. Symptoms of heart disease vary depending on the specific type of heart disease. A classic symptom of heart disease is chest pain. However, with some forms of heart disease, such as atherosclerosis, there may be no symptoms in some people until life-threatening complications develop. Any of a number of conditions that can be affects the heart. The data mining is the process of finding the hidden knowledge from the data base or any other information repositories. The main purpose of the health care industry is to improving the quality of healthcare data by reducing the missing values and removing the noise in the data base. Several data mining techniques are used in the diagnosis of heart disease such as naïve bayes, decision tree, and neural network, kernel density, bagging algorithm, k-mean clustering and support vector machine showing different levels of accuracies [5-11]

K-means clustering is one of the most popular and well know clustering techniques. Its simplicity and reliable behavior made it popular in many applications [12, 18]. Several researchers have identified that age, blood pressure and cholesterol are critical risk factors associated with heart disease [14, 16-17].

Mr. Atul Kumar Pandey Assistant Professor of Computer Science, Department of Physics, Govt. PG Science College, Rewa(M.P.)-India, Mobile No-09424944538,

Dr. Prabhat Pandey OSD, Additional Directorate, Higher Education, Division Rewa (M.P.)-India, Mobile No-09425183334.,

Dr. K.L. Jaiswal Assistant Professor and In charge of BCA, DCA & PGDCA, Department of Physics, Govt. PG Science College, Rewa(M.P.)-India-486001, Mobile No-09424746167.,

Mr. Ashish Kumar Sen Assistant Professor, Department of Mathematics and Computer Science, Govt. PG Science College, Rewa(M.P.)-India, Mobile No-09893658054,

In identifying the attributes that will be used in the clustering, these attributes are obvious clustering attributes for heart disease patients.

2. BACKGROUND

Researchers have been investigating the use of statistical analysis and data mining techniques to help healthcare professionals in the diagnosis of heart disease. Statistical analysis has identified the risk factors associated with heart disease to be age, blood pressure, smoking [13], cholesterol [15], diabetes [16], and hypertension, family history of heart disease [17], obesity, and lack of physical activity [18]. Knowledge of the risk factors associated with heart disease helps health care professionals to identify patients at high risk of having heart disease.

The Clustering is the process of grouping the similar data items [20]. It is the unsupervised learning techniques, in which the class label will not be provided. The Clustering methods are Partitioned clustering, Hierarchical methods, Density based clustering, Sub Space Clustering. Hierarchical algorithms find successive clusters using previously established clusters. These algorithms usually are either agglomerative ("bottom-up") or divisive ("top-down"). Agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters. Partitioned algorithms typically determine all clusters at once, but can also be used as divisive algorithms in the hierarchical clustering [21]. D e n s i t y - b a s e d clustering algorithms are devised to discover arbitrary-shaped clusters. In this approach, a cluster is regarded as a region in which the density of data objects exceeds a threshold. DBSCAN and OPTICS are two typical algorithms of this kind. Subspace clustering methods look for clusters that can only be seen in a particular projection (subspace, manifold) of the data. These methods thus can ignore irrelevant attributes. The general problem is also known as Correlation clustering while the special case of axis-parallel subspaces is also known as Two-way clustering, co-clustering or bi clustering: in these methods not only the objects are clustered but also the features of the objects, i.e., if the data is represented in a data matrix, the rows and columns are clustered simultaneously [19]. They usually do not however work with arbitrary feature combinations as in general subspace methods. But this special case deserves attention due to its applications in bioinformatics. Conceptual clustering is a machine learning paradigm for unsupervised classification developed mainly during the 1980s. It is distinguished from ordinary data clustering by generating a concept description for each generated class [21]. Most conceptual clustering methods are capable of generating hierarchical category structures; see Categorization for more information on hierarchy. Conceptual clustering is closely related to formal concept analysis, decision tree learning, and mixture model learning.

3. METHODOLOGY

I. HEART DISEASE DATASET

The data used in this study is the Cleveland Clinic Foundation Heart disease data set available at http://archive.ics.uci.edu/ml/datasets /Heart+Disease. The data set has 76 raw attributes. However, all of the published experiments only refer to 13 of them. The data set contains 303 rows of which 297 are complete. Six rows contain missing values and they are removed from the experiment

II. PROPOSED FRAMEWORK

The Proposed Framework has two major categories Attribute Selection method and Clustering Algorithm. In Attribute Selection Method they have two stages first is Common Features Subset evaluator (CFs) and second one is Best-First Search Method. We have applied five clustering algorithms and classified the heart patients in Classes to Clusters Evaluation Mode against the last attribute nom.

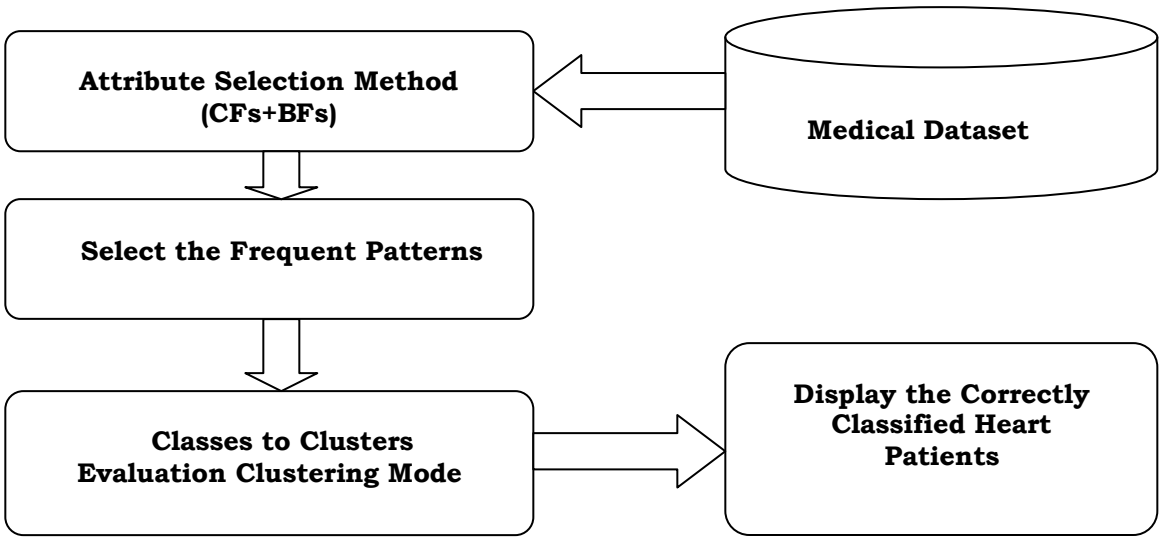

Fig 1: Proposed Framework

III. ATTRIBUTE SUBSET SELECTION METHOD

In weka the preprocessing contains two filters of supervised and unsupervised filters. The attribute selection is one of the supervised filters. A supervised attribute filter that can be used to select attributes [24]. It is very flexible and allows various search and evaluation methods to be combined. In this filter uses the CfsSubsetEval for the evation and the best first for the searching.

Options:-evaluator -- Determines how attributes/attribute subsets are **evaluated search** -- Determines the search method.

[1] CfsSubsetEval

Evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low Inter-co-relation are preferred.

Options:-locallyPredictive -- Identify locally predictive attributes. Iteratively adds attributes with the highest correlation with the class as long as there is not already an attribute in the subset that has a higher correlation with the attribute in question **missingSeparate** --Treat missing as a separate value. Otherwise, counts for missing values are distributed across other values in proportion to their frequency.

[2] Best First

Searches the space of attribute subsets by greedy hill climbing augmented with a backtracking facility. Setting the number of consecutive non-improving nodes allowed controls the level of backtracking done. Best first may start with the empty set of attributes and search forward, or start with the full set of attributes and search backward, or start at any point and search in both directions (by considering all possible single attribute additions and deletions at a given point).

Options:-direction -- Set the direction of the search. **lookupCacheSize** --Set the maximum size of the lookup cache of evaluated subsets.This is expressed as a multiplier of the number of attributes inthe data set. (default = 1).**searchTermination** -- Set the amount of backtracking. Specify the number of **startSet** -- Set the startpoint for the search. This is specified as a comma seperated listoff attribute indexes starting at 1.

4. PERFORMANCE EVALUATION

The table 1 shows the performance of clustering algorithms

using heart disease data. The attributes will be evaluated based on the prediction accuracy of the algorithms.

TABLE 1: PREDICTION ACCURACY OF CLUSTERS BEFORE FEATURES SELECTION METHOD

Clusters Algorithms	Correctly Classified Instance	In correctly Classified Instance	Prediction Accuracy %
COBWEB	6	297	1.9802%
EM	247	56	81.5182
Farthest First	223	80	73.5974
Make Density Based Clusters	247	56	81.5182
Simple K-Means	245	58	80.8581

The following evaluation graph shows the performance of the clustering algorithms in fig 2. In data mining, the clustering algorithms EM and Make Density Based Clusters having the highest prediction accuracy comparing to the remaining clustering algorithms and their accuracy are similar.

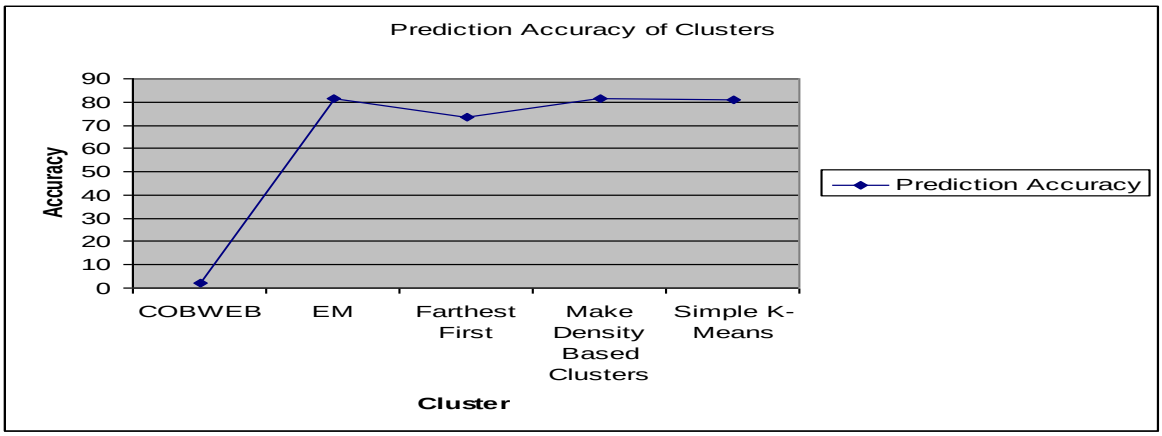

Fig 2:- Evaluation Graph without Attributes Selection Method

TABLE 2: PREDICTION ACCURACY OF CLUSTERS AFTER FEARURE SELECTION METHOD

Clusters Algorithms	Correctly Classified Instance	In correctly Classified Instance	Prediction Accuracy %
COBWEB	10	293	3.3003
EM	250	60	80.198
Farthest First	226	67	77.8878
Make Density Based Clusters	250	43	85.8086
Simple K-Means	248	60	80.198

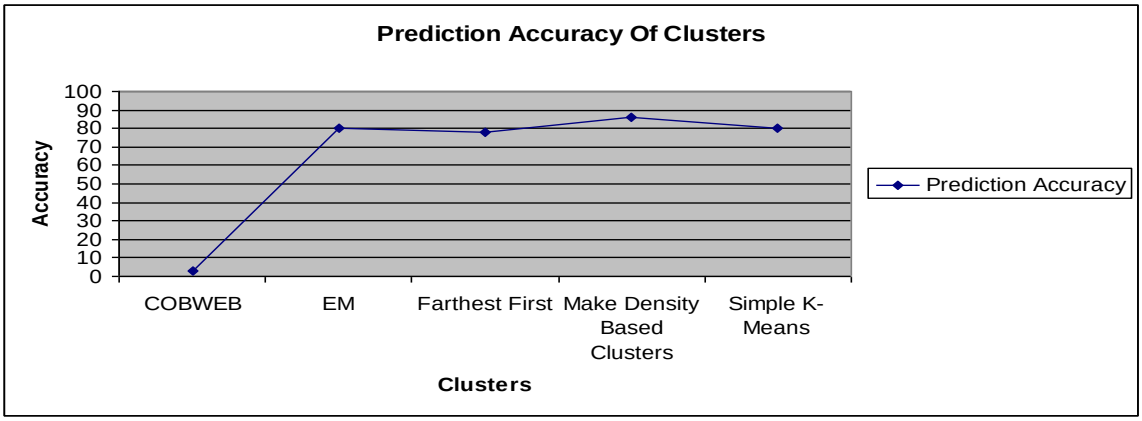

Fig 3: Evaluation Graph after Feature Selection Method

The clustering algorithm, Make Density Based Clusters having the highest prediction accuracy compare to other clustering algorithms after applying the attribute selection method in table 2. In attribute selection method CFs Subset Evaluator and BestFirst search method results the selected eight attributes which gives better results than the previous one. These selected attributes are cp, restecg, thalach, exang, oldpeak, ca, thal, num. Make Density Based Clusters algorithm having the highest prediction accuracy of 85.80%. The evaluation graph shows the performance of the clustering algorithms in fig 3.

In Figure 4, shows the result of Make Density Based Clusters Algorithm on Weka Clusterer Visualize.

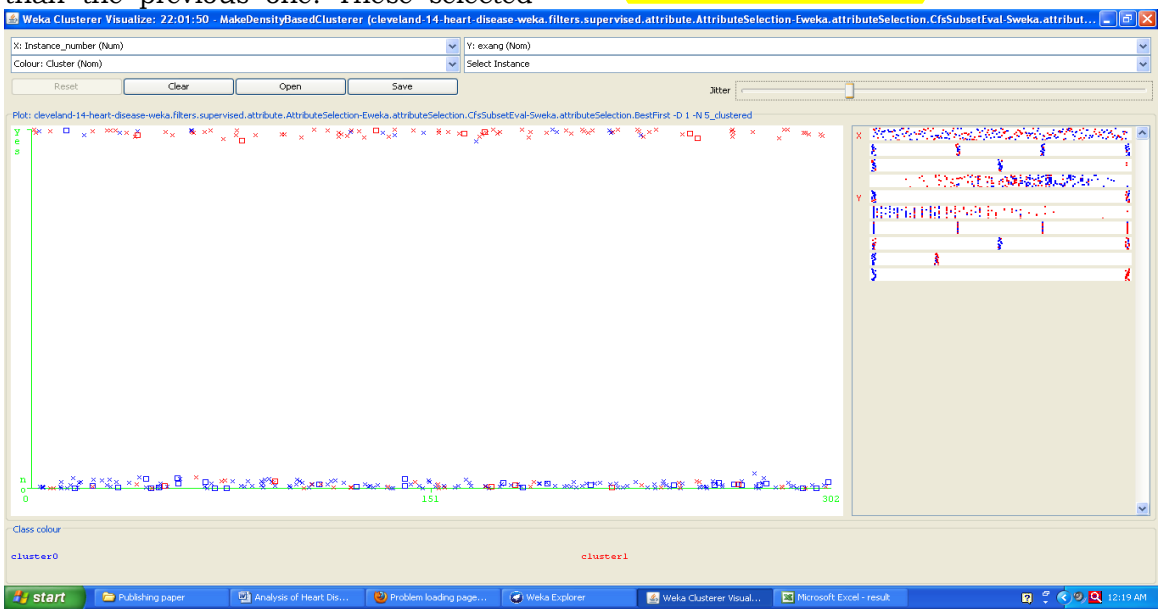

Fig 4: Weka Clusterer Visualize of Make Density Based Cluster Algorithm

5. SUMMARY

Heart disease is the leading cause of death all over the world in the past ten years. Motivated by the world-wide increasing mortality of heart disease patients each year and the availability of huge amount of patients' data that could be used to extract useful knowledge, researchers have been using data mining techniques to help health care professionals in the diagnosis of heart disease. In this paper we have applied the different clustering algorithms on Heart disease dataset. The results shows that EM and Make Density Based Clusters having the highest prediction accuracy comparing to the remaining clustering algorithms and their accuracy are similar to each other without applying the attribute selection method. After applying the attributes selection method CFs Subset evaluator and BestFirst Search method results the selected eight attributes. These selected eight attributes are cp, restecg, thalach, exang, oldpeak, ca, thal, and num. The clustering algorithm **Make Density Based Clusters** having the 85.80% of highest Prediction Accuracy after applying the attributes selection method on a Cleveland dataset to investigate its efficiency in the diagnosis of heart disease. We also investigated if integrating **Attribute Selection** with **Make Density Based Clusters** could enhance its accuracy even further.

REFERENCES

[1] Liao, S.-C. and I.-N. Lee, Appropriate medical data categorization for data mining classification techniques. MED. INFORM., 2002. Vol. 27, no. 1, 59–67,

[2] Porter, T. and B. Green, Identifying Diabetic Patients: A Data Mining Approach. Americas Conference on Information Systems, 2009.

[3] Panzarasa, S., et al., Data mining techniques for analyzing stroke care 2007

processes. Proceedings of the 13th World Congress on Medical Informatics, 2010.
[4] Li L, T.H., Wu Z, Gong J, Gruidl M, Zou J, Tockman M, Clark RA, Data mining techniques for cancer detection using serum proteomic profiling. Artificial Intelligence in Medicine, Elsevier, 2004.
[5] Das, R., I. Turkoglu, and A. Sengur, Effective diagnosis of heart disease through neural networks ensembles. Expert Systems with Applications, Elsevier, 2009. 36 (2009): p. 7675–7680.
[6] Andreeva, P., Data Modelling and Specific Rule Generation via Data Mining Techniques. International Conference on Computer Systems and Technologies - CompSysTech, 2006.
[7] Hara, A. and T. Ichimura, Data Mining by Soft Computing Methods for The Coronary Heart Disease Database. Fourth International Workshop on Computational Intelligence & Applications, IEEE, 2008.
[8] Rajkumar, A. and G.S. Reena, Diagnosis Of Heart Disease Using Datamining Algorithm. Global Journal of Computer Science and Technology, 2010. Vol. 10 (Issue 10).
[9] Sitar-Taut, V.A., et al., Using machine learning algorithms in cardiovascular disease risk evaluation. Journal of Applied Computer Science & Mathematics, 2009.
[10] Srinivas, K., B.K. Rani, and A. Govrdhan, Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. International Journal on Computer Science and Engineering (IJCSE), 2010. Vol. 02, No. 02: p. 250-255.
[11] Yan, H., et al., Development of a decision support system for heart disease diagnosis using multilayer perceptron. Proceedings of the 2003 International Symposium on, 2003. vol.5: p. pp. V-709- V-712.
[12] Wu, X., et al., Top 10 algorithms in data mining analysis. Knowl. Inf. Syst., 2007.
[13] Heller, R.F., et al., How well can we predict coronary heart disease? Findings in the United Kingdom Heart Disease Prevention Project. BRITISH MEDICAL JOURNAL, 1984.
[14] Wilson, P.W.F., et al., Prediction of Coronary Heart Disease Using Risk Factor Categories. American Heart Association Journal, 1998.
[15] Simons, L.A., et al., Risk functions for prediction of cardiovascular disease in elderly Australians: the Dubbo Study. Medical Journal of Australia, 2003. 178.
[16] Salahuddin and F. Rabbi, Statistical Analysis of Risk Factors for Cardiovascular disease in Malakand Division. Pak. j. stat. oper. res., 2006. Vol.II: p. pp49-56.
[17] Shahwan-Akl, L., Cardiovascular Disease Risk Factors among Adult Australian-Lebanese in Melbourne. International Journal of Research in Nursing, 2010. 6 (1).
[18] Bramer, M., Principles of data mining. 2007: Springer.
[19] Varun Kumar, Nisha Rathee,"Knowledge Discovery from Database using an Integration of clustering and Classification", IJACSA, vol 2 No.3,PP. 29-33,March 2011.
[20] Ritu Chauhan, Harleen Kaur, M.Afshar Alam, "Data Clustering Method for Discovering Clusters in Spatial Cancer Databases", International Journal of Computer Applications (0975 – 8887) Volume 10– No.6, November 2010.
[21] Witten, I.H., Frank, E, "Data Mining: Practical Machine Learning Tools and Techniques", 2nd edn. Morgan Kaufmann, San Francisco (2005).
[22] G.Karraz,G.Magenes,"Automatic Classification of Heart beats using Neural Network Classifier based on a Bayesian Frame Work", IEEE, Vol 1,2006.
[23] N.A.Setiawan, A.F.M.Hani,"Missing Attribute Value Prediction Based on Artificial Neural Network and Rough Set Theory", IEEE, Vol 1, pp.306-310,2008
[24] Weka, "Data Mining Machine Learning Software, [Online]Available : http://www.cs.waikato.ac.nz/ml/.