Integrating Decision Tree and K-Means Clustering with Different Initial Centroid Selection Methods in the Diagnosis of Heart Disease Patients

Mai Shouman, Tim Turner, Rob Stocker
School of Engineering and Information Technology
University of New South Wales at the Australian Defence Force Academy
Northcott Drive, Canberra ACT 2600
mai.shouman@student.adfa.edu.au, t.turner@adfa.edu.au, r.stocker@adfa.edu.au

Abstract—*Heart disease is the leading cause of death in the world over the past 10 years. Researchers have been using several data mining techniques to help health care professionals in the diagnosis of heart disease patients. Decision Tree is one of the data mining techniques used in the diagnosis of heart disease showing considerable success. K-means clustering is one of the most popular clustering techniques; however initial centroid selection strongly affects its results. This paper investigates integrating k-means clustering with decision tree in the diagnosis of heart disease patients. It also investigates different methods of initial centroid selection of the k-means clustering such as inlier, outlier, range, random attribute values, and random row methods in the diagnosis of heart disease patients. The results show that integrating k-means clustering with decision tree with different initial centroid selection could enhance the decision tree accuracy in the diagnosing heart disease patients. It also showed that the inlier initial centroid selection method could achieve higher accuracy than other initial centroid selection methods in the diagnosis of heart disease patients.*

Keywords-Data Mining, K-Means Clustering, Initial Centroid Selection Methods, Decision Tree, Heart Disease Diagnosis.

1. INTRODUCTION

Heart disease is the leading cause of death in the world over the past 10 years. Moreover, the World Health Organization has reported that heart disease is the first leading cause of death in both high and low income countries [1]. The European Public Health Alliance reports that heart attacks and other circulatory diseases account for 41% of all deaths [2]. The Economical and Social Commission of Asia and the Pacific found that in one fifth of Asian countries, most lives are lost to non-communicable diseases such as cardiovascular, cancers, and diabetes diseases [3]. Statistics of South Africa report that heart and circulatory system diseases are the third leading cause of death in Africa [4]. The Australian Bureau of Statistics reported that heart and circulatory system diseases are the first leading cause of death in Australia, causing 33.7% of all deaths [5].

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 [6-7]. Data mining is an essential step in knowledge discovery. It is the exploration of large datasets to extract hidden and previously unknown patterns, relationships and knowledge that are difficult to detect with traditional statistical methods [8-12]. The application of data mining is rapidly spreading in a wide range of sectors such as analysis of organic compounds, financial forecasting, healthcare and weather forecasting [13].

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 [14]. Researchers are using data mining techniques in the medical diagnosis of several diseases such as diabetes [15], stroke [16], cancer [17], and heart disease [18]. Several data mining techniques are used in the diagnosis of heart disease such as naïve bayes, decision tree, neural network, kernel density, bagging algorithm, and support vector machine showing different levels of accuracies [18-24]

Decision Tree is one of the successful data mining techniques used in the diagnosis of heart disease patients [19, 22, 25]. Although researchers are investigating enhancing decision tree performance in classification problems, less research is done on enhancing decision tree performance in disease diagnosis. This research investigates enhancing decision tree performance in the diagnosis of heart disease patients through integrating clustering as a preprocessing step to decision tree classification.

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 [26]. Initial centroid selection is a critical issue in k-means clustering and strongly affects its results [27]. This paper investigates integrating k-means clustering using different initial centroid selection methods with decision tree in the diagnosis of heart disease patients. The rest of the paper is divided as follows: the background section investigates applying data mining techniques in the diagnosis of heart disease; the

methodology section explains k-means clustering, different initial centroid selection methods, and decision tree used in the diagnosis of heart disease patients; the heart disease data section explains the data used; the results section presents the results of integrating k-means clustering and decision tree; followed by the summary section.

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 [28], cholesterol [29], diabetes [30], hypertension, family history of heart disease [31], obesity, and lack of physical activity [32]. Knowledge of the risk factors associated with heart disease helps health care professionals to identify patients at high risk of having heart disease.

Researchers have been applying different data mining techniques over different heart disease datasets to help health care professionals in the diagnosis of heart disease [18-19, 22-25]. The results of the different data mining research cannot be compared because they have used different datasets. However, over time a benchmark data set has arisen in the literature: the Cleveland Heart Disease Dataset (CHDD).

Decision tree is one of the data mining techniques showing considerable success compared to other data mining techniques over different heart disease datasets [19, 21-22, 25]. Applying decision tree in diagnosing heart disease patients showed different accuracies on different datasets that ranged between 60.4% and 94.93%[22, 33]. Tu et al. applied decision tree classifier on the Cleveland heart disease dataset showing accuracy of 78.9% [25].

Recently researchers are investigating enhancing decision tree performance in classification problems. Anbarasi et al. investigated enhancing decision tree performance through integrating genetic algorithm as a feature subset selection method in the diagnosis of heart disease patients [34]. This paper investigates enhancing decision tree performance in the diagnosis of heart disease patients through the integration of k-means clustering.

This paper investigates if integrating k-means clustering with decision tree can enhance the classifier's performance in diagnosing heart disease patients. Importantly, the research involves a systematic investigation of which initial centroid selection method can provide better performance in diagnosing heart disease patients. It also investigates if applying different numbers of clusters can provide different performance in diagnosing heart disease patients and which number of clusters will provide the better performance.

3. METHODOLOGY

The methodology section discusses k-means clustering with five initial centroid selection methods. It also discusses the Decision Tree classifier used in the diagnosis of heart disease patients (Figure 1).

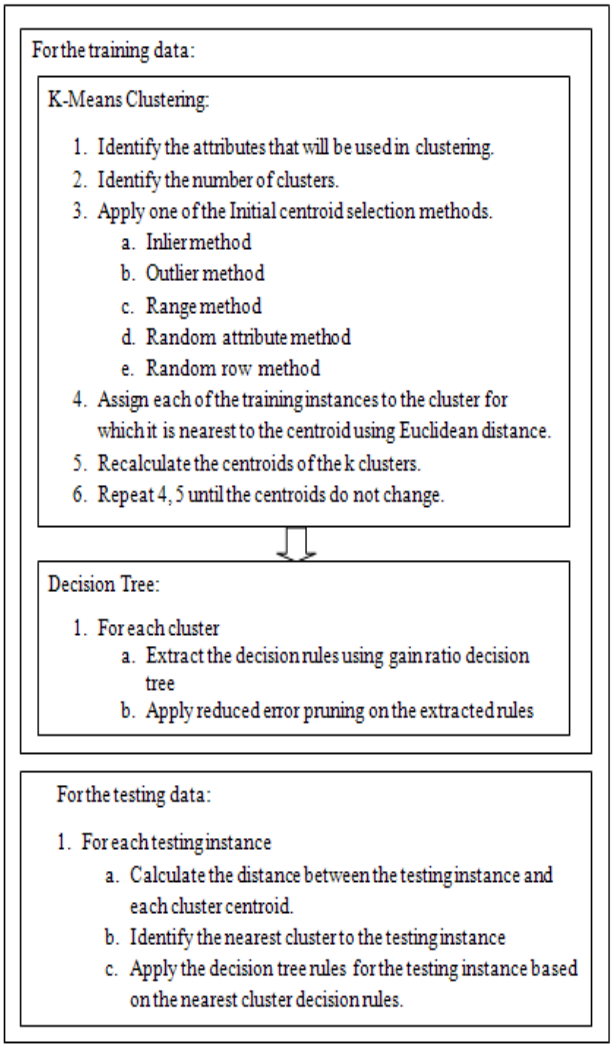

Figure 1: Integrating K-means Clustering and Decision Tree

3.1 Discretization

Decision Tree cannot deal with continuous attributes so they need to be converted into discrete ones, a process called discretization. Dougherty et al. carried out a comparative study between two unsupervised and two supervised discretization methods using 16 data sets showing that differences between the classification accuracies achieved by different discretization methods are not statistically significant [35]. Equal frequency discretization is a popular and successful unsupervised discretization method [36]. Previous related research has shown that this discretization method provides marginally better accuracy when applied on the CHHD [37]. So it is used as a preprocessing step to convert the continuous heart disease attributes to discrete ones.

3.2 K-Means Clustering

K-means clustering is one of the most popular and well know clustering techniques because of its simplicity and good behavior in many applications [26, 36]. The steps used in k-means clustering are shown in Figure 1.

Several researchers have identified that age, blood pressure and cholesterol are critical risk factors associated with heart disease [28, 31-32] . In identifying the attributes that will be used in the clustering, these attributes are obvious clustering attributes for heart disease patients. The number of clusters used in the k-means in this investigation ranged between two and five clusters. The difference between the initial centroid methods is discussed in the following section.

3.3 Initial Centroid Selection

Initial centroid selection is an important matter in k-means clustering and strongly affects its results [27]. This section discusses the generation of initial centroids based on actual sample data points using inlier method, outlier method, range method, random attribute method, and random row method [38].

3.3.1 Inlier Method

In generating the initial K centroids using the inlier method the following equations are used:

$$Ci = Min(X) - i \quad \text{where } 0 \le i \le k \qquad (1)$$

$$Cj = Min(Y) - j \quad \text{where } 0 \le j \le k \qquad (2)$$

Where the initial centroid is C (ci, cj) and min (X) and min (Y) are the minimum value of attribute X, and attribute Y respectively. K represents the number of clusters.

3.3.2 Outlier Method

In generating the initial K centroids using the outlier method the following equations are used:

$$Ci = Max(X) - i \quad \text{where } 0 \le i \le k \qquad (3)$$

$$Cj = Max(Y) - j \quad \text{where } 0 \le j \le k \qquad (4)$$

Where the initial centroid is C (ci, cj) and max (X) and max (Y) are the maximum value of attribute X, and attribute Y respectively. K represents the number of clusters.

3.3.3 Range Method

In generating the initial K centroids using the range method the following equations are used:

$$Ci = ((Max(X) - Min(X)) / K) * n \quad \text{where } 0 \le i \le k \qquad (5)$$

$$Cj = ((Max(Y) - Min(Y)) / K) * n \quad \text{where } 0 \le j \le \qquad (6)$$

The initial centroid is C (ci, cj). Where max (X) and min (X) are maximum and minimum values of attribute X, max (Y) and min (Y) are maximum and minimum values of attribute Y respectively. K represents the number of clusters.

3.3.4 Random Attribute Method

In generating the initial K centroids using the random attribute method the following equations are used:

$$Ci = random(X) \quad \text{where } 1 \le i \le k \qquad (7)$$

$$Cj = random(Y) \quad \text{where } 1 \le j \le k \qquad (8)$$

The initial centroid is C (ci, cj). The values of 'i', and 'j' vary from 1 to 'k'.

3.3.5 Random Row Method

In generating the initial K centroids using the random row method the following equations are used:

$$I = random(V) \quad \text{where } 1 \le V \le N \qquad (9)$$

$$Ci = X(I) \qquad (10)$$

$$Cj = Y(I) \qquad (11)$$

The initial centroid is C (ci, cj). N is the number of instances in the training dataset. X (I) and Y(I) are the values of the attributes X and Y respectively for the instance I.

3.4 Decision Tree

The decision tree type used in this research is the gain ratio decision tree. The gain ratio decision tree is based on the entropy (information gain) approach, which selects the splitting attribute that minimizes the value of entropy, thus maximizing the information gain [36]. To identify the splitting attribute of the decision tree, one must calculate the information gain for each attribute and then select the attribute that maximizes the information gain. The information gain for each attribute is calculated using the following formula [8, 36]:

$$E = \sum_{i=1}^{k} Pi \log_2 Pi \qquad (12)$$

Where k is the number of classes of the target attribute

Pi is the number of occurrences of class i divided by the total number of instances (i.e. the probability of i occurring).

To reduce the effect of bias resulting from the use of information gain, a variant known as gain ratio was introduced by the Australian academic Ross Quinlan [36]. The information gain measure is biased toward tests with many outcomes. That is, it prefers to select attributes having a large number of values [8]. Gain Ratio adjusts the information gain for each attribute to allow for the breadth and uniformity of the attribute values.

Gain Ratio = Information Gain / Split Information (13)

Where the split information is a value based on the column sums of the frequency table [36].

After extracting the decision tree rules, reduced error pruning was used to prune the extracted decision rules. Reduced error pruning is one of the fastest pruning methods and known to produce both accurate and small decision rules [39]. Applying reduced error pruning provides more compact decision rules and reduces the number of extracted rules.

3.5 10 Fold Cross Validation

To measure the stability of the proposed model, the data is divided into training and testing data with 10-fold cross validation. To evaluate the performance of the proposed model the sensitivity, specificity, and accuracy are calculated. The sensitivity is the proportion of positive instances that are correctly classified as positive (e.g. the proportion of sick people that are classified as sick). The specificity is the proportion of negative instances that are correctly classified as negative (e.g. the proportion of healthy people that are classified as healthy). The accuracy is the proportion of instances that are correctly classified [36].

Sensitivity = True Positive / Positive (14)

Specificity = True Negative / Negative (15)

Accuracy = (True Positive + True Negative) / (Positive + Negative) (16)

4. HEART DISEASE DATA

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. The attributes used in this study are shown in Table 1.

Table 1: Selected Cleveland Heart Disease Data Set Attributes

Name	Type	Description
Age	Continuous	Age in years
Sex	Discrete	1 = male 0 = female
Cp	Discrete	Chest pain type: 1 = typical angina 2 = atypical angina 3 = non-anginal pain 4 =asymptomatic
Trestbps	Continuous	Resting blood pressure (in mm Hg)
Chol	Continuous	Serum cholesterol in mg/dl
Fbs	Discrete	Fasting blood sugar > 120 mg/dl: 1 = true 0 = false
Restecg	Discrete	Resting electrocardiographic results: 0 = normal 1 = having ST-T wave abnormality 2 =showing probable or define left ventricular hypertrophy
Thalach	Continuous	Maximum heart rate achieved
Exang	Discrete	Exercise induced angina: 1 = yes 0 = no
Old peak ST	Continuous	Depression induced by exercise relative to rest
Slope	Discrete	The slope of the peak exercise segment : 1 = up sloping 2 = flat 3= down sloping
Ca	Discrete	Number of major vessels colored by fluoroscopy that ranged between 0 and 3.
Thal	Discrete	3 = normal 6= fixed defect 7= reversible defect
Diagnosis	Discrete	Diagnosis classes: 0 = healthy 1= patient who is subject to possible heart disease

5. RESULTS

A range of single and combined number of clustering attributes is applied in the experiment involving age, blood pressure and cholesterol attributes. However, best results are found using single attribute which is the age attribute. So K-means clustering is applied using the age attribute then the decision tree is applied on the thirteen attributes. The results of sensitivity, specificity, and accuracy in the diagnosis of heart disease using k-means clustering and decision tree with different initial centroids selection methods and different numbers of clusters are shown in Table 2. For the random attribute and random row methods, ten runs are executed and the average and best for each method are calculated and shown in Table 2. Tables 2 show that the best accuracy achieved is 83.9% by the inlier method with two clusters. The range method with different numbers of clusters did not show any enhancement in the decision tree accuracy in the diagnosis of heart disease patients.

Increasing the number of clusters with the inlier method did not show any enhancement in the accuracy as shown in Figure 2. Increasing the number of clusters with the outlier method could enhance its accuracy and showed the best accuracy with three clusters as shown in Figure 3. Increasing the number of clusters with the range method could enhance its accuracy and showed the best accuracy with four clusters as shown in Figure 4. However these accuracies are still less than that achieved by the inlier method with two clusters. Increasing the number of clusters for the random attribute and the random row could achieve slight enhancement in the

accuracy but it is still less than that achieved by them with two clusters as shown in Figure 5, and 6 respectively.

Table 2: Integrating different initial centroid selection for k-means clustering with Decision tree in diagnosing heart disease patients

No of Clusters	Initial Centroid Selection Method		Sensitivity	Specificity	Accuracy
No of clusters = 2	Inlier Method		81.6	83	83.9
	Outlier Method		71.6	76.2	76
	Range Method		71.6	76.2	76
	Random Attribute	Avg	75.85	78.85	79.29
		Best	77.7	83.3	82.2
	Random Row	Avg	76.94	79.51	80.14
		Best	81.6	83	83.9
No of clusters = 3	Inlier Method		76.6	80.2	80.9
	Outlier Method		78.1	79.6	81.2
	Range Method		69.8	77.8	76.3
	Random Attribute	Avg	73.17	78.07	78.33
		Best	76	79.9	80.3
	Random Row	Avg	72.71	78.06	77.95
		Best	76.2	78.9	79.8
No of clusters = 4	Inlier Method		72.8	80	78.5
	Outlier Method		74.1	80.3	79.9
	Range Method		72.8	80	78.5
	Random Attribute	Avg	72.31	79.05	78.01
		Best	74.2	80.7	80.5
	Random Row	Avg	72.07	79.14	78.05
		Best	72.3	81.1	79.9
No of clusters = 5	Inlier Method		78.2	77.2	75.9
	Outlier Method		68.1	72	73.6
	Range Method		72.8	77.2	75.9
	Random Attribute	Avg	71.59	76.2	75.56
		Best	73.5	77.3	78.1
	Random Row	Avg	72.3	75.86	75.6
		Best	76.7	75	78

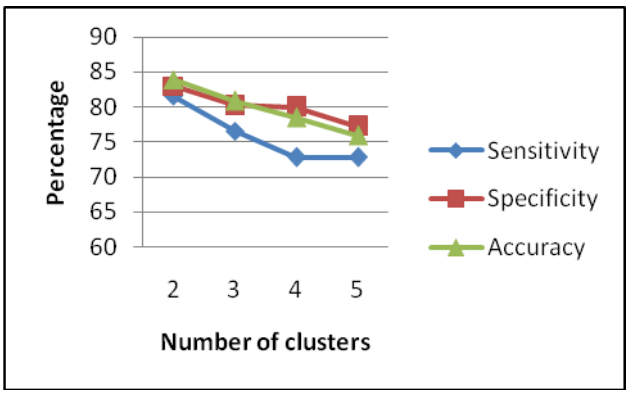

Figure 2: Different Number of Clusters Performance for Inlier Method

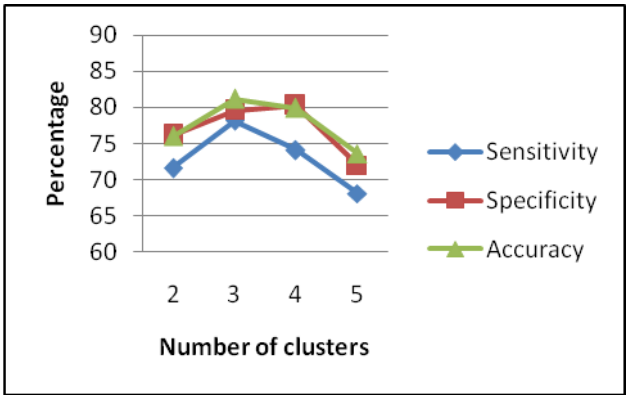

Figure 3: Different Number of Clusters Performance for Outlier Method

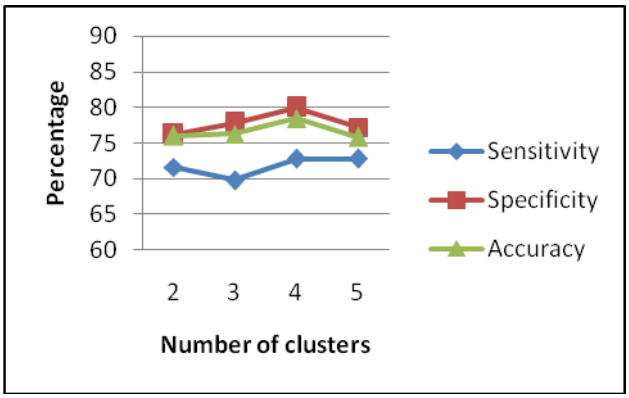

Figure 4: Different Number of Clusters Performance for Range Method

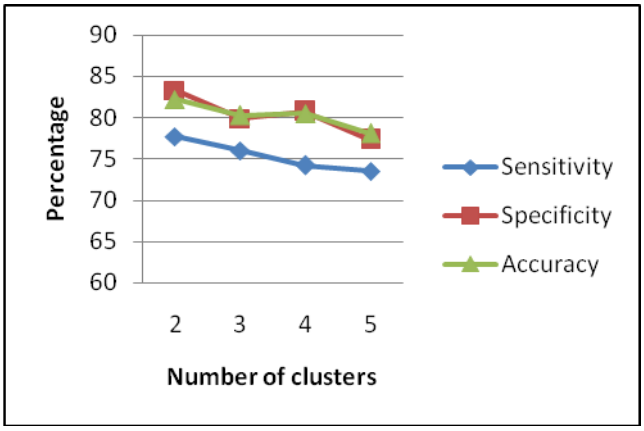

Figure 5: Different Number of Clusters Performance for Random Attribute Method

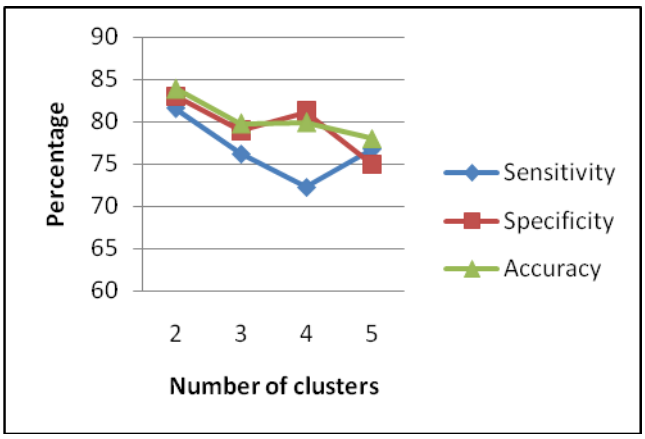

Figure 6: Different Number of Clusters Performance for Random Row Method

Why do two clusters show better performance than other numbers of clusters in the diagnosis of heart disease patients? The number of instances is relatively small in the CHHD. A larger dataset is needed to identify if two clusters will still provide the best results. Also, the target attribute of the Cleveland heart disease dataset has two values. Further investigation is also needed to identify if there is a relationship between the number of clusters showing best results and the number of values of the target attribute.

When comparing integrating k-means clustering and decision tree with traditional decision tree applied previously on the same dataset, integrating k-means clustering with decision tree could enhance the accuracy of decision tree in diagnosing heart disease patients as shown in Table 3. In Addition, integrating k-means clustering and decision tree could achieve higher accuracy than the bagging algorithm in the diagnosis of heart disease patients as shown in Table 3.

Table 3: Comparing integrating k-means clustering and decision tree with traditional decision tree and other data mining techniques

Author/ Year	Technique	Accuracy
Tu, et al., 2009	Decision tree	78.91%
	Bagging Algorithm	81.41%
Our work	Two clusters Inlier initial centroid selection k-means clustering decision tree	83.9%

6. SUMMARY

Heart disease is the leading cause of death all over the world. Researchers have been investigating applying different data mining techniques to help health care professionals in the diagnosis of heart disease patients. Decision Tree is one of the successful data mining techniques used in the diagnosis of heart disease patients. This paper investigated integrating k-means clustering with decision tree in the diagnosis of heart disease patients. Initial centroid selection is a critical issue that strongly affects k-means clustering results. Our research systematically investigated applying different methods of initial centroid selection such as range, inlier, outlier, random attribute values, and random row methods for the k-means clustering technique in the diagnosis of heart disease patients. The results show that integrating k-means clustering and decision tree can enhance decision tree accuracy in the diagnosis of heart disease patients. The results also show that the best accuracy achieved is 83.9% by the inlier method with two clusters. Finally, some limitations on this work are noted as pointers for future research.

7. REFERENCES

[1] World Health Organization. 2007 7-Febuary 2011]; Available from: http://www.who.int/mediacentre/factsheets/fs310.pdf.

[2] European Public Health Alliance. 2010 7-February-2011]; Available from: http://www.epha.org/a/2352

[3] ESCAP. 2010 7-February-2011]; Available from: http://www.unescap.org/stat/data/syb2009/9.Health-risks-causes-of-death.asp.

[4] Statistics South Africa. 2008 7-February-2011]; Available from: http://www.statssa.gov.za/publications/P03093/P030932006.pdf

[5] Australian Bureau of Statistics. 2010 7-February-2011]; Available from: http://www.ausstats.abs.gov.au/Ausstats/subscriber.nsf/0/E8510D1C8DC1AE1CCA2576F600139288/$File/33030_2008.pdf

[6] Helma, C., E. Gottmann, and S. Kramer, Knowledge discovery and data mining in toxicology. Statistical Methods in Medical Research, 2000.

[7] Podgorelec, V., et al., Decision Trees: An Overview and Their Use in Medicine. Journal of Medical Systems, 2002. Vol. 26.

[8] Han, j. and M. Kamber, Data Mining Concepts and Techniques. 2006: Morgan Kaufmann Publishers.

[9] Lee, I.-N., S.-C. Liao, and M. Embrechts, Data mining techniques applied to medical information. Med. inform, 2000.

[10] Obenshain, M.K., Application of Data Mining Techniques to Healthcare Data. Infection Control and Hospital Epidemiology, 2004.

[11] Sandhya, J., et al., Classification of Neurodegenerative Disorders Based on Major Risk Factors Employing Machine Learning Techniques. International Journal of Engineering and Technology, 2010. Vol.2, No.4.

[12] Thuraisingham, B., A Primer for Understanding and Applying Data Mining. IT Professional IEEE, 2000.

[13] Ashby, D. and A. Smith, The Best Medicine? Plus Magazine - Living Mathematics., 2005.

[14] 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, .

[15] Porter, T. and B. Green, Identifying Diabetic Patients: A Data Mining Approach. Americas Conference on Information Systems, 2009.
[16] Panzarasa, S., et al., Data mining techniques for analyzing stroke care processes. Proceedings of the 13th World Congress on Medical Informatics, 2010.
[17] 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.
[18] 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.
[19] Andreeva, P., Data Modelling and Specific Rule Generation via Data Mining Techniques. International Conference on Computer Systems and Technologies - CompSysTech, 2006.
[20] 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.
[21] 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).
[22] Sitar-Taut, V.A., et al., Using machine learning algorithms in cardiovascular disease risk evaluation. Journal of Applied Computer Science & Mathematics, 2009.
[23] 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.
[24] 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.
[25] Tu, M.C., D. Shin, and D. Shin, Effective Diagnosis of Heart Disease through Bagging Approach. Biomedical Engineering and Informatics, IEEE, 2009.
[26] Wu, X., et al., Top 10 algorithms in data mining analysis. Knowl. Inf. Syst., 2007.
[27] Tajunisha, N. and V. Saravanan, A new approach to improve the clustering accuracy using informative genes for unsupervised microarray data sets. International Journal of Advanced Science and Technology, 2011.
[28] 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.
[29] Wilson, P.W.F., et al., Prediction of Coronary Heart Disease Using Risk Factor Categories. American Heart Association Journal, 1998.
[30] Simons, L.A., et al., Risk functions for prediction of cardiovascular disease in elderly Australians: the Dubbo Study. Medical Journal of Australia, 2003. 178.
[31] 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.
[32] Shahwan-Akl, L., Cardiovascular Disease Risk Factors among Adult Australian-Lebanese in Melbourne. International Journal of Research in Nursing, 2010. 6 (1).
[33] Palaniappan, S. and R. Awang, Web-Based Heart Disease Decision Support System using Data Mining Classification Modeling Techniques. Proceedings of iiWAS, 2007.
[34] Anbarasi, M., E. Anupriya, and N.C.S.N. Iyengar, Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm. International Journal of Engineering Science and Technology, 2010. Vol. 2(10).
[35] Dougherty, J., R. Kohavi, and M. Sahami, Supervised and unsupervised discretization of continuous features. In: Proceedings of the 12th international conference on machine learning. San Francisco: Morgan Kaufmann, 1995: p. p. 194–202.
[36] Bramer, M., Principles of data mining. 2007: Springer.
[37] Shouman, M., T. Turner, and R. Stocker, Using decision tree for diagnosing heart disease patients. 9th Australasian Data Mining Conference 2011. 121.
[38] Khan, D.M. and N. Mohamudally, A Multiagent System (MAS) for the Generation of Initial Centroids for kmeans Clustering Data Mining Algorithm based on Actual Sample Datapoints. Journal of Next Generation Information Technology, August, 2010. Volume 1, Number 2.
[39] Esposito, F., D. Malerba, and G. Semeraro A Comparative Analysis of Methods for Pruning Decision Trees. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, 1997. VOL. 19, NO. 5.