Asian Journal of Applied Science and Technology (AJAST)
(Peer Reviewed Quarterly International Journal) Volume 2, Issue 4, Pages 221-226, Oct-Dec 2018

A Hybrid Classification Paradigm for Disease Prediction

S. Ambika[1] & Mr. Saravana Kumar[2]

[1]Research Scholar, Department of Computer Science, RVS Technical Campus (Affiliated to Anna University), Coimbatore-641402, Tamil Nadu, India.
[2]Assistant Professor, Department of Computer Science, RVS Technical campus (Affiliated to Anna University), Coimbatore-641402, Tamil Nadu, India.

Article Received: 30 May 2018 Article Accepted: 29 September 2018 Article Published: 21 November 2018

ABSTRACT

Most hospitals today employ some sort of hospital information systems to manage their healthcare or patient data. 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. 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) based on genetic algorithm and Naive Bayes algorithm. Results show that each technique has its unique strength in realizing the objectives of the defined mining goals. IHDPS can answer queries which cannot be constructed by a single question or information seeking query. A set of interlinked queries have to be processed to get the desired decision making remedy from the prediction revealing system. 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 Solution rendering Knowledge base where effective retrieval techniques have to be used for making the best use of the untapped knowledge that was updated over the period of past years.

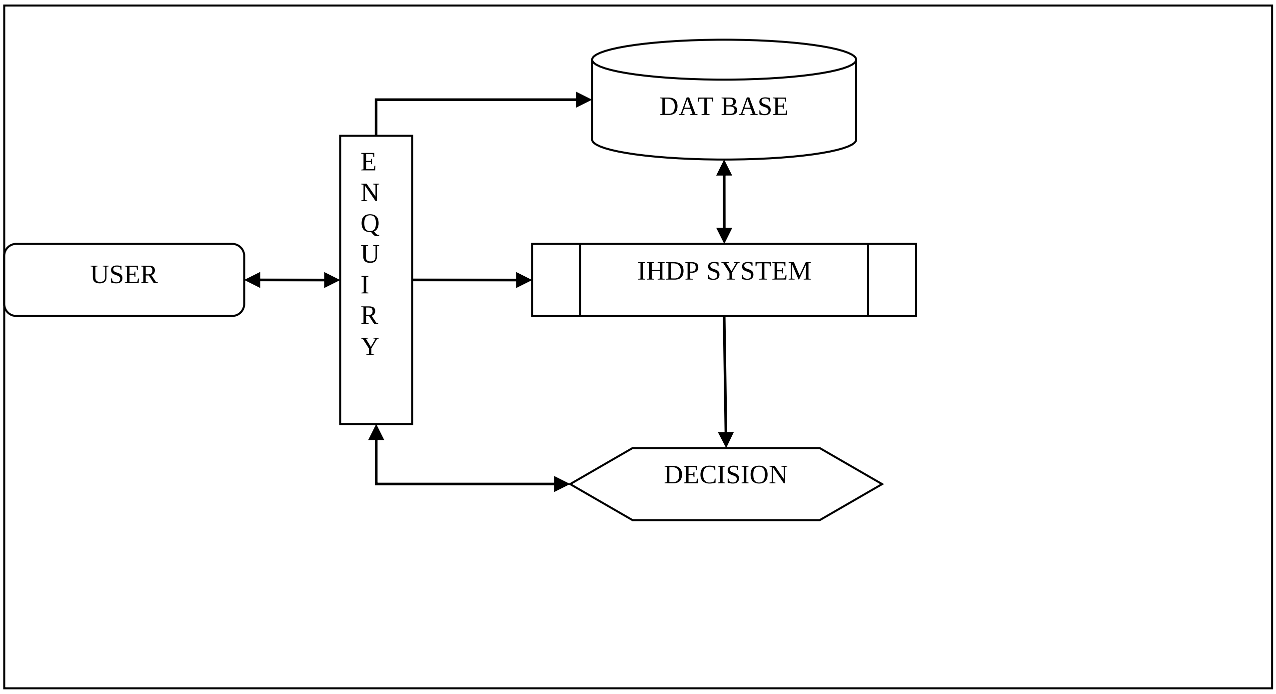
1. INTRODUCTION

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. 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 project.

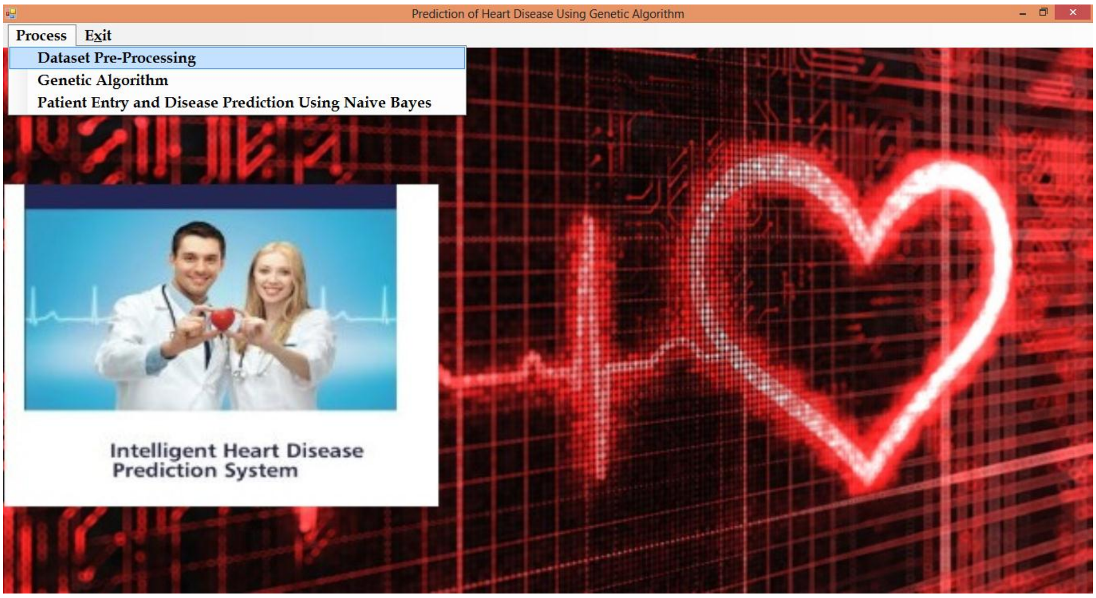
2. PROBLEM STATEMENT

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. 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. This suggestion is promising as data modeling 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.

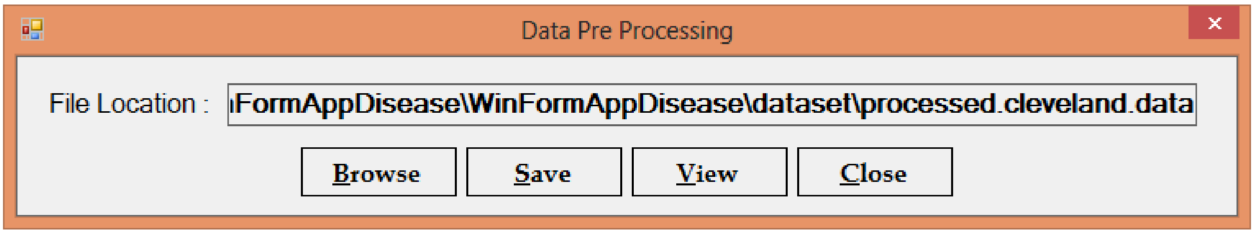


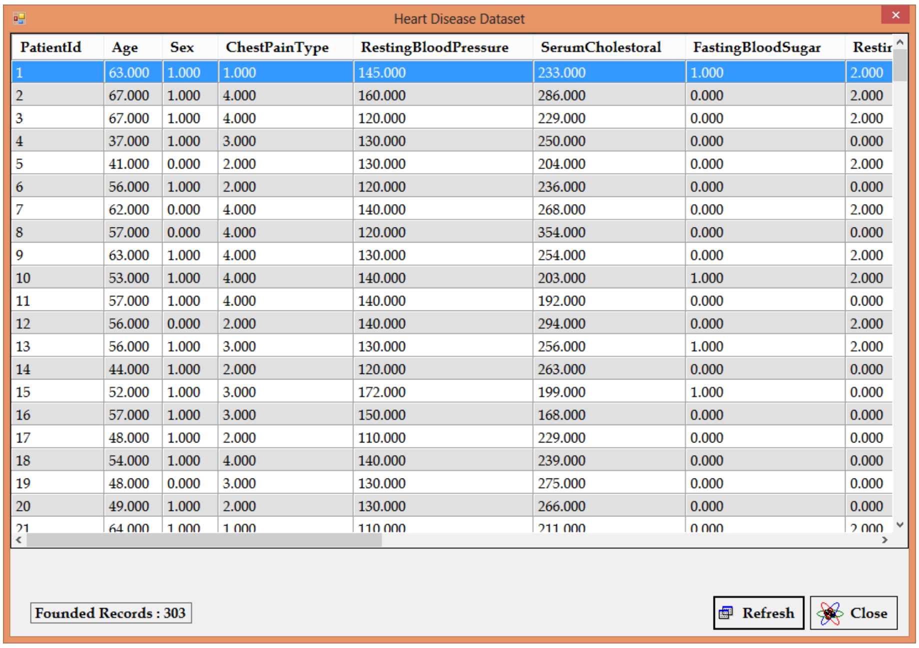
3. METHODS

3.1. PREPROCESSING OF THE PATIENTS DATABASES FROM DIFFERENT DEPARTMENTS



This facility enables the system to consolidate required information's from different depots and optimize loading over multiple trailers and to remove the dirty information's as well. After pre processing the data has been saved in the table in data base.

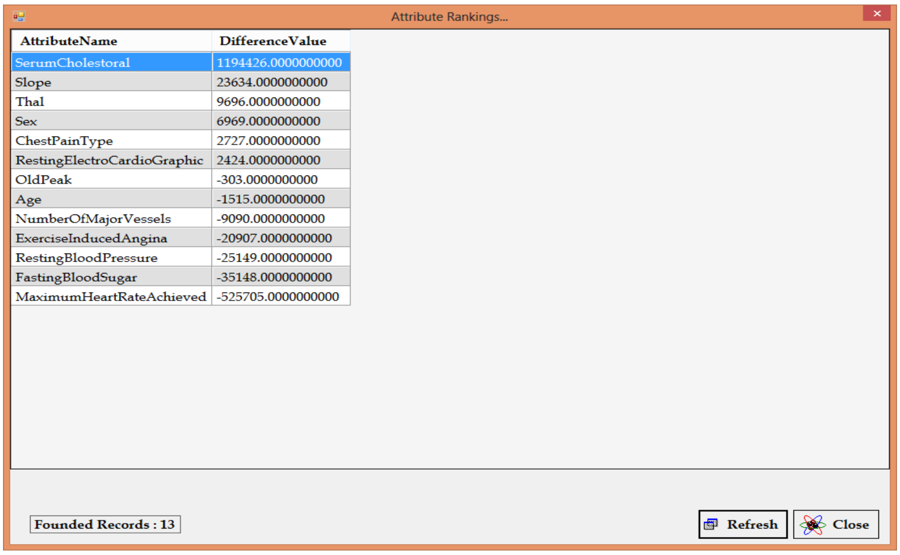




PatientId	Age	Sex	ChestPainType	RestingBloodPressure	SerumCholestoral	FastingBloodSugar	Restir
1	63.000	1.000	1.000	145.000	233.000	1.000	2.000
2	67.000	1.000	4.000	160.000	286.000	0.000	2.000
3	67.000	1.000	4.000	120.000	229.000	0.000	2.000
4	37.000	1.000	3.000	130.000	250.000	0.000	0.000
5	41.000	0.000	2.000	130.000	204.000	0.000	2.000
6	56.000	1.000	2.000	120.000	236.000	0.000	0.000
7	62.000	0.000	4.000	140.000	268.000	0.000	2.000
8	57.000	0.000	4.000	120.000	354.000	0.000	0.000
9	63.000	1.000	4.000	130.000	254.000	0.000	2.000
10	53.000	1.000	4.000	140.000	203.000	1.000	2.000
11	57.000	1.000	4.000	140.000	192.000	0.000	0.000
12	56.000	0.000	2.000	140.000	294.000	0.000	2.000
13	56.000	1.000	3.000	130.000	256.000	1.000	2.000
14	44.000	1.000	2.000	120.000	263.000	0.000	0.000
15	52.000	1.000	3.000	172.000	199.000	1.000	0.000
16	57.000	1.000	3.000	150.000	168.000	0.000	0.000
17	48.000	1.000	2.000	110.000	229.000	0.000	0.000
18	54.000	1.000	4.000	140.000	239.000	0.000	0.000
19	48.000	0.000	3.000	130.000	275.000	0.000	0.000
20	49.000	1.000	2.000	130.000	266.000	0.000	0.000
21	64.000	1.000	1.000	110.000	211.000	0.000	2.000

3.2. IDENTIFYING OF IMPORTANT ATTRIBUTE USING GA

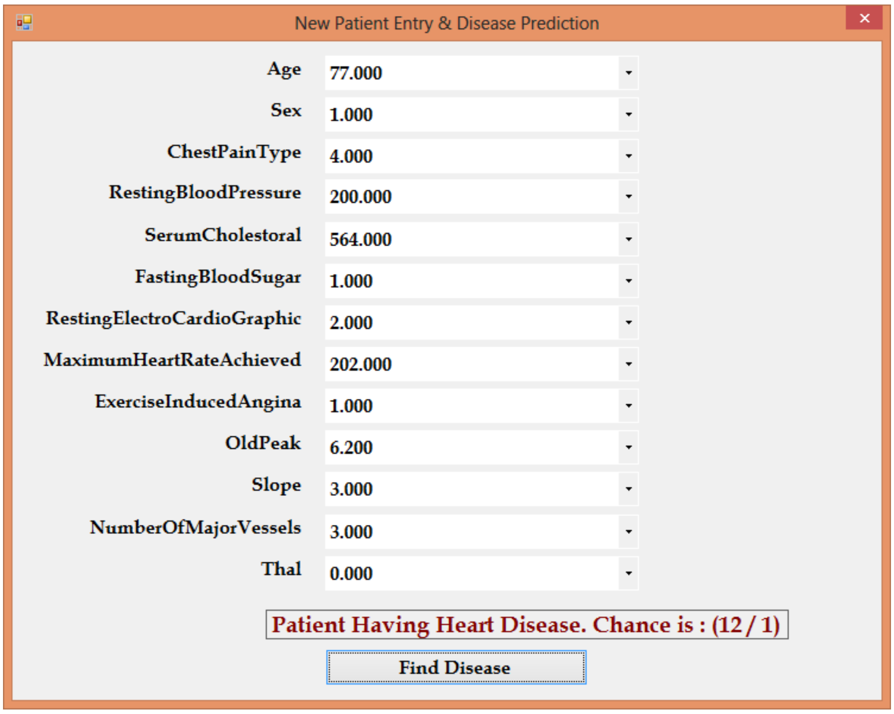
The pre process data has been saved in the tables. Genetic algorithm is used for attribute (column) ranking. We are going to find important attribute using GA.



AttributeName	DifferenceValue
SerumCholestoral	1194426.0000000000
Slope	23634.0000000000
Thal	9696.0000000000
Sex	6969.0000000000
ChestPainType	2727.0000000000
RestingElectroCardioGraphic	2424.0000000000
OldPeak	-303.0000000000
Age	-1515.0000000000
NumberOfMajorVessels	-9090.0000000000
ExerciseInducedAngina	-20907.0000000000
RestingBloodPressure	-25149.0000000000
FastingBloodSugar	-35148.0000000000
MaximumHeartRateAchieved	-525705.0000000000

3.3. DISEASE PREDICTION USING NAÏVE BAYES

The Naïve Bayes – for disease prediction this rule (algorithm) is used to create models with predictive capabilities. It provides new ways of exploring and understanding data. This module improves the performance and decision accuracy of Naïve Bayes algorithms with improved Heart disease prediction system.



4 ALGORITHMS USED

4.1 GENETIC ALGORITHM

Genetic algorithms are based on the principle of genetic modification, mutation and natural selection. These are algorithmic optimization strategies inspired by the principles observed in natural evolution. The genetic algorithm creates a number of random solutions to the problem. All these solutions may not be good, a group of solutions can be skipped entirely, and it can come down to the overlapping solutions. Poor solutions are discarded, and the good ones retained. A good solution is then being hybridized, and then the whole process is repeated. Finally, similar to the process of natural selection, only the best solutions remain. So, from the set of potential solutions to the problems that competes with each other, the best solutions are chosen and combined with each other in order to obtain a universal solution from the set of solutions that will become better and better, similar to the process of evolution of organisms. Genetic algorithms are used in data mining to formulate hypotheses about the dependencies between variables in the form of association rules or other internal formalism.

4.2 NAÏVE BAYES

Naïve Bayes, a data mining modeling technique to develop a Heart Disease Prediction System which is implemented as web-based questionnaire application. It uses an historical heart disease database to generate relationships and thus, predict the risk level based on the values entered by the user. Bayes' Rule is used to create models that have predictive capabilities.

4.3 DATABASE DESIGN

A Database design is a collection of stored data organized in such a way that the data requirements are specified by the database. The general objective is to make information access easy, quick, inexpensive and flexible for the user. There are also some specific objectives like controlled redundancy from failure, privacy, security and performance.

FIELD	DATATYPE
Age	numeric(7,3)
Sex	numeric(7,3)
Chest Pain Type	numeric(7,3)
Resting Blood Pressure	numeric(7,3)
Serum Cholesterol	numeric(7,3)
Fasting Blood Sugar	numeric(7,3)
Resting Electro Cardio Graphic	numeric(7,3)
Maximum Heart Rate Achieved	numeric(7,3)
Exercise Induced Angina	numeric(7,3)
Old Peak	numeric(7,3)
Slope	numeric(7,3)
Number of Major Vessels	numeric(7,3)
Thal	numeric(7,3)
Diagnosis of Heart Disease	numeric(7,3)

Table 4.1 Database design for processed dataset

5. CONCLUSION

A prototype heart disease prediction system is developed using GA, Navie Bayes. The system extracts hidden knowledge from a historical heart disease database. DMX (Data Mining Extensions) query language and functions are used to build and access the models. The models are trained and validated against a test dataset. The most effective model to predict patients with heart disease appears to be Naïve Bayes. The goals are evaluated against the trained models. All the models discussed could answer complex queries, each with its own strength with respect to ease of model interpretation, access to detailed information and accuracy. This technique further improves in future enhancement.

REFERENCES

[1.] Blake, C.L., Mertz, C.J.: "UCI Mach. Learning Databases", http://mlearn.ics.uci.edu/databases/heartdisease/, 2004.

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

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

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

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

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

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

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

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