Intelligent Heart Disease Prediction System Using Data Mining Techniques

Sellappan Palaniappan, Rafiah Awang
Department of Information Technology
Malaysia University of Science and Technology
Block C, Kelana Square, Jalan SS7/26 Kelana Jaya,
47301 Petaling Jaya, Selangor, Malaysia
sell@must.edu.my , rafyea99@yahoo.com

Abstract

The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not "mined" to discover hidden information for effective decision making. Discovery of hidden patterns and relationships often goes unexploited. Advanced data mining techniques can help remedy this situation. This research has developed a prototype Intelligent Heart Disease Prediction System (IHDPS) using data mining techniques, namely, Decision Trees, Naïve Bayes and Neural Network. Results show that each technique has its unique strength in realizing the objectives of the defined mining goals. IHDPS can answer complex "what if" queries which traditional decision support systems cannot. Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease. It enables significant knowledge, e.g. patterns, relationships between medical factors related to heart disease, to be established. IHDPS is Web-based, user-friendly, scalable, reliable and expandable. It is implemented on the .NET platform.

1. Motivation

A major challenge facing healthcare organizations (hospitals, medical centers) is the provision of quality services at affordable costs. Quality service implies diagnosing patients correctly and administering treatments that are effective. Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable. Hospitals must also minimize the cost of clinical tests. They can achieve these results by employing appropriate computer-based information and/or decision support systems.

Most hospitals today employ some sort of hospital information systems to manage their healthcare or patient data [12]. These systems typically generate huge amounts of data which take the form of numbers, text, charts and images. Unfortunately, these data are rarely used to support clinical decision making. There is a wealth of hidden information in these data that is largely untapped. This raises an important question: "How can we turn data into useful information that can enable healthcare practitioners to make intelligent clinical decisions?" This is the main motivation for this research.

2. Problem statement

Many hospital information systems are designed to support patient billing, inventory management and generation of simple statistics. Some hospitals use decision support systems, but they are largely limited. They can answer simple queries like "What is the average age of patients who have heart disease?", "How many surgeries had resulted in hospital stays longer than 10 days?", "Identify the female patients who are single, above 30 years old, and who have been treated for cancer." However, they cannot answer complex queries like "Identify the important preoperative predictors that increase the length of hospital stay", "Given patient records on cancer, should treatment include chemotherapy alone, radiation alone, or both chemotherapy and radiation?", and "Given patient records, predict the probability of patients getting a heart disease."

Clinical decisions are often made based on doctors' intuition and experience rather than on the knowledge-rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. Wu, et al proposed that integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome [17]. This suggestion is promising as data modelling and analysis tools, e.g., data mining, have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions.

978-1-4244-1968-5/08/$25.00 ©2008 IEEE

3. Research objectives

The main objective of this research is to develop a prototype Intelligent Heart Disease Prediction System (IHDPS) using three data mining modeling techniques, namely, Decision Trees, Naïve Bayes and Neural Network. IHDPS can discover and extract hidden knowledge (patterns and relationships) associated with heart disease from a historical heart disease database. It can answer complex queries for diagnosing heart disease and thus assist healthcare practitioners to make intelligent clinical decisions which traditional decision support systems cannot. By providing effective treatments, it also helps to reduce treatment costs. To enhance visualization and ease of interpretation, it displays the results both in tabular and graphical forms.

4. Data mining review

Although data mining has been around for more than two decades, its potential is only being realized now. Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases [15]. Fayyad defines data mining as "a process of nontrivial extraction of implicit, previously unknown and potentially useful information from the data stored in a database" [4]. Giudici defines it as "a process of selection, exploration and modelling of large quantities of data to discover regularities or relations that are at first unknown with the aim of obtaining clear and useful results for the owner of database" [5].

Data mining uses two strategies: supervised and unsupervised learning. In supervised learning, a training set is used to learn model parameters whereas in unsupervised learning no training set is used (e.g., k-means clustering is unsupervised) [12].

Each data mining technique serves a different purpose depending on the modelling objective. The two most common modelling objectives are classification and prediction. Classification models predict categorical labels (discrete, unordered) while prediction models predict continuous-valued functions [6]. Decision Trees and Neural Networks use classification algorithms while Regression, Association Rules and Clustering use prediction algorithms [3].

Decision Tree algorithms include CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3) and C4.5. These algorithms differ in selection of splits, when to stop a node from splitting, and assignment of class to a non-split node [7]. CART uses Gini index to measure the impurity of a partition or set of training tuples [6]. It can handle high dimensional categorical data. Decision Trees can also handle continuous data (as in regression) but they must be converted to categorical data.

Naive Bayes or Bayes' Rule is the basis for many machine-learning and data mining methods [14]. The rule (algorithm) is used to create models with predictive capabilities. It provides new ways of exploring and understanding data. It learns from the "evidence" by calculating the correlation between the target (i.e., dependent) and other (i.e., independent) variables.

Neural Networks consists of three layers: input, hidden and output units (variables). Connection between input units and hidden and output units are based on relevance of the assigned value (weight) of that particular input unit. The higher the weight the more important it is. Neural Network algorithms use Linear and Sigmoid transfer functions. Neural Networks are suitable for training large amounts of data with few inputs. It is used when other techniques are unsatisfactory.

5. Methodology

IHDPS uses the CRISP-DM methodology to build the mining models. It consists of six major phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Business understanding phase focuses on understanding the objectives and requirements from a business perspective, converting this knowledge into a data mining problem definition, and designing a preliminary plan to achieve the objectives. Data understanding phase uses the raw the data and proceeds to understand the data, identify its quality, gain preliminary insights, and detect interesting subsets to form hypotheses for hidden information. Data preparation phase constructs the final dataset that will be fed into the modeling tools. This includes table, record, and attribute selection as well as data cleaning and transformation. The modeling phase selects and applies various techniques, and calibrates their parameters to optimal values. The evaluation phase evaluates the model to ensure that it achieves the business objectives. The deployment phase specifies the tasks that are needed to use the models [3].

Data Mining Extension (DMX), a SQL-style query language for data mining, is used for building and accessing the models' contents. Tabular and graphical visualizations are incorporated to enhance analysis and interpretation of results.

5.1. Data source

A total of 909 records with 15 medical attributes (factors) were obtained from the Cleveland Heart Disease database [1]. Figure 1 lists the attributes. The records were split equally into two datasets: training dataset (455 records) and testing dataset (454 records). To avoid bias, the records for each set were selected randomly.

For the sake of consistency, only categorical attributes were used for all the three models. All the non-categorical medical attributes were transformed to categorical data.

The attribute "Diagnosis" was identified as the predictable attribute with value "1" for patients with heart disease and value "0" for patients with no heart disease. The attribute "PatientID" was used as the key; the rest are input attributes. It is assumed that problems such as missing data, inconsistent data, and duplicate data have all been resolved.

Predictable attribute

1. Diagnosis (value 0: < 50% diameter narrowing (no heart disease); value 1: > 50% diameter narrowing (has heart disease))

Key attribute

1. PatientID – Patient's identification number

Input attributes

1. Sex (value 1: Male; value 0 : Female)
2. Chest Pain Type (value 1: typical type 1 angina, value 2: typical type angina, value 3: non-angina pain; value 4: asymptomatic)
3. Fasting Blood Sugar (value 1: > 120 mg/dl; value 0: < 120 mg/dl)
4. Restecg – resting electrographic results (value 0: normal; value 1: 1 having ST-T wave abnormality; value 2: showing probable or definite left ventricular hypertrophy)
5. Exang – exercise induced angina (value 1: yes; value 0: no)
6. Slope – the slope of the peak exercise ST segment (value 1: unsloping; value 2: flat; value 3: downsloping)
7. CA – number of major vessels colored by floursopy (value 0 – 3)
8. Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)
9. Trest Blood Pressure (mm Hg on admission to the hospital)
10. Serum Cholesterol (mg/dl)
11. Thalach – maximum heart rate achieved
12. Oldpeak – ST depression induced by exercise relative to rest
13. Age in Year

Figure 1. Description of attributes

5.2. Mining models

Data Mining Extension (DMX) query language was used for model creation, model training, model prediction and model content access. All parameters were set to the default setting except for parameters "Minimum Support = 1" for Decision Tree and "Minimum Dependency Probability = 0.005" for Naïve Bayes [10]. The trained models were evaluated against the test datasets for accuracy and effectiveness before they were deployed in IHDPS. The models were validated using Lift Chart and Classification Matrix.

5.3. Validating model effectiveness

The effectiveness of models was tested using two methods: Lift Chart and Classification Matrix. The purpose was to determine which model gave the highest percentage of correct predictions for diagnosing patients with a heart disease.

Lift Chart with predictable value. To determine if there was sufficient information to learn patterns in response to the predictable attribute, columns in the trained model were mapped to columns in the test dataset. The model, predictable column to chart against, and the state of the column to predict patients with heart disease (predict value = 1) were also selected. Figure 2 shows the Lift Chart output. The X-axis shows the percentage of the test dataset used to compare predictions while the Y-axis shows the percentage of values predicted to the specified state. The blue and green lines show the results for random-guess and ideal model respectively. The purple, yellow and red lines show the results of Neural Network, Naïve Bayes and Decision Tree models respectively.

The top green line shows the ideal model; it captured 100% of the target population for patients with heart disease using 46% of the test dataset. The bottom blue line shows the random line which is always a 45-degree line across the chart. It shows that if we randomly guess the result for each case, 50% of the target population would be captured using 50% of the test dataset. All three model lines (purple, yellow and red) fall between the random-guess and ideal model lines, showing that all three have sufficient information to learn patterns in response to the predictable state.

Lift Chart with no predictable value. The steps for producing Lift Chart are similar to the above except that the state of the predictable column is left blank. It does not include a line for the random-guess model. It tells how well each model fared at predicting the correct number of the predictable attribute. Figure 3 shows the Lift Chart output. The X-axis shows the

percentage of test dataset used to compare predictions while the Y-axis shows the percentage of predictions that are correct. The blue, purple, green and red lines show the ideal, Neural Network, Naïve Bayes and Decision Trees models respectively. The chart shows the performance of the models across all possible states. The model ideal line (blue) is at 45-degree angle, showing that if 50% of the test dataset is processed, 50% of test dataset is predicted correctly.

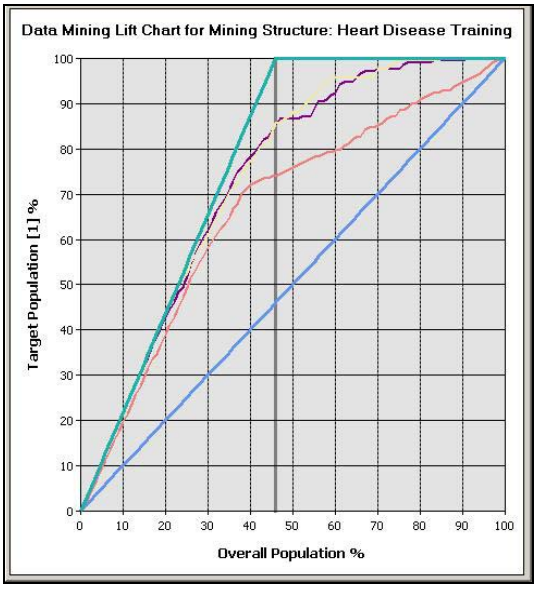

Figure 2. Result of Lift Chart with predictable value

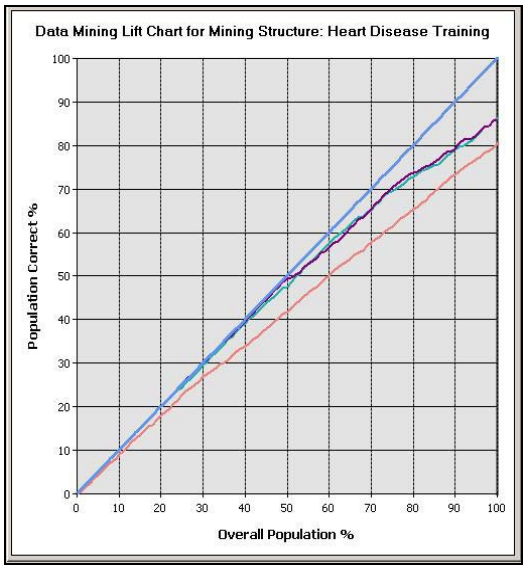

Figure 3. Result of Lift Chart without predictable value

The chart shows that if 50% of the population is processed, Neural Network gives the highest percentage of correct predictions (49.34%) followed by Naïve Bayes (47.58%) and Decision Trees (41.85%). If the entire population is processed, Naïve Bayes model appears to perform better than the other two as it gives the highest number of correct predictions (86.12%) followed by Neural Network (85.68%) and Decision Trees (80.4%).

Processing less than 50% of the population causes the Lift lines for Neural Network and Naïve Bayes to be always higher than that for Decision Trees, indicating that Neural Network and Naïve Bayes are better at making high percentage of correct predictions than Decision Trees. Along the X-axis the Lift lines for Neural Network and Naïve Bayes overlap, indicating that both models are equally good for predicting correctly. When more than 50% of population is processed, Neural Network and Naïve Bayes appear to perform better as they give high percentage of correct predictions than Decision Trees. This is because the Lift line for Decision Trees is always below that of Neural Network and Naïve Bayes. For some population range, Neural Network appears to fare better than Naives Bayes and vice-versa.

Classification Matrix. Classification Matrix displays the frequency of correct and incorrect predictions. It compares the actual values in the test dataset with the predicted values in the trained model. In this example, the test dataset contained 208 patients with heart disease and 246 patients without heart disease. Figure 4 shows the results of the Classification Matrix for all the three models. The rows represent predicted values while the columns represent actual values (1 for patients with heart disease, '0' for patients with no heart disease). The left-most columns show values predicted by the models. The diagonal values show correct predictions.

Counts for Decision Tree on Diagnosis Group

	Predicted	0 (Actual)	1 (Actual)
	0	219	62
	1	27	146

Counts for Naive Bayes on Diagnosis Group

	Predicted	0 (Actual)	1 (Actual)
	0	211	28
	1	35	180

Counts for Neural Network on Diagnosis Group

	Predicted	0 (Actual)	1 (Actual)
	0	211	30
	1	35	178

Figure 4. Results of Classification Matrix for all the three models

Figure 5 summarizes the results of all three models. Naïve Bayes appears to be most effective as it has the highest percentage of correct predictions (86.53%) for patients with heart disease, followed by Neural Network (with a difference of less than 1%) and Decision Trees. Decision Trees, however, appears to be most effective for predicting patients with no heart disease (89%) compared to the other two models.

Model Type	Prediction Attributes	No. of cases	Prediction
Decision Tree	+WHD, +PHD	146	Correct
	–WHD, +PHD	27	Incorrect
	–WHD, –PHD	219	Correct
	+WHD, –PHD	62	Incorrect
Naïve Bayes	+WHD, +PHD	180	Correct
	–WHD, +PHD	35	Incorrect
	–WHD, –PHD	211	Correct
	+WHD, –PHD	28	Incorrect
Neural Network	+WHD, +PHD	178	Correct
	–WHD, +PHD	35	Incorrect
	–WHD, –PHD	211	Correct
	+WHD, –PHD	30	Incorrect

Legend
+WHD: Patients with heart disease
–WHD: Patients with no heart disease
+PHD: Patients predicted as having heart disease
–PHD: Patients predicted as having no heart disease

Figure 5. Model results

5.4. Evaluation of Mining Goals

Five mining goals were defined based on exploration of the heart disease dataset and objectives of this research. They were evaluated against the trained models. Results show that all three models had achieved the stated goals, suggesting that they could be used to provide decision support to doctors for diagnosing patients and discovering medical factors associated with heart disease. The goals are as follows:

Goal 1: Given patients' medical profiles, predict those who are likely to be diagnosed with heart disease. All three models were able to answer this question using singleton query and batch or prediction join query. Both queries could predict on single input cases and multiple input cases respectively. IHDPS supports prediction using "what if" scenarios. Users enter values of medical attributes to diagnose patients with heart disease. For example, entering values Age = 70, CA = 2, Chest Pain Type = 4, Sex = M, Slope = 2 and Thal = 3 into the models, would produce the output in Figure 6. All three models showed that this patient has a heart disease. Naïve Bayes gives the highest probability (95%) with 432 supporting cases, followed closely by Decision Tree (94.93%) with 106 supporting cases and Neural Network (93.54%) with 298 supporting cases. As these values are high, doctors could recommend that the patient should undergo further heart examination. Thus performing "what if" scenarios can help prevent a potential heart attack.

Goal 2: Identify the significant influences and relationships in the medical inputs associated with the predictable state – heart disease. The Dependency viewer in Decision Trees and Naïve Bayes models shows the results from the most significant to the least significant (weakest) medical predictors. The viewer is especially useful when there are many predictable attributes. Figures 7 and 8 show that in both models, the most significant factor influencing heart disease is "Chest Pain Type". Other significant factors include Thal, CA and Exang. Decision Trees model shows 'Trest Blood Pressure" as the weakest factor while Naïve Bayes model shows 'Fasting Blood Sugar' as the weakest factor. Naïve Bayes appears to fare better than Decision Trees as it shows the significance of all input attributes. Doctors can use this information to further analyze the strengths and weaknesses of the medical attributes associated with heart disease.

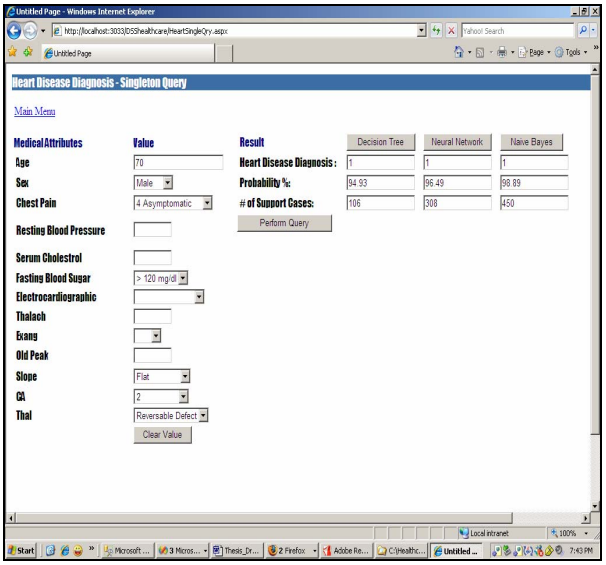

Figure 6. Output for singleton query module

Goal 3: Identify the impact and relationship between the medical attributes in relation to the predictable state – heart disease. Identifying the impact and relationship between the medical attributes in relation to heart disease is only found in Decision

Trees viewer (Figure 9). It gives a high probability (99.61%) that patients with heart disease are found in the relationship between the attributes (nodes): "Chest Pain Type = 4 and CA = 0 and Exang = 0 and Trest Blood Pressure >= 146.362 and < 158.036." Doctors can use this information to perform medical screening on these four attributes instead of on all attributes on patients who are likely to be diagnosed with heart disease. This will reduce medical expenses, administrative costs, and diagnosis time. Information on least impact (5.88%) is found in the relationship between the attributes: "Chest Pain Type not = 4 and Sex = F". Also given is the relationship between attributes for patients with no heart disease. Results show that the relationship between the attributes: "Chest Pain Type not = 4 and Sex = F" has the highest impact (92.58%). The least impact (0.2%) is found in the attributes: "Chest Pain Type = 4 and CA = 0 and Exang = 0 and Trest Blood Pressure >= 146.362 and < 158.036". Additional information such as identifying patients' medical profiles based selected nodes can also be obtained by using the drill through function. Doctors can use the Decision Tree viewer to perform further analysis.

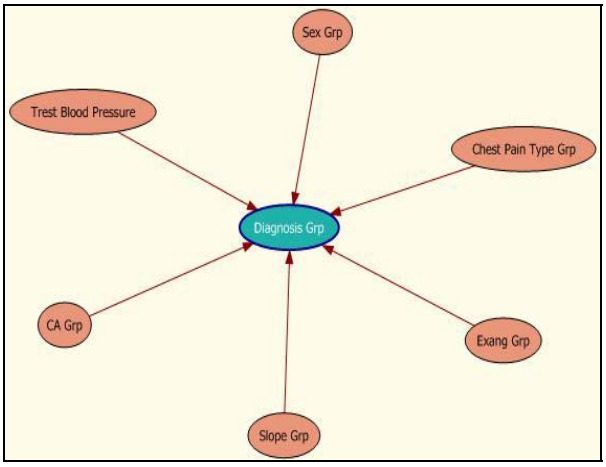

Figure 7. Decision Trees dependency network

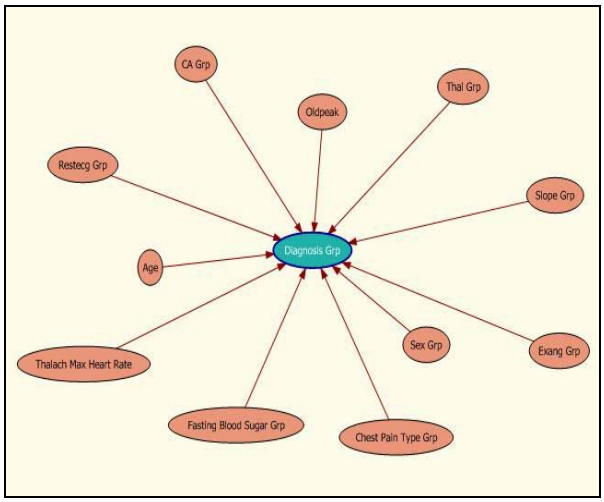

Figure 8. Dependency network for Naïve Bayes

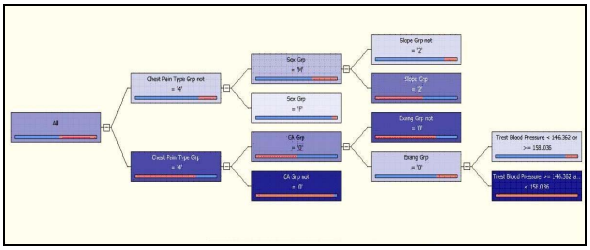

Figure 9. Decision Trees Viewer

Goal 4: Identify characteristics of patients with heart disease. Only Naïve Bayes model identifies the characteristics of patients with heart disease. It shows the probability of each input attribute for the predictable state. Figure 10 shows that 80% of the heart disease patients are males (Sex = 1) of which 43% are between ages 56 and 63. Other significant characteristics are: high probability in fasting blood sugar with less than 120 mg/dl reading, chest pain type is asymptomatic, slope of peak exercise is flat, etc.

Figure 11 shows the characteristics of patients with no heart disease with high probability in fasting blood sugar with less than 120 mg/dl reading, no exercise induced, number of major vessels is zero, etc. These results can be further analyzed.

Attributes	Values	Probability %
FastingBloodSugar	FastingBloodSugar = 0	86.179
Exang	Exang = 0	83.74
CA	ca = 0	80.488
Thal	thal = 3	79.268
Oldpeak	Oldpeak < 0.63	67.073
Slope	slope = 1	65.854
Restecg	Restecg = 0	57.724
Sex	Sex = 1	56.911
Sex	Sex = 0	43.089
Restecg	Restecg = 2	41.463
Chest	ChestPainType = 3	41.057
ThalachMaxHeartRate	ThalachMaxHeartRate >= 167.58	38.211
1 2 3 4		

Figure 10. Naïve Bayes Attribute Characteristics Viewer in descending order for patients with heart disease

Attributes	Values	Probability %
FastingBloodSugar	FastingBloodSugar = 0	86.179
Exang	Exang = 0	83.74
CA	ca = 0	80.488
Thal	thal = 3	79.268
Oldpeak	Oldpeak < 0.63	67.073
Slope	slope = 1	65.854
Restecg	Restecg = 0	57.724
Sex	Sex = 1	56.911
Sex	Sex = 0	43.089
Restecg	Restecg = 2	41.463
Chest	ChestPainType = 3	41.057
ThalachMaxHeartRate	ThalachMaxHeartRate >= 167.58	38.211
1 2 3 4		

Figure 11. Naïve Bayes Attribute Characteristic Viewer in descending order for patients with no heart disease

Goal 5: Determine the attribute values that differentiate nodes favoring and disfavoring the predictable states: (1) patients with heart disease (2) patients with no heart disease. This query can be answered by analyzing the results of attribute discrimination viewer of Naïve Bayes and Neural Network models. The viewer provides information on the impact of all attribute values that relate to the predictable state. Naïve Bayes model (Figure 12) shows the most important attribute favoring patients with heart disease: "Chest Pain Type = 4" with 158 cases and 56 patients with no heart disease. The input attributes "Thal = 7" with 123 (75.00%) patients, "Exang = 1" with 112 (73.68%) patients," Slope =2" with 138 (66.34%) patients, etc. also favor predictable state. In contrast, the attributes "Thal = 3" with 195 (73.86%) patients, "CA = 0" with 198 (73.06%) patients, "Exang = 0" with 206 (67.98%), etc. favor predictable state for patients with no heart disease.

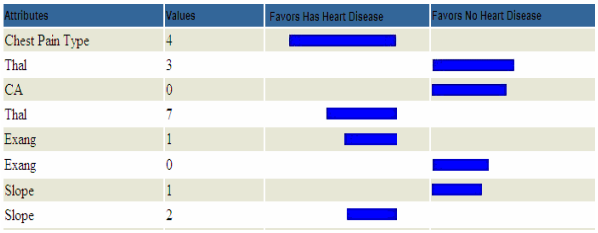

Figure 12. A Tornado Chart for Attribute Discrimination Viewer in descending order for Naïve Bayes

Neural Network model (Figure 13) shows that the most important attribute value that favors patients with heart disease is "Old peak = 3.05 – 3.81" (98%). Other attributes that favor heart disease include "Old peak >= 3.81", "CA=2", "CA=3", etc. Attributes like "Serum Cholesterol >= 382.37", "Chest Pain Type = 2", "CA =0", etc. also favor the predictable state for patients with no heart disease.

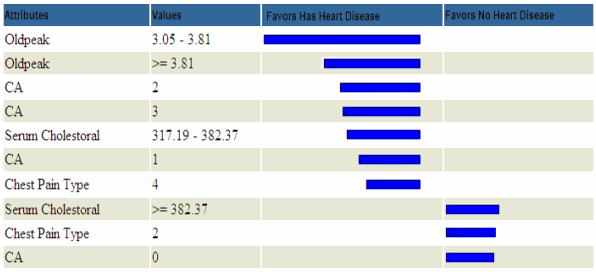

Figure 13. Attribute Discrimination Viewer in descending order for Neural Network

6. Benefits and limitations

IHDPS can serve a training tool to train nurses and medical students to diagnose patients with heart disease. It can also provide decision support to assist doctors to make better clinical decisions or at least provide a "second opinion."

The current version of IHDPS is based on the 15 attributes listed in Figure 1. This list may need to be expanded to provide a more comprehensive diagnosis system. Another limitation is that it only uses categorical data. For some diagnosis, the use of continuous data may be necessary. Another limitation is that it only uses three data mining techniques. Additional data mining techniques can be incorporated to provide better diagnosis. The size of the dataset used in this research is still quite small. A large dataset would definitely give better results. It is also necessary to test the system extensively with input from doctors, especially cardiologists, before it can be deployed in hospitals. [Access to the system is currently restricted to stakeholders.]

7. Conclusion

A prototype heart disease prediction system is developed using three data mining classification modeling techniques. The system extracts hidden knowledge from a historical heart disease database. DMX query language and functions are used to build and access the models. The models are trained and validated against a test dataset. Lift Chart and Classification Matrix methods are used to evaluate the effectiveness of the models. All three models are able to extract patterns in response to the predictable state. The most effective model to predict patients with heart disease appears to be Naïve Bayes followed by Neural Network and Decision Trees.

Five mining goals are defined based on business intelligence and data exploration. The goals are evaluated against the trained models. All three models could answer complex queries, each with its own strength with respect to ease of model interpretation, access to detailed information and accuracy. Naïve Bayes could answer four out of the five goals; Decision Trees, three; and Neural Network, two. Although not the most effective model, Decision Trees results are easier to read and interpret. The drill through feature to access detailed patients' profiles is only available in Decision Trees. Naïve Bayes fared better than Decision Trees as it could identify all the significant medical predictors. The relationship between attributes produced by Neural Network is more difficult to understand.

IHDPS can be further enhanced and expanded. For example, it can incorporate other medical attributes besides the 15 listed in Figure 1. It can also incorporate other data mining techniques, e.g., Time Series, Clustering and Association Rules. Continuous data can also be used instead of just categorical data. Another area is to use Text Mining to mine the vast amount of unstructured data available in healthcare databases. Another challenge would be to integrate data mining and text mining [16].

7. References

[1] Blake, C.L., Mertz, C.J.: "UCI Machine Learning Databases", http://mlearn.ics.uci.edu/databases/heart-disease/, 2004.

[2] Chapman, P., Clinton, J., Kerber, R. Khabeza, T., Reinartz, T., Shearer, C., Wirth, R.: "CRISP-DM 1.0: Step by step data mining guide", SPSS, 1-78, 2000.

[3] Charly, K.: "Data Mining for the Enterprise", 31st Annual Hawaii Int. Conf. on System Sciences, IEEE Computer, 7, 295-304, 1998.

[4] Fayyad, U: "Data Mining and Knowledge Discovery in Databases: Implications fro scientific databases", Proc. of the 9th Int. Conf. on Scientific and Statistical Database Management, Olympia, Washington, USA, 2-11, 1997.

[5] Giudici, P.: "Applied Data Mining: Statistical Methods for Business and Industry", New York: John Wiley, 2003.

[6] Han, J., Kamber, M.: "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2006.

[7] Ho, T. J.: "Data Mining and Data Warehousing", Prentice Hall, 2005.

[8] Kaur, H., Wasan, S. K.: "Empirical Study on Applications of Data Mining Techniques in Healthcare", Journal of Computer Science 2(2), 194-200, 2006.

[9] Mehmed, K.: "Data mining: Concepts, Models, Methods and Algorithms", New Jersey: John Wiley, 2003.

[10] Mohd, H., Mohamed, S. H. S.: "Acceptance Model of Electronic Medical Record", Journal of Advancing Information and Management Studies. 2(1), 75-92, 2005.

[11] Microsoft Developer Network (MSDN). http://msdn2.microsoft.com/en-us/virtuallabs/aa740409.aspx, 2007.

[12] Obenshain, M.K: "Application of Data Mining Techniques to Healthcare Data", Infection Control and Hospital Epidemiology, 25(8), 690–695, 2004.

[13] Sellappan, P., Chua, S.L.: "Model-based Healthcare Decision Support System", Proc. Of Int. Conf. on Information Technology in Asia CITA'05, 45-50, Kuching, Sarawak, Malaysia, 2005

[14] Tang, Z. H., MacLennan, J.: "Data Mining with SQL Server 2005", Indianapolis: Wiley, 2005.

[15] Thuraisingham, B.: "A Primer for Understanding and Applying Data Mining", IT Professional, 28-31, 2000.

[16] Weiguo, F., Wallace, L., Rich, S., Zhongju, Z.: "Tapping the Power of Text Mining", Communication of the ACM. 49(9), 77-82, 2006.

[17] Wu, R., Peters, W., Morgan, M.W.: "The Next Generation Clinical Decision Support: Linking Evidence to Best Practice", Journal Healthcare Information Management. 16(4), 50-55, 2002.