Volume 2, Issue 3, March 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Decision Support in Heart Disease System With Subset Selection Using a Hybrid Bacteriological Algorithm

Mrs. M.Rose Margaret[1]
[1]Assistant Professor
Department of Computer Application
CMS College of Science and Commerce, Coimbatore
Email: rosemargaret@gmail.com

Dr.A.V.Senthil Kumar [2]
[2]Director
Department of Computer Application
The Hindustan College of Science and Commerce, Coimbatore

Abstract: The successful application of data mining in highly visible fields like e-business, marketing and retail has led to its application in other industries and sectors. Among these sectors just discovering is healthcare. There is a wealth of data available within the healthcare systems. However, there is a lack of effective analysis tools to discover hidden relationships and trends in data. In our work, bacteriological algorithm is used to determine the attributes which contribute more towards the diagnosis of heart ailments which indirectly reduces the number of tests which are needed to be taken by a patient. An efficient approach for the prediction of heart attack risk levels from heart disease. This algorithm is trained with the selected significant pattern s for the effective prediction of heart attack . This research will be helpful in making a Decision Support in Heart Disease Prediction System (DSHDPS) using data mining modeling technique, namely, Naïve Bayes. Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease.

Keywords: data mining, decision support, heart disease, Naïve Bayes

I. Introduction:

Medical data mining has great potential for exploring the hidden patterns in the data sets of the medical domain. These patterns can be utilized for clinical diagnosis. However, the available raw medical data are widely distributed, heterogeneous in nature, and voluminous. These data need to be collected in an organized form. This collected data can be then integrated to form a hospital information system. Data mining technology provides a user oriented approach to novel and hidden patterns in the data. The term Heart disease encompasses the diverse diseases that affect the heart. Heart disease was the major cause of casualties in the different countries including India. Coronary heart disease, Cardiomyopathy and Cardiovascular disease are some categories of heart diseases. The term "cardiovascular disease" includes a wide range of conditions that affect the heart and the blood vessels and the manner in which blood is pumped and circulated through the body. Cardiovascular disease (CVD) results in several illness, disability, and death. The diagnosis of diseases is a vital and intricate job in medicine. Medical diagnosis is regarded as an important yet complicated task that needs to be executed accurately and efficiently. The automation of this system would be extremely advantageous. Regrettably all doctors do not possess expertise in every sub specialty and moreover there is a shortage of resource persons at certain places. Therefore, an automatic medical diagnosis system would probably be exceedingly beneficial by bringing all of them together. Appropriate computer-based information and/or decision support systems can aid in achieving clinical tests at a reduced cost. Our work attempts to predict efficiently diagnosis with reduced number of factors that contribute more towards the cardiac disease using classification data mining technique[3][4].

II. Basic Classification Concepts

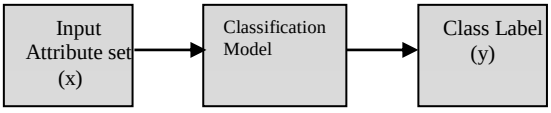
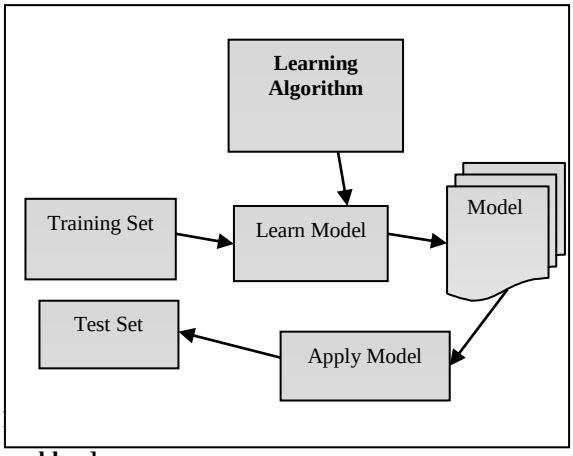


Figure I. Classification Model

Classification which is the task of assigning objects to one of several predefined categories, is a pervasive problem that encompasses many diverse applications. Classification as the task of mapping an input attribute set x into its class label y. A classification model can be used to predict the class label of unknown records. A classification model can be treated as a black box that automatically assigns a class label when presented with the attribute set of an unknown record[7].

A Classification technique is a systematic approach to building classification models from an input dataset. Examples include decision tree classifiers, rule based classifiers, neural networks, support vector machines and naïve Bayes classifiers. Each technique employs a learning algorithm to identify a model that best fits the relationship between the attribute set and class label of the input data. Figure II shows a general approach for solving classification problems[7].



problem]

III. Methodology Used

A. Subset selection with Bacteriological Algorithm

Feature Extraction is the process of detecting and eliminating irrelevant, weakly relevant or redundant attributes or dimensions in a given data set. The goal of feature selection is to find the minimal subset of attributes such that the resulting probability distribution of data classes is close to original distribution obtained using all attributes. Search for an optimal subset would be highly expensive especially when n and the number of data classes increases. Sometimes it may be infeasible. Therefore most of the feature selection techniques are heuristic methods. These heuristic methods are greedy in nature and try to explore possible reduced search space. Feature selection techniques fall under two categories. First, feature ranking techniques and second, feature subset selection techniques. In the former, all features are ranked by a metric like information gain, chi-square etc. The features that do not achieve the adequate score are eliminated. In the later, the search is for optimal subset of features that would be equivalent to original subset of features. The subset of features are evaluated more commonly based on distance metrics like Euclidean, Hamming etc or filter metrics like Entropy or Probabilistic distance. Common search approaches include greedy forward attribute selection, backward attribute selection, simulated annealing, and Genetic Algorithms.

- choose an initial population calculate the fitness value for each individual
- reproduction
- crossover
- mutation on one or several individuals
- several stopping criteria : x number of generations, a given fitness value reached

The global process of a GA

B. The Bacteriological Model

The bacteriological approach is more an adaptive approach than an optimization approach as with GAs. It aims at mutating the initial population to adapt it to a particular environment. The adaptation is only based on small changes in the individuals. The individuals in the population are called *bacteria* and correspond to *atomic units*. Unlike the genetic model the bacteria cannot be divided. The crossover operation cannot be used anymore. Bacteria can only be reproduced and altered to improve the population.

C. Process of Hybrid Bacteriological Model

- choose an initial population
- calculate the fitness value for each individual
- reproduction
- mutation on one or several individuals
- several stopping criteria : x number of generations, a given fitness value reached

…

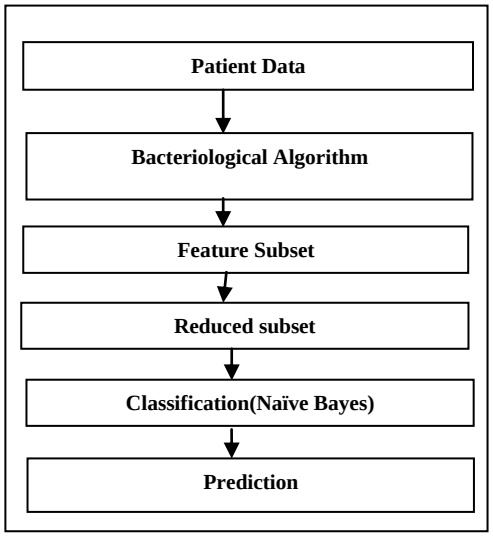
Process of Hybrid Bacteriological Model

As with the genetic model, a *fitness function* is necessary to choose bacteria for reproduction. With this function a global iterative process to adapt an initial population is given Starting from this population, the fitness function allows the algorithm to select the best bacteria. Then these bacteria are saved and reproduced to generate a new population. Several bacteria in this population

are mutated, then the best ones are selected again to produce another generation. This process stops after a number of generations or when the memorized population has reached an optimum fitness value.

IV. Implementation of Bayesian Classification

The Naïve Bayes Classifier technique is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often out perform more sophisticated classification methods. Naïve Bayes model identifies the characteristics of patients with heart disease. It shows the probability of each input attribute for the predictable state. Naive Bayes or Bayes' Rule is the basis for many machine-learning and data mining methods. The rule (algorithm) is used to create models with predictive capabilities. It provides new ways of exploring and understanding data.



Overview of the System
[Prediction of Heart Disease]

V. Experiments And Results

Data set with 13 attributes(bacteria) were used. All attributes are made categorical and inconsistencies are resolved for simplicity. To enhance the prediction of classifiers, bacteriological algorithm is incorporated. Each attribute is treated as a bacteria. Algorithm is applied to reduce the number of bacteria's that contribute more for the heart disease.

	C1	C2
C1	True Positives 57	False Negatives 2
C2	False Positives 1	True Negatives 50

Table 2 Classification matrix for the classifier

Input Attributes

1. Sex
2. Chest Pain Type
3. Fasting Blood Sugar
4. Restecg – resting electrographic
5. Exang – exercise induced angina
6. Slope – the slope of the peak exercise ST segment
7. CA – number of major vessels colored by floursopy
8. Thal
9. Trest Blood Pressure
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

Reduced Input Attributes as a result of bacteriological Algorithm

1. Type - Chest Pain Type
2. Rbp - Resting blood pressure
3. Eia - Exercise induced angina
4. Oldpk - Old peak
5. Vsl - No. of vessels colored
6. Thal -Maximum heart rate achieved
7. Serum Cholesterol (mg/dl)

Naïve Bayes classifier was used for diagnosis of heart disease and range values were shown below for reduced subset

Type Chest Pain Type value 1: typical type 1 angina, value 2:typical type angina, value 3: non-angina pain; value 4: asymptomatic

Trest bps Resting blood pressure in (mmHg) BP<80 BP-Normal BP<90 BP-Normal-to-High ,BP>90 BP-High

Chol Serum cholestral in mg/dl Chol<200 Chol-Normal Chol<400 Chol-High Chol>400 Chol-Severe

Thalach Maximum heart rate achieved (value) <100 MHR-Normal ,MHR-High,MHR-Severe

Old peak St depression induced by exercise relative to rest 0 or1

Exang Exercise induced angina (value) <1.5 DE-Normal(value) <2.5 DE-Normal-to-High (value) >2.5 DE-High

Vsl number of major vessels colored (value 0 – 3)

Input Attributes list and description

The classifiers was fed with reduced data set with 7 attributes (Table 2). Data set of 110 records were used. Results are shown in Table 3. Observations exhibit that the Naïve Bayes classifier performs good accuracy with incorporating feature subset selection using bacteriological algorithm.

Classifier	Naïve Bayes
No Of Attributes	13
No of Attributes in the reduced set	7
Accuracy	97.2%
Error rate	0.027

Table 3 Result of the classifier

6. Conclusion:

Decision Support in Heart Disease Prediction System is developed using Naive Bayesian Classification technique. The system extracts hidden knowledge from a historical heart disease database. This is the most effective model to predict patients with heart disease. The system predicts with good accuracy with less no of attributes. DSHDPS can be further enhanced and expanded by applying with other data mining techniques for subset selection. It can also incorporate other medical attributes besides the above list.

References

[1] Bressan, M. and J. Vitria (2003): On the selection and classification of independent features, Pattern Analysis and Machine Intelligence,IEEE Transactions. pp. 1312-1317.

[2] Carlos Ordonez (2006): Comparing Association Rules and Decision Trees for Disease Prediction, *ACM, HIKM*

[3] Sellappan Palaniappan and Rafiah Awang (2008): Intelligent Heart Disease Prediction System Using Data Mining Techniques, 978-1-4244-1968- 5/08/ IEEE.

[4] Shantakumar B.Patil and Y.S.Kumaraswamy (2009): Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network, European Journal of Scientific Research ISSN 1450- 216X Vol.31 No.4, pp. 642-656.

[5] Carloz Ordonez, *Association Rule Discovery with Train and Test approach for heart disease prediction*, IEEE Transactions on Information Technology in Biomedicine,
Volume 10, No. 2, April 2006.pp 334-343.

[6] Mrs.G.Subbalakshmi (M.Tech), Mr. K. Ramesh M.Tech, Asst. Professor, Mr. M. Chinna Rao ,*Decision Support in*
Heart Disease Prediction System using Naive Bayes, Indian Journal of Computer Science and Engineering (IJCSE) Vol. 2 No. 2 Apr-May 2011.

[7] Tan, Steinbach, Kumar Lecture Notes for Data Mining Classification: Basic Concepts, Decision Trees, and Model Evaluation ,Chapter 4.

Author Biographies

M.Rose Margaret presently working as assistant professor Department of computer Appliction in CMS College of Science and Commerce. She is presently pursuing her Research in Department of Computer Science, Bharathiar University,Coimbatore . Her field of research includes Data Mining, Artificial intelligence and Fuzzy Logic.

Dr. A.V.Senthil Kumar obtained his BSc Degree (Physics) in 1987, P.G.Diploma in Computer Applications in 1988, MCA in 1991 from Bharathiar University. He obtained his Master of Philosophy in Computer Science from Bharathidasan University, Trichy during 2005 and his Ph.D in Computer Science from Vinayaka Missions University during 2009. To his credit he had industrial experience for five years as System Analyst in a Garment Export Company. Later he took up teaching and attached to CMS College of Science and Commerce, Coimbatore and now he working as Director & Professor in the Department of MCA since 05/03/2010. He has to his credit 3 Book Chapters, 10 papers in International Journals, 2 papers in National Journals, 13 papers in International Conferences, 5 papers in National Conferences, and edited a book in Data Mining (IGI Global, USA) and a book in Mobile Computing (IGI Global, USA). He is an Editor-in-Chief for International Journal titled "International Journal of Data Mining and Emerging Technologies", "International Journal of Image Processing and Applications", "International Journal of Advances in Knowledge Engineering & Computer Science" and "International

Journal of Research and Reviews in Computer Science".. **Key Member** for India, Machine Intelligence Research Lab (MIR Labs). He is an Editorial Board Member and Reviewer for various International Journals. He is also a Committee member for various International Conferences. He is a Life member of International Association of Engineers (IAENG), Systems Society of India (SSI), member of The Indian Science Congress Association, member of Internet Society (ISOC), International Association of Computer Science and Information Technology (IACSIT), Indian Association for Research in Computing Science (IARCS), and committee member for various International Conferences. He has got many awards from National and International Societies. Also a freelance writer for Tamil Computer (a fortnightly) and PC Friend (monthly).

