Proceedings of the IASTED International Conference
Computational Intelligence (CI 2009)
August 17 - 19, 2009 Honolulu, Hawaii, USA

DIAGNOSIS HEART DISEASE USING AN ASSOCIATION RULE DISCOVERY APPROACH

Jesmin Nahar[1], Kevin S. Tickle[1], Shawkat Ali[1], Yi-Ping Phoebe Chen[2]
[1]School of Computing Sciences, Central Queensland University, Queensland, Australia
E-mail :{j.nahar, k.tickle, s.ali@cqu.edu.au}
[2]Faculty of Science and Technology, Deakin University, Victoria, Australia
E-mail: phoebe@deakin.edu.au

ABSTRACT
Around the world heart disease is the number one killer of humans. The most common cause of heart disease is narrowing or blockage of the coronary arteries, the blood vessels that supply blood to the heart itself. Heart problems occur as a result of genetic heredity or daily life style factors. This research involves an investigation of association rules predicting healthy and sick heart status using heart disease data. This research is significant in that useful rules are identified that enable the prediction of healthy/unhealthy heart status from real life factors.

KEY WORDS
Heart Disease, Classification, Association Rule Mining.

1. Introduction

Heart disease makes many individuals around the world vulnerable to experiencing early death. Heart disease is constituted by disorders that significantly decrease the heart's capacity to function normally. The heart is the life of the body, pumping blood and its life-giving oxygen and nutrients to all parts of the body. If the systematic pumping action of the heart becomes inefficient, the functioning of essential organs like the brain and kidneys can suffer. In addition, if the heart stops working altogether, death occurs within minutes. So, life itself is absolutely reliant on the efficient operation of the heart (1).

Different forms of heart disease include: Alcoholic Cardiomyopathy, Aortic Regurgitation, Aortic Stenosis, Arrhythmias, Cardiogenic Shock, Congenital Heart Disease, Coronary Artery Disease (CAD) Dilated Cardiomyopathy, Endocarditis, Heart Attack (myocardial infarction), Heart Failure, Heart Tumor, Hypertrophic Cardiomyopathy, Idiopathic Cardiomyopathy, Ischemic Cardiomyopathy, Acute Mitral regurgitation, Chronic Mitral Regurgitation, Mitral Stenosis, Mitral Valve Prolapse, Peripartum Cardiomyopathy, Pulmonary Stenosis, Stable Angina and Unstable Angina, and Tricuspid Regurgitation (2). Early detection of heart diseases is an extremely important area of medical research (3). A recent statistical study reported that out of more than 71 million American adults, 27,400,000 over the age of 65 years are suffering one or more types of cardiovascular disease (CVD) (4). Timely treatment of heart disease can be lifesaving; however, correct diagnoses must first be made (5). Unfortunately, accurate diagnosis of heart diseases has certainly not been a simple process. In fact, a range of factors can make it difficult to diagnose types of heart diseases, frequently resulting in a delay in making a correct diagnostic decision. To reduce the time required to make a diagnosis and to improve the diagnostic accuracy, it has become an increasing challenge to develop dependable and powerful medical diagnostic systems that will support the complex diagnostic decision-making process (5).

2. Related Work

Gamboa et al (6) suggested the application of Fuzzy Support Vector Clustering to the identification of heart diseases and their severity, so that warning signs can be provided regarding patients' health status. The algorithm employs a kernel induced metric to allot each piece of data to a cluster and the SVM density assessment algorithm to parameterize clusters (to identify membership degrees matrix). Palaniappan & Awang (7) developed a prototype Intelligent Heart Disease Prediction System (IHDPS) using data mining techniques, namely, Decision Trees, Naïve Bayes and Neural Network. The experimental results demonstrated that each technique has its inimitable strength in realizing the objectives of the defined mining goals. El-Hanjouri et al (8) suggested the use of an improved diagnostic system which uses heart sounds to identify different heart diseases. In their experiment wavelet decomposition and me1cepstrum were used for feature extraction. They used Hidden Markov Models (HMM) to classify different types of heart disease. Ordonez (9) established an algorithm that uses search constraints to lessen the number of rules, searches for association rules on a training set, and at last authenticates them on an independent test set. In this way, important rules with high confidence, high lift, or both, that remain applicable on the test set on several runs, are established. These experimental rules symbolize important medical knowledge.

Obayya & Abou-Chadi (10) utilised a multi-layer feed forward neural network to distinguish between ordinary subjects and patients with such diseases as congestive heart failure (CHF) and myocardial infarction. In their experiment three different techniques were used to choose the inputs used to predict classification categories. The experimental result showed the average classification rate achieved to be 96.36% using non-linear techniques. To develop the classification rate, data fusion at feature extraction level was adopted. It has been established that the average classification rate has been improved to reach 98.18%. Their research recommends strongly the data fusion approach for classifying heart diseases through the use of the heart rate variability signals. Avci (11) studied an intelligent system based on genetic-support vector machines (GSVM), an approach which is offered for classification of the Doppler signals of disease of the heart valve. In their research the intelligent system dealt with a combination of the feature extraction and classification from calculated Doppler signal waveforms at the heart valve using the Doppler ultrasound. The GSVM chooses the most appropriate wavelet filter type for the problem, the wavelet entropy parameter, the optimal kernel function type, the kernel function parameter, and the soft margin constant C penalty parameter of support vector machines (SVM) classifier. The investigation results showed that this GSVM system is effective in detecting Doppler heart sounds whereas the averaged rate of a correct classification is 95%.

Maglogiannis et al (12) recommends the use of an automated detection system for the identification of heart valve diseases based on the Support Vector Machines (SVM) classification of heart sounds. In the beginning the heart sounds were fruitfully categorized using an SVM classifier as normal or disease-related and then the parallel murmurs in the unhealthy cases were classified as systolic or diastolic. Zheng et al (5) proposed the use of committee machines (CM) based on ensembles of multilayer perceptron (MLP) to distinguish the detection of five major heart diseases. In this research particularly, two ensemble systems - random subspace and bagging were considered. Ordonez et al (13) focused on mapping medical data to a transaction format appropriate for mining association rules and classifying useful constraints. In their research they introduced an improved algorithm to discover constrained association rules. Kim et al (14) proposed to combine observation and the investigative skills of oriental medicine used in prevention, to create a system that diagnoses and categorizes cardiovascular diseases using IT technology. Youn et al (15) proposed the use of a system reflecting development in diagnosis and system performance. The experimental results suggested that the proposed physio-grid system offers logical medical services guaranteeing high system performance in data management and high competence in diagnosis.

Shi et al (16) investigated the classification of five different shapes of ST (Stress Test) segment using fuzzy adaptive resonance theory mapping (ARTMAP) of neural networks. The planned method was established using by the data from the standard MIT/BIH ECG database. Results showed that the fuzzy ARTMAP might be used to differentiate the shapes of ST segment successfully. Soman et al (17) analysed Ischemic –heart disease (IHD) data from 65 patients to aid decision making at diagnosis. They found that decision trees provide potent structural information about the data, serving as a powerful data mining tool. On the other hand support Vector machines provide excellent classification with high accuracy prediction.

3. Data collection

This is Dr. Detrano's database modified to a real (18) MIXED dataset. In that data set the original attributes is : age; sex (1,0); cp (1-4); trestbps; chol; fbs (1,0); restecg (0,1,2); thalach; exang (1,0); oldpeak; slope (1,2,3); ca; thal (3,6,7); class att: 0 is healthy, 1,2,3,4 is sick.

Dr. Detrano's (18) modified the original data to a real mixed dataset from Cleveland, Hungary, Switzerland, and the VA Long Beach. As a result, the 14 factors identified as most useful were coded in the following way in the current study.

1. Age: age in years
2. Sex: (1 = male; 0 = female)
3. Chest pain type (CP), Value 1: typical angina; Value 2: atypical angina; Value 3: non-anginal pain, Value 4: asymptomatic
4. Resting blood pressure (trestbps) (in mm Hg on admission to the hospital)
5. Serum cholestoral (chol) in mg/dl
6. Fasting blood sugar (fbs) > 120 mg/dl) (1 = true; 0 = false)
7. Resting electrocardiographic (restecg) results
8. Maximum heart rate achieved (thalach)
9. Exercise induced angina (exang) (1 = yes; 0 = no)
10. Stress Test (oldpeak) depression induced by exercise relative to rest
11. The slope of the peak exercise ST segment (slope), Value 1: upsloping; Value 2: flat; Value 3: downsloping
12. Number of major vessels (0-3) colored by flourosopy (ca)
13. Thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
14. The predicted attribute (num): diagnosis of heart disease (angiographic disease status)

-- Value 0: < 50% diameter narrowing; -- Value 1: > 50% diameter narrowing.

4. Association Rule Mining

Association rule mining (ARM) is being applied to search for interesting relationships between healthy and sick hearts from data in the heart disease database. In this section the standard definition of association rules (19, 20, 21, 22, and 23) is introduced.

The basic formulation of ARM can be summarized as follows (19).

- $I = \{i_1, i_2, ..., i_n\}$ – set of causes of heart disease; Healthy heart
- D = set of test outcome T; $T \subseteq I$;
- X = set of causes from I, T contains X.
- An association rule is a pair $X \Rightarrow Y$, where $X \subseteq I$, $Y \subseteq I$, $X \cap Y = \phi$.

The strength of rules is assessed using support and confidence levels.

- The support of rule $X \Rightarrow Y$ is defined as the percentage of transactions containing both X and Y in D.
- The confidence of $X \Rightarrow Y$ is defined as the percentage of transactions containing X that also contain Y in D.

The sets of items X and Y are called antecedent (left-hand-side or LHS) and consequent (right-hand-side or RHS) to the rule, respectively. The association rule generations are used to satisfy a user-specified minimum support and user-specified minimum confidence.

The task for association rule mining can be divided into two fields:

- Find all combinations of the attribute for a specific heart disease whose supports are greater than a user-specified minimum support level (threshold).
- Use the attribute of heart disease from frequent heart disease data set to generate the desired rules. Generally the confidence of each rule is computed, and if it is above the confidence threshold, then the rule in the system is retrieved.

Amongst the long list of association learning algorithms Apriori (19), Predictive Apriori (24), Tertius (25) are the most popular algorithms in the machine learning community. The Apriori algorithm is a state of the art algorithm; with most of the association rule algorithms constituting variations of this algorithm (19). The apriori, predictive apriori, and tertius algorithms were utilized in our experiment. More details can be found in Witten and Frank (26).

5. Experimental Results

We considered three association rule mining algorithms: Apriori, Predictive Apriori and Tertius for our experiment. The apriori algorithm was found to be the optimal algorithm for the extraction of useful rules in the heart disease data set. Rules were selected based on confidence levels, with all the reported rules obtaining confidence levels above 90%. On the basis of experimental results the generated rules for identification of healthy and sick heart are described in the section below.

5.1 Healthy Rules

Healthy rule: If **{**female ∩ exercise_induced_angina (false) ∩ number_of_vessels_colored (0)} ⇒ then the people might be healthy, not suffering for heart disease. Conf: (0.98).

Healthy rule: If **{**female ∩ exercise_induced_angina (false) ∩ number_of_vessels_colored (0) ∩ thal (normal)}⇒ then the people might be healthy, not suffering for heart disease. Conf -0.98

Healthy rule: If **{**female ∩ fasting_blood_sugar (false) ∩ exercise_induced_angina (false) ∩ number_of_vessels_colored (0)}⇒ then the people might be healthy, not suffering for heart disease. Conf: (0.98).

Healthy rule: If **{**female ∩ fasting_blood_sugar (false) ∩ exercise_induced_angina (false) ∩ thal (normal)} ⇒ then the people might be healthy, not suffering for heart disease. Conf :(0.95).

Healthy rule: If **{**female ∩ number_of_vessels_colored (0) ∩ thal (normal) ∩ Fasting blood sugar (false) } ⇒ then the people might be healthy, not suffering for heart disease. Conf :(0.95).

Healthy rule: If **{**female ∩ number_of_vessels_colored (0) ∩ fasting_blood_sugar (false)} ⇒ then the people might be healthy, not suffering for heart disease. Conf :(0.95).

Healthy rule: If **{**slope (up) ∩ number_of_vessels_colored (0) ∩ thal (normal)}⇒ then the people might be healthy, not suffering for heart disease. Conf :(0.94).

5.2 Sick Rules

Sick rule: If **{**chest_pain_type(asymptomatic) ∩ slope (flat) ∩ thal (reversible)} ⇒ then the people might be sick and suffering for heart disease. Conf :(0.96).

Sick rule: If **{**chest_pain_type (asymptomatic) ∩ exercise_induced_angina (TRUE) ∩ thal (reversible)} ⇒ then the people might be sick and suffering for heart disease. Conf :(0.94).

Sick rule: If **{**fasting_blood_sugar (false) ∩ exercise_induced_angina (TRUE) ∩ chest_pain_type (asymptomatic)} ⇒ then the people is sick and suffering for heart disease. Conf :(0.94).

Sick rule: If **{**exercise_induced_angina (TRUE) ∩ thal (reversible) ∩ chest_pain_type (asymptomatic)} ⇒ then the people might be sick and suffering for heart disease. Conf :(0.92).

Sick rule: If **{**male ∩ fasting_blood_sugar (false) ∩ exercise_induced_angina (TRUE) ∩ chest_pain_type (asymptomatic)} ⇒ then the people might be sick and suffering for heart disease. Conf :(0.92).

Sick rule: If **{**exercise_induced_angina (TRUE) ∩ chest_pain_type (asympt)} ⇒ then the people might be sick and suffering for heart disease. Conf :(0.92).

6. Conclusion

In this experiment we used the three association rule mining algorithms of apriori, predictive apriori and tertius. Finally, on the basis of experimental results we suggested that the Apriori algorithm is the best suited algorithm for this type of task. The experimental results show that all the generated rules hold the highest (above 90 %) confidence level. Therefore, we observed the generated rules to be very useful in identification of heart disease and the maintenance of healthy hearts. In future, we have planned to consider rules based on supervised learning algorithms to extract significant factors for different heart diseases.

References

[1] Heart (2009). http://heartdisease.about.com/od/coronaryarterydisease/a/heartdisease.htm, accessed 18th May, 2009.

[2] Heart disease (2009). http://www.healthline.com/channel/heart-disease.html, 18th May.2009.

[3] M. Akay, Y. M. Akay, & W. Welkowitz, Neural networks for the diagnosis of coronary artery disease, *International Joint Conference on Neural Networks*, IJCNN, (2), 1992, 419–424.

[4] T. Thom et al., "Heart disease and stroke statistics – 2006 update. A report from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee," Circulation, 114(23), 2006, e630. (doi: 10.1 161/CIRCULATIONAHA.105.171600).

[5] J. Zheng, Y. Jiang, & H. Yan, Committee Machines with Ensembles of Multilayer Perceptron for the Support of Diagnosis of Heart Diseases. *Proc. International Conference on, Communications, Circuits and Systems*, (3), 2006, 2046 – 2050.

[6] A. L. G. Gamboa, M.G. Mendoza, R. E. I. Orozco, J. M. Vargas, & N. H. Gress, Hybrid Fuzzy-SV Clustering for Heart Disease Identification**,** Computational Intelligence for Modelling, *International Conference on Control and Automation, 2006 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce,* 2006, 121-121.

[7] Palaniappan & Awang. Intelligent heart disease prediction system using data mining techniques, *International Conference on Computer Systems and Applications, AICCSA, IEEE/ACS,* 2008,108 – 115.

[8] M. El-Hanjouri, W. Alkhaldi, N. Hamdy, & O.A. Alim, Heart diseases diagnosis using HMM. 11th Mediterrane *Electrotechnical Conference, MELECON,* 2002. 489 – 492.

[9] Ordonez, C. Association rule discovery with the train and test approach for heart disease prediction. *IEEE Transactions on Information Technology in Biomedicine,* 10(2), 2006. 334 – 343.

[10] M. Obayya, & F. Abou-Chadi, Data fusion for heart diseases classification using multi-layer feed forward neural network**,** *International Conference on Computer Engineering & Systems, ICCES*, 2008. 67 – 70.

[11] E. Avci, A new intelligent diagnosis system for the heart valve diseases by using genetic-SVM classifier, *Expert Systems with Applications*, 36 (7), 2009, 10618-10626.

[12] I. Maglogiannis, E. Loukis, E. Zafiropoulos, & S. Stasis, Support Vectors Machine-based identification of heart valve diseases using heart sounds, *Computer Methods and Programs in Biomedicine*, 95(1), 2009, 47-61.

[13] C. Ordonez, E. Omiecinski, L. de Braal, C. A. Santana, N. Ezquerra, J. A. Taboada, D. Cooke, E. Krawczynska, & E. V. Garcia, Mining constrained association rules to predict heart disease. *Proc. IEEE International Conference on , Data Mining, ICDM* , 2001, 433 – 440.

[14] B-h. Kim, S-h. Lee, D-U. Cho, & S-Y. Oh, A Proposal of Heart Diseases Diagnosis Method Using Analysis of Face Color, *International Conference on Advanced Language Processing and Web Information Technology, ALPIT '* 2008. 220 – 225.

[15] C-H. Youn, C.H. Han, Y. Han, S.S. Kwon, & E.B. m, Physio-grid system for advanced identification of heart disease, *e-health Networking, Applications and Services, 10th International Conference on Health Com,* 2008. 13 – 17.

[16] L. Shi, Z. Sun, H. Li, & W. Liu, Research on Diagnosing Coronary Heart Disease using Fuzzy Adaptive Resonance Theory Mapping Neural Networks, *IEEE International Conference on Control and Automation, ICCA,* 2007. 1126 – 1128.

[17] K.P. Soman, D.M. Shyam, & P. Madhavdas, Efficient classification and analysis of ischemic heart disease using proximal support vector machines based decision trees, *TENCON, Conference on Convergent Technologies for Asia-Pacific Region,* (1), 2003, 214 – 217.

[18] Heart disease dataset (2009). http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/cleve.mod, accessed 18th May, 2009.

[19] R. T. Agrawal, L. Imielinski, & A.N. Swami, Mining association rules between sets of items in large databases. *Proc. of the 1993 ACM SIGMOD international conference on Management of data,* 1993, 207–216.

[20] C. Ordonez, & E. Omiecinski, Discovering association rules based on image content. *In IEEE*

Advances in Digital Libraries Conference (ADL), 1999, 38–49.

[21] C. Ordonez, C. A. Santana, & L. de. Braal, Discovering interesting association rules in medical data. In *ACM DMKD Workshop*, 2000, 78–85.

[22] J. Nahar, S. Ali, Y.P.P. Chen, Microarray Data Classification Using Automatic SVM Kernel Selection, *DNA and Cell Biology*, 26(10), 2007, 707-712.

[23] T. Scheffer, Finding Association Rules that Trade Support Optimally Against Confidence. *Proc. of the 5th European Conference on Principles and Practice of Knowlege Discovery in Databases(PKDD'01),* Freiburg, Germany : Springer- Verlag, 2001, 424-435.

[24] P.A. Flach, & N. Lachiche, *Confirmation-guided discovery of first-order rules with Tertius*, Kluwer Academic Publishers. The Netherlands, 42, 2001, 61-95.

[25] I. H. Witten, & E. Frank, *Data Mining: Practical machine learning tools and techniques,* 2nd Edition, Morgan Kaufmann, 2000, San Francisco.