Using Decision Tree for Diagnosing 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. Decision Tree is one of the successful data mining techniques used. However, most research has applied J4.8 Decision Tree, based on Gain Ratio and binary discretization. Gini Index and Information Gain are two other successful types of Decision Trees that are less used in the diagnosis of heart disease. Also other discretization techniques, voting method, and reduced error pruning are known to produce more accurate Decision Trees. This research investigates applying a range of techniques to different types of Decision Trees seeking better performance in heart disease diagnosis. A widely used benchmark data set is used in this research. To evaluate the performance of the alternative Decision Trees the sensitivity, specificity, and accuracy are calculated. The research proposes a model that outperforms J4.8 Decision Tree and Bagging algorithm in the diagnosis of heart disease patients.

Keywords: Data Mining, Decision Tree, Discretization, Heart Disease.

1 Introduction

Heart disease is the leading cause of death in the world over the past 10 years (World Health Organization 2007). The European Public Health Alliance reported that heart attacks, strokes and other circulatory diseases account for 41% of all deaths (European Public Health Alliance 2010). The Economical and Social Commission of Asia and the Pacific reported that in one fifth of Asian countries, most lives are lost to non-communicable diseases such as cardiovascular diseases, cancers, diabetes and chronic respiratory diseases (ESCAP 2010). The Australian Bureau of Statistics reported that heart and circulatory system diseases are the first leading cause of death in Australia, causing 33.7% all deaths (Australian Bureau of Statistics 2010).

Motivated by the world-wide increasing mortality of heart disease patients each year and the availability of huge amount of patients' data from which to extract useful knowledge, researchers have been using data mining techniques to help health care professionals in the diagnosis of heart disease (Helma, Gottmann et al. 2000; Podgorelec, Kokol et al. 2002). Data mining 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 (Lee, Liao et al. 2000; Thuraisingham 2000; Obenshain 2004; Han and Kamber 2006; Sandhya, P. Deepa Shenoy et al. 2010).

Data mining in healthcare is an emerging field of high importance for providing prognosis and a deeper understanding of medical data. Data mining applications in healthcare include analysis of health care centres for better health policy-making and prevention of hospital errors, early detection, prevention of diseases and preventable hospital deaths, more value for money and cost savings, and detection of fraudulent insurance claims (Ruben 2009). Researchers are using data mining techniques in the medical diagnosis of several diseases such as diabetes (Porter and Green 2009), stroke (Panzarasa, Quaglini et al. 2010), cancer (Li L 2004), and heart disease (Das, Turkoglu et al. 2009). Several data mining techniques are used in the diagnosis of heart disease such as Naïve Bayes, Decision Tree, neural network, kernel density, automatically defined groups, bagging algorithm, and support vector machine showing different levels of accuracies (Yan, Zheng et al. 2003; Andreeva 2006; Das, Turkoglu et al. 2009; Sitar-Taut, Zdrenghea et al. 2009; Rajkumar and Reena 2010; Srinivas, Rani et al. 2010)

This paper presents a new model that enhances the Decision Tree accuracy in identifying heart disease patients. The model integrates a multiple classifiers voting technique with different types of discretization methods and different types of Decision Trees. 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 the proposed methodology for enhancing the Decision Tree accuracy in diagnosing heart disease, and the results section is followed by a summary section.

2 Background

Knowledge of the risk factors associated with heart disease helps health care professionals to identify patients at high risk of having heart disease. Statistical analysis has identified the risk factors associated with heart disease to be age, blood pressure, smoking habit (Heller, Chinn et al. 1984), total cholesterol (Wilson, D'Agostino et al. 1998), diabetes (Simons, Simons et al. 2003),

hypertension, family history of heart disease (Salahuddin and Rabbi 2006), obesity, and lack of physical activity (Shahwan-Akl 2010).

Researchers have been applying different data mining techniques to help health care professionals with improved accuracy in the diagnosis of heart disease. Neural network, Naïve Bayes, Genetic algorithm, Decision Tree, classification via clustering, and direct kernel self-organizing map are some techniques used in the diagnosis of heart disease (De Beule, Maesa et al. 2007; Tantimongcolwat, Naenna et al. 2008; Das, Turkoglu et al. 2009; Anbarasi, Anupriya et al. 2010; Kavitha, Ramakrishnan et al. 2010; Srinivas, Rani et al. 2010).

In particular, researchers have been investigating the application of the Decision Tree technique in the diagnosis of heart disease with considerable success. Andreeva used C4.5 Decision Tree in the diagnosis of heart disease showing accuracy of 75.73% (Andreeva 2006). Sitair-Taut et al. used the weka tool to investigate applying Naïve Bayes and J4.8 Decision Trees for the detection of coronary heart disease. The results showed that there is no significant difference between Naïve Bayes and Decision Trees in the ability to realize a correct prediction of coronary heart disease (Sitar-Taut, Zdrenghea et al. 2009). Tu et al. used the bagging algorithm in the weka tool and compared it with J4.8 Decision Tree in the diagnosis of heart disease. The bagging algorithm showed the better accuracy of 81.41% while the Decision Tree showed an accuracy of 78.91% (Tu, Shin et al. 2009).

Applying Decision Tree techniques has shown useful accuracy in the diagnosis of heart disease. But assisting health care professionals in the diagnosis of the world's biggest killer demands higher accuracy. Our research seeks to improve diagnosis accuracy to improve health outcomes. Most Decision Tree types used such as J4.8 and C4.5 Decision Trees are based on Gain Ratio in the extraction of Decision Tree rules. However there are other Decision Tree types such as Information Gain and Gini Index that have been less used in the diagnosis of heart disease.

Decision Tree is one of the data mining techniques that cannot handle continuous variables directly so the continuous attributes must be converted to discrete attributes, a process called discretization (Kotsiantis and Kanellopoulos 2006). J4.8 and C4.5 Decision Trees use binary discretization for continuous-valued features. However, multi-interval discretization methods are known to produce more accurate Decision Trees than binary discretization (Fayyad and Keki 1992; Perner and Trautzsch 1998), yet are less used in heart disease diagnosis research. Other important accuracy improving is applying multiple classifier voting and reduced error pruning to Decision Tree in the diagnosis of heart disease patients. Intuitively, more complex models might be expected to produce more accurate results, but which combination of techniques is best?

Seeking to thoroughly investigate options for accuracy improvements in heart disease diagnosis this paper systematically investigates applying multiple classifiers voting technique with different multi-interval discretization methods such as equal width, equal frequency, chi merge and entropy with different types of Decision Tree such as Information Gain, Gini Index, and Gain Ratio.

3 Methodology

There are two main issues that affect the performance of Decision Trees; the data discretization method used and the type of Decision Tree used. Reduced error pruning is shown to further improve decision tree performance. The proposed methodology involves systematically testing different discretization techniques, multiple classifiers voting technique and different Decision Trees type in the diagnosis of heart disease patients. Different combinations of discretization methods, decision tree types and voting are tested to identify which combination will provide the best performance in diagnosing heart disease patients. A test harness was implemented using Microsoft Visual Studio 2008.

3.1 Data Discretization

Discretization methods are categorised as supervised or unsupervised (Dougherty, Kohavi et al. 1995). The unsupervised discretization methods do not make use of class membership information during the discretization process. The supervised discretization methods use the class labels for carrying out discretization process such as chi-square based methods and entropy based methods (Kotsiantis and Kanellopoulos 2006). All the discretization methods are used as a pre-processing step to convert the continuous attributes in the data set to discrete attributes. The number of intervals used by the discretization techniques is five. Each method was used to pre-process the benchmark data set for trials of each decision tree type (hence left-most column of Tables 2 and 3).

3.1.1 Unsupervised Discretization

In unsupervised discretization, equal-width interval and equal-frequency methods are used. The equal-width discretization algorithm determines the minimum and maximum values of the discretized attribute and then divides the range into the user-defined number of equal-width discrete intervals. The equal-frequency algorithm determines the minimum and maximum values of the discretized attribute, sorts all values in ascending order, and divides the range into a user-defined number of intervals so that every interval contains the same number of sorted values (Dougherty, Kohavi et al. 1995).

3.1.2 Supervised Discretization

In supervised discretization chi merge and entropy are two of the most well-known discretization methods (Bramer 2007). The chi merge discretization uses $\chi 2$ statistic to determine the independence of the class from the two adjacent intervals, combining them if they are dependent, and allowing them to be separate otherwise (Kerber 1992). This algorithm merges the pair of intervals with the lowest value of $\chi 2$ as long as the number of intervals is more than predefined maximum number of intervals.

The entropy discretization used in this model is that developed by Fayyad and Keki (Fayyad and Keki 1992).

Entropy is an information-theoretic measure of the 'uncertainty' contained in a training set (Han and Kamber 2006). It evaluates candidate cut points through an entropy-based method to select boundaries for discretization. Instances are sorted into ascending numerical order and then the entropy for each candidate cut point is calculated. Cut points are recursively selected to minimize entropy until a stopping criterion is achieved. In this model the stop criterion is achieving five intervals of the attribute.

3.2 Voting

Multiple classifier voting involves dividing the training data into smaller equal subsets of data and building a Decision Tree classifier for each subset of data. Voting is based on plurality or majority voting; each individual classifier contributes a single vote (Hall, Bowyer et al. 2000). Applying voting to classification algorithms is showing successful improvement in the accuracy of these classifiers (Paris, Affendey et al. 2010). The research here tested voting subsets where the data was divided between three and eleven subsets for each discretization method for each decision tree type. The most successful division (nine subsets) is reported here.

3.3 Decision Tree Type

There are many types of Decision Trees. The difference between them is the mathematical model that is used in selecting the splitting attribute in extracting the Decision Tree rules. The research tests the three most commonly-used types: Information Gain, Gini Index, and Gain Ratio Decision Trees, each described below.

3.3.1 Information Gain

The entropy (Information Gain) approach selects the splitting attribute that minimizes the value of entropy, thus maximising the Information Gain. 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 (Han and Kamber 2006; Bramer 2007):

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

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).

3.3.2 Gini Index

The Gini Index measures the impurity of data. The Gini Index is calculated for each attribute in the data set. If there are k classes of the target attribute, with the probability of the ith class being Pi, the Gini Index is defined as (Bramer 2007):

$$\text{Gini Index} = 1 - \sum_{i=1}^{k} Pi^2 \quad (2)$$

The splitting attribute is the attribute with the largest reduction in the value of the Gini Index.

3.3.3 Gain Ratio

To reduce the effect of the bias resulting from the use of Information Gain, a variant known as Gain Ratio was introduced by the Australian academic Ross Quinlan (Bramer 2007). The Information Gain measure is biased toward tests with many outcomes. That is, it prefers to select attributes having a large number of values (Han and Kamber 2006). 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 (3)

Where the split information is a value based on the column sums of the frequency table (Bramer 2007).

3.4 Pruning

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 (Esposito, Malerba et al. 1997). Applying reduced error pruning provides more compact decision rules and reduces the number of extracted rules.

3.5 Performance Evaluation

To evaluate the performance of each combination the sensitivity, specificity, and accuracy were calculated. The sensitivity is proportion of positive instances that are correctly classified as positive (i.e. the proportion of sick people that are classified as sick). The specificity is the proportion of negative instances that are correctly classified as negative (i.e. the proportion of healthy people that are classified as healthy). The accuracy is the proportion of instances that are correctly classified (Bramer 2007). To measure the stability of the performance of the proposed model the data is divided into training and testing data with 10-fold cross validation.

Sensitivity = True Positive/Positive (4)

Specificity = True Negative/ Negative (5)

Accuracy = (True Positive + True Negative) / (Positive + Negative) (6)

3.6 Summary

The research process involves data discretization, data partitioning, Decision Tree type selection, and the application of reduced error pruning to produce pruned Decision Trees. The data discretization is divided into supervised and unsupervised methods. The unsupervised methods involve equal width and equal frequency while the supervised discretization methods involve chi merge and entropy. The data partitioning involves testing with and without voting. Three Decision Tree types are tested: Information Gain, Gini Index, and Gain Ratio. Finally, reduced error pruning is applied on all the Decision Tree rules extracted from the training data. Figure 1 summarizes the components of the research process.

The actual testing involved executing each variant of each element in combination against the whole data set. Twelve Decision Tree variants were created by mixing discretization approaches with different Decision Tree

types. Each variant was then tested on its own and through different voting partitioning schemes (three, five, seven, nine and eleven partitions). The result of each variant through each voting partition had reduced error pruning applied. Overall, more than 70 Decision Trees were executed over the one data set to compile the findings presented here

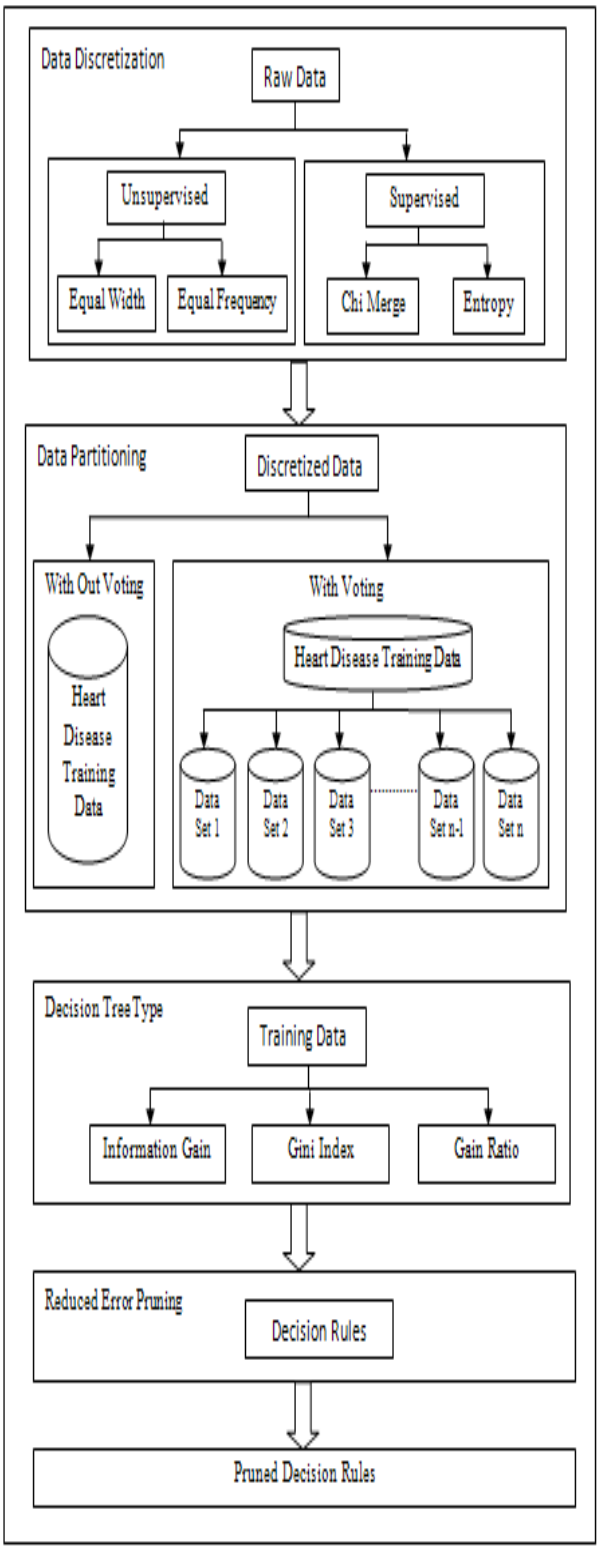

Figure 1: Research Process Used to Assess Alternative Decision Tree Techniques

4 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. Consequently, to allow comparison with the literature, we restricted testing to these same attributes (see Table 1). The data set contains 303 rows of which 297 are complete. Six rows contain missing values and they are removed from the experiment.

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 by Estes'criteria
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

The results of sensitivity, specificity, and accuracy in the diagnosis of heart disease using equal width, equal frequency, chi merge, and entropy discretization with Information Gain, Gini Index, and Gain Ratio Decision Trees and reduced error pruning are shown in Table 2 and Table 3. Table 2 shows the results without applying voting to the Decision Tree. The highest accuracy achieved is 79.1% by the equal width discretization Information Gain Decision Tree. Different partitions of voting were applied to the data. The nine subsets voting showed the best performance and is the only iteration reported here (Table 3). The highest accuracy achieved is 84.1% by the equal frequency discretization Gain Ratio Decision Tree. Table 4 shows the difference in accuracy when applying the nine subsets voting scheme. The highest increase in the accuracy is achieved by the equal frequency discretization Gain Ratio Decision Tree and is 6.4%.

Table 2: Without Voting Decision Tree Results

		Sensitivity	Specificity	Accuracy
Equal Width	Info Gain	78.1%	79.4%	79.1%
	Gini Index	76.4%	83.4%	78.8%
	Gain Ratio	66.1%	80.5%	75.5%
Equal Frequency	Info Gain	76%	75.7%	76.3%
	Gini Index	75.5%	77.7%	76.9%
	Gain Ratio	71.4%	79.9%	77.7%
ChiMerge	Info Gain	71.7%	80.5%	77.1%
	Gini Index	73%	81.1%	78.2%
	Gain Ratio	63.3%	81.4%	75.1%
Entropy	Info Gain	78.1%	79.7%	78.1%
	Gini Index	77.1%	80.7%	78.4%
	Gain Ratio	68%	82.4%	76.5%

The chimerge and entropy supervised discretization methods do not show any enhancement in the Decision Tree accuracy either with or without voting. Applying the voting is showing an increase in the accuracy of different types of Decision Tree. When comparing the best results with the J4.8 Decision Tree and bagging algorithm that used the same data set (Tu, Shin et al. 2009), this research achieved higher sensitivity, specificity, and accuracy than J4.8 Decision Tree and achieved higher sensitivity, and accuracy than bagging algorithm as shown in Table 5.

Table 3: Nine Voting Decision Tree Results

		Sensitivity	Specificity	Accuracy
Equal Width	Info Gain	73%	89.7%	82.6%
	Gini Index	69%	89.6%	81.5%
	Gain Ratio	70.3%	90.6%	81%
Equal Frequency	Info Gain	69.3%	86.3%	82%
	Gini Index	68.4%	88.1%	81.4%
	Gain Ratio	77.9%	85.2%	84.1%
ChiMerge	Info Gain	71.6%	83.2%	78.3%
	Gini Index	72.7%	82.7%	79.4%
	Gain Ratio	66.2%	85.5%	79.1%
Entropy	Info Gain	69%	90.3%	80.9%
	Gini Index	73.4%	92.8%	83.9%
	Gain Ratio	67.5%	89.8%	79.9%

Table 4: Increased Accuracy after Applying the Voting

		Increase Accuracy
Equal Width	Info Gain	3.5%
	Gini Index	2.7%
	Gain Ratio	5.5%
Equal Frequency	Info Gain	5.7%
	Gini Index	4.5%
	Gain Ratio	6.4%
ChiMerge	Info Gain	1.2%
	Gini Index	1.2%
	Gain Ratio	4%
Entropy	Info Gain	2.8%
	Gini Index	5.5%
	Gain Ratio	3.4%

Table 5: Comparing Proposed Model with Previous Results

		Sensitivity	Specificity	Accuracy
Proposed Model (Nine Voting Equal Frequency Discretization Gain Ratio Decision Tree)		77.9%	85.2%	84.1%
Tu et al., 2009	**J4.8 Decision Tree**	72.01%	84.48%	78.9%
	Bagging Algorithm	74.93%	86.64%	81.41%

From these results it is concluded that although most researchers are using binary discretization with Gain Ratio Decision Tree in the diagnosis of heart disease, applying multi-interval equal frequency discretization with nine voting Gain Ratio Decision Tree provides better results in the diagnosis of heart disease patients. . We surmise that the improvement in accuracy arises from the increased granularity in splitting attributes offered by multi-interval discretization. Combined with Gain Ratio calculations, this likely increases the accuracy of the probability calculation for any given attribute value. Having that higher probability validated by the voting across multiple similar trees further enhances the selection of useful splitting attribute values. These results would benefit from further testing on much larger data sets.

6 Summary

Decision Tree is one of the successful data mining techniques used in the diagnosis of heart disease. Yet its accuracy is not perfect. Most research applies the J4.8 Decision Tree that is based on Gain Ratio and binary discretization. This research systematically tested combinations of discretization, decision tree type and voting to identify a more robust, more accurate method. The supervised discretization methods do not show any enhancement in the Decision Tree accuracy either with or without voting. Applying voting shows increase in the accuracy of different types of Decision Tree. Systematic testing against a widely-used benchmark data set shows that nine voting with equal frequency discretization and Gain Ratio Decision Tree can enhance the accuracy of the diagnosis of heart disease.

7 References

Anbarasi, M., E. Anupriya, et al. (2010). "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm." International Journal of Engineering Science and Technology Vol. 2(10).

Andreeva, P. (2006). "Data Modelling and Specific Rule Generation via Data Mining Techniques." International Conference on Computer Systems and Technologies - CompSysTech.

Australian Bureau of Statistics. (2010). Retrieved 7-February-2011, from http://www.ausstats.abs.gov.au/Ausstats/subscriber.nsf/0/E8510D1C8DC1AE1CCA2576F600139288/$File/33030_2008.pdf

Bramer, M. (2007). Principles of data mining, Springer.

Das, R., I. Turkoglu, et al. (2009). "Effective diagnosis of heart disease through neural networks ensembles." Expert Systems with Applications, Elsevier 36 (2009): 7675–7680.

De Beule, M., E. Maesa, et al. (2007). "Artificial neural networks and risk stratification: A promising combination." Mathematical and Computer Modelling, Elsevier.

Dougherty, J., R. Kohavi, et al. (1995). "Supervised and unsupervised discretization of continuous features." In: Proceedings of the 12th international conference on machine learning. San Francisco: Morgan Kaufmann: p. 194–202.

ESCAP. (2010). Retrieved 7-February-2011, from http://www.unescap.org/stat/data/syb2009/9.Health-risks-causes-of-death.asp.

Esposito, F., D. Malerba, et al. (1997). "A Comparative Analysis of Methods for Pruning Decision Trees." IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE VOL. 19, NO. 5.

European Public Health Alliance. (2010). Retrieved 7-February-2011, from http://www.epha.org/a/2352

Fayyad, U. M. and B. I. Keki (1992). "On the handing of Continuous-Valued Attributes in Decision Tree Generation." Machine Learning 8 (87-102).

Hall, L. O., K. W. Bowyer, et al. (2000). "Distributed Learning on Very Large Data Sets." In Workshop on Distributed and Parallel Knowledge Discover.

Han, j. and M. Kamber (2006). Data Mining Concepts and Techniques, Morgan Kaufmann Publishers.

Heller, R. F., S. Chinn, et al. (1984). "How well can we predict coronary heart disease? Findings in the United Kingdom Heart Disease Prevention Project." BRITISH MEDICAL JOURNAL.

Helma, C., E. Gottmann, et al. (2000). "Knowledge discovery and data mining in toxicology." Statistical Methods in Medical Research.

Kavitha, K. S., K. V. Ramakrishnan, et al. (2010). "Modeling and design of evolutionary neural network for heart disease detection." International Journal of Computer Science Issues (IJCSI) Vol. 7, Issue 5.

Kerber, R. (1992). "ChiMerge: Discretization of Numeric Attributes." In Proceedings of the Tenth National Conference on Articial Intelligence

Kotsiantis, S. and D. Kanellopoulos (2006). "Discretization Techniques: A recent survey." International Transactions on Computer Science and Engineering Vol.32 (1) pp. 47-58.

Lee, I.-N., S.-C. Liao, et al. (2000). "Data mining techniques applied to medical information." Med. inform.

Li L, T. H., Wu Z, Gong J, Gruidl M, Zou J, Tockman M, Clark RA (2004). "Data mining techniques for cancer

detection using serum proteomic profiling." Artificial Intelligence in Medicine, Elsevier.

Obenshain, M. K. (2004). "Application of Data Mining Techniques to Healthcare Data." Infection Control and Hospital Epidemiology.

Panzarasa, S., S. Quaglini, et al. (2010). "Data mining techniques for analyzing stroke care processes." Proceedings of the 13th World Congress on Medical Informatics.

Paris, I. H. M., L. S. Affendey, et al. (2010). "Improving Academic Performance Prediction using Voting Technique in Data Mining." World Academy of Science, Engineering and Technology 62.

Perner, P. and S. Trautzsch (1998). "Multi-Interval Discretization Methods for Decision Tree Learning." Advances in Pattern Recognition, A. Amin, D. Dori, P. Pudil, and H. Freeman (Eds.),LNCS 1451, Springer Verlag S. 475-482.

Podgorelec, V., P. Kokol, et al. (2002). "Decision Trees: An Overview and Their Use in Medicine." Journal of Medical Systems Vol. 26.

Porter, T. and B. Green (2009). "Identifying Diabetic Patients: A Data Mining Approach." Americas Conference on Information Systems.

Rajkumar, A. and G. S. Reena (2010). "Diagnosis Of Heart Disease Using Datamining Algorithm." Global Journal of Computer Science and Technology Vol. 10 (Issue 10).

Ruben, D. C. J. (2009). "Data Mining in Healthcare: Current Applications and Issues."

Salahuddin and F. Rabbi (2006). "Statistical Analysis of Risk Factors for Cardiovascular disease in Malakand Division." Pak. j. stat. oper. res. Vol.II: pp49-56.

Sandhya, J., P. Deepa Shenoy, et al. (2010). "Classification of Neurodegenerative Disorders Based on Major Risk Factors Employing Machine Learning Techniques." International Journal of Engineering and Technology Vol.2, No.4.

Shahwan-Akl, L. (2010). "Cardiovascular Disease Risk Factors among Adult Australian-Lebanese in Melbourne." International Journal of Research in Nursing 6 (1).

Simons, L. A., J. Simons, et al. (2003). "Risk functions for prediction of cardiovascular disease in elderly Australians: the Dubbo Study." Medical Journal of Australia 178.

Sitar-Taut, V. A., D. Zdrenghea, et al. (2009). "Using machine learning algorithms in cardiovascular disease risk evaluation." Journal of Applied Computer Science & Mathematics.

Srinivas, K., B. K. Rani, et al. (2010). "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks." International Journal on Computer Science and Engineering (IJCSE) Vol. 02, No. 02: 250-255.

Tantimongcolwat, T., T. Naenna, et al. (2008). "Identification of ischemic heart disease via machine learning analysis on magnetocardiograms." Computers in Biology and Medicine, Elsevier 38 (2008): 817 – 825.

Thuraisingham, B. (2000). "A Primer for Understanding and Applying Data Mining." IT Professional IEEE.

Tu, M. C., D. Shin, et al. (2009). "Effective Diagnosis of Heart Disease through Bagging Approach." Biomedical Engineering and Informatics, IEEE.

Wilson, P. W. F., R. B. D'Agostino, et al. (1998). "Prediction of Coronary Heart Disease Using Risk Factor Categories." American Heart Association Journal.

World Health Organization. (2007). Retrieved 7-Febuary 2011, from http://www.who.int/mediacentre/factsheets/fs310.pdf.

Yan, H., J. Zheng, et al. (2003). "Development of a decision support system for heart disease diagnosis using multilayer perceptron." Proceedings of the 2003 International Symposium on vol.5: pp. V-709- V-712.