PREDICTION OF HEART DISEASE USING GENETIC ALGORITHM FOR SELECTION OF OPTIMAL REDUCED SET OF ATTRIBUTES

[1]SHRUTI RATNAKAR, [2]K. RAJESWARI, [3]ROSE JACOB

Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune, Maharashtra, India

Abstract: The healthcare industry collects huge amounts of healthcare data which is not feasible to handle manually. Due to advancements in technology, today, many hospitals use information systems to store and manage data. These large amounts of data are very important in the field of data mining to extract useful information and generate relationships amongst the attributes. Thus, data mining is used to develop a mechanism to predict risk of heart disease. With the tremendously growing population, the doctors and experts available are not in proportion with the population. Also, symptoms of heart disease may not be significant and thus, may often be neglected. So we propose an Intelligent Heart Disease Decision Support System to help the doctors reach out to those people who are deprived of these medical services. In general, it can serve as a training tool to train nurses and medical students to diagnose patients having risk of heart disease.
In this paper we have discussed two modeling techniques: Naïve Bayes' Rule and Genetic Algorithm which predict the risk level of heart disease. In Genetic Algorithm, optimal reduced set of attributes are found using Genetic Search method. The 13 attributes in the original list have been reduced to 6. In Naïve Bayes' technique, a historical heart disease database is used to generate relationships amongst the attributes using the concepts of conditional probability.

Keywords: Data mining, Genetic Algorithm, Naïve Bayes' Rule, Decision trees

I. INTRODUCTION

In the present era many technologies and scientific innovations like decision support systems, image and scanning systems are available to support the doctors in clinical decision making. But these services are costly and are not able to reach the remote areas. As a result, the poor and the needy are not able to access these services. Also, the clinical decisions are often made based on doctors' knowledge and experience. This may lead to erroneous conclusions due to misinterpretations which is really a serious matter.
World Health Organization in the year 2012 reported that 11.8% of the total global deaths (in US) are due to Cardio Vascular Disease. It is very important to develop a system which will help us derive at right conclusions.
Data mining is the solution to this serious problem. Data mining is an essential step in the process of knowledge discovery in databases. Thus data mining refers to extracting or mining knowledge from large amounts of data.
In this paper, we discuss the paper based on Genetic algorithm and compare it with the paper based on Naïve Bayes' Rule.
The first paper[1] discusses the classification of data mining technique to predict diagnosis of heart ailments efficiently, with reduced number of factors (i.e. attributes) that contribute more towards the cardiac disease.
The second paper[2] discusses Naïve Bayes, a data mining modeling technique which uses the concepts of conditional probability and a historical data set.
The paper is organized as follows:
Section 2 discusses the first paper and Section 3 discusses the second paper mentioned above. The drawbacks of both the papers have been discussed and a new system has been proposed to overcome these drawbacks.

II. CLASSIFICATION TECHNIQUES

In this paper[1], Genetic Algorithm is used to determine the attributes which contribute more towards the diagnosis of heart diseases.

A. Genetic Algorithm

In the field of artificial intelligence, a genetic algorithm (GA) is a search heuristic that imitates the process of natural evolution. This heuristic is routinely used to generate useful solutions to optimization and search problems. Genetic algorithms belong to the larger class of evolutionary algorithms (EA), which generate optimized solutions using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover. A typical genetic algorithm requires:

1. A genetic representation of the solution domain,
2. A fitness function to evaluate the solution domain.

A standard representation of the solution is as an array of bits. The fitness function is defined over the genetic representation and measures the quality of the represented solution. The fitness function is always problem dependent. Initially many individual solutions are (usually) randomly generated to form an initial population. During each successive generation, a proportion of the existing population is selected to breed a new generation. Individual solutions are selected through a fitness-based process, where fitter solutions (as measured by a fitness function) are typically more likely to be selected. The next step is to generate a second generation population of solutions from those selected through genetic

operators: crossover (also called recombination) and mutation.

B. Dataset

The dataset under consideration has been taken from University of California Irvin (UCI). Originally 13 attributes were involved in predicting the heart disease. The thirteen attributes have been shown in Table 1. Feature extraction method has been used to find the minimal subset of attributes that is equivalent to original set of attributes. The number of attributes has been reduced to 6 using Genetic Search. This reduced data is fed to the three classification models. The genetic search starts with zero attribute and an initial population with randomly generated rules. Based on the concept of survival of fittest new population is constructed to obey the fittest rule in the current population, as well as offspring of these rules. The process of generation continues until it evolves a population P where every rule in P satisfies the fitness threshold. With initial population of 20 instances, generation continued till the twentieth generation with cross over probability of 0.6 and mutation probability of 0.033. In this way, genetic search lead to the selection of 6 attributes out of the 13 mentioned (Refer Table 2).

Classification models are used to extract models describing important data classes or to predict future trends. The three classification models used are: Decision Trees and Naïve Bayes.

Decision Trees, which are simple and easy to implement can handle high dimensional data. The performance of this method suffers from repetition and replication for which necessary steps are required.

Naïve Bayes classifier has minimum error rate but inaccuracies are noticed which are caused by assumptions. This method assumes no dependency between attributes.

Following classifier evaluation methods have been discussed in this paper which includes sensitivity, specificity, precision and accuracy.

Sensitivity = t_pos/pos
Specificity = t_neg/neg
Precision = t_pos/(t_pos + f_pos)
Accuracy = Sensitivity pos/(pos+neg)+ Specificity neg/(pos+neg)

where
t_pos: no. of true positives(healthy samples that were correctly classified)
t_neg: no. of true negatives(sick samples
that were correctly classified)
pos: no. of positive(healthy) samples
neg: no. of positive(sick) samples
f_pos: no. of false positives (sick samples that were incorrectly labeled as healthy)
Precision: percentage of samples labeled as healthy

Considering these values for the three classification models, some conclusions have been derived. The Decision Tree data mining technique performs better than the other technique with relatively high model construction time. Naïve Bayes' performs consistently before and after reduction of attributes with the same model construction time. In real time systems, this paper does not provide consistency and, the prediction of the intensity of the disease is not predictable.

III. NAÏVE BAYES

This paper uses 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. "Evidence" forms an integral part of this method.

Table 1:UCI heart dataset

Predictable Atttribute Diagnosis (value 0: <50% diameter narrowing (no heart disease); value 1: >50% diameter narrowing (has heart disease))
Key Attribute Patient ID – Patient's identification number
Input Attributes 1. Age in Year 2. Sex (value 1: Male; value 0: Female) 3. Chest Pain Type (value 1:typical type 1 angina, value 2: typical type angina, value 3:non-angina pain; value 4: asymptomatic) 4. Fasting Blood Sugar (value 1: >120 mg/dl; value 0: <120 mg/dl) 5. Restecg – resting electrographic results (value 0:normal; value 1: having ST-T wave abnormality; value 2: showing probable or definite left ventricular

hypertrophy)
6. Exang - exercise induced angina (value 1: yes; value 0: no)
7. Slope – the slope of the peak exercise ST segment (value 1:unsloping; value 2: flat; value 3: downsloping)
8. CA – number of major vessels colored by floursopy (value 0-3)
9. Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)
10. Trest Blood Pressure (mm Hg on admission to the hospital)
11. Serum Cholestrol (mg/dl)
12. Thalach – maximum heart rate achieved
13. Oldpeak – ST depression induced by exercise

Table 2: List of reduced attributes

Predictable Attribute **Diagnosis** Value 0: No heart disease Value 1: Has heart disease
Reduced Input Attributes Type :Chest Pain Type Rbp :Reduced blood pressure Eia :Exercise Induced Angina Oldpk :Old Peak Vsal :No. of vessels coloured Thal :Maximum Heart Rate achieved

Table 3: Comparison of the two classification models

Modelling Techniques	Accuracy (%)	Model Construction Time(seconds)
Naïve Bayes	96.5	0.02
Decision Tree	99.2	0.09

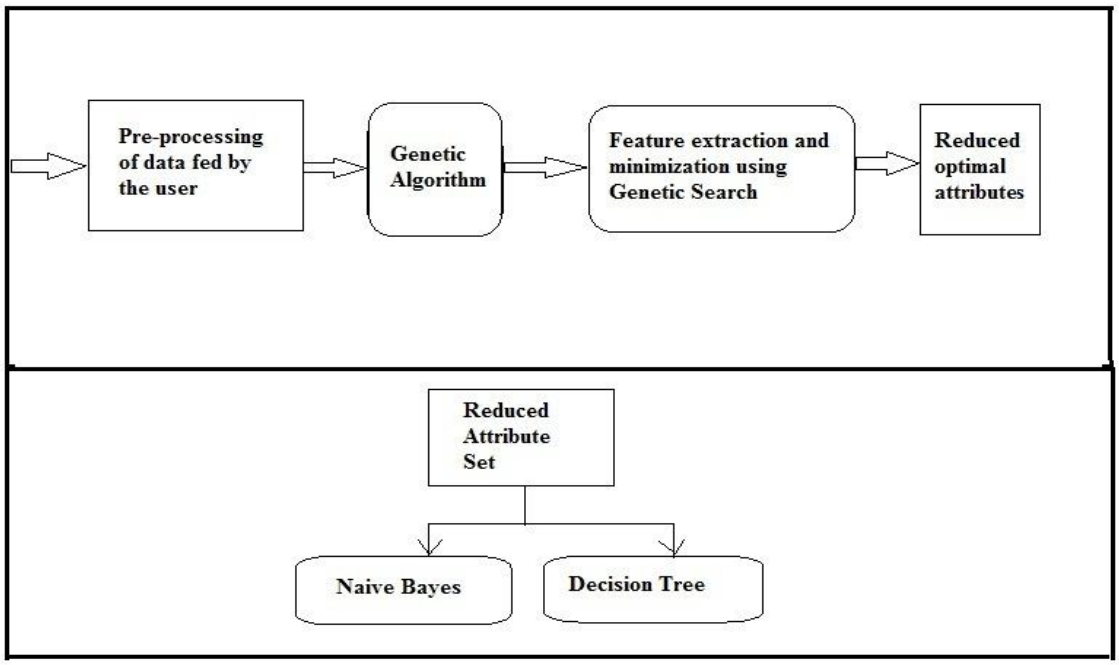


Figure 1: Overview of the system generating reduced optimal attributes and using these generated attributes as an input to the classification models.

A. Dataset

The recordset used in this paper is the Cleveland Heart Disease Database. The attribute to be predicted-"Diagnosis" uses a value "1" for patients with heart disease and value "0" for patients with no heart

disease. The key used is Patient ID. These attributes along with input attributes are shown in Table 1.

Bayes' Rule states that a conditional probability is the likelihood of some conclusion, C, given some evidence / observation, E, where a dependency relationship exists between C and E. The probability is denoted by P(C/E) given by

$$P(C/E) = \frac{P(E/C)\,P(C)}{P(E)}$$

Naïve Bayes classification algorithm works as follows:

1. Let D be a training set of tuples and their associated class labels. Each tuple is represented by a n-dimensional attribute vector, T= (t_1, t_2, ... , t_n), depicting n-measurements made on the tuple from n-attributes, respectively A_1, A_2, A_3, ... , A_n.
2. Suppose that there are m-classes, C_1, C_2, ... , C_m. Given a tuple, T, the classifier will predict that T belongs to the class having the highest posterior probability, conditioned on T. That is the Naïve Bayes' classifier predicts that tuple t belongs to the class C_i if and only if P (C_i/T) > P (C_j/T) for $1 \le j \le m$, $j \ne i$ Thus, we have to maximize P(C_i/T). The class for which P(C_i/T) is maximized is called the maximum posteriori hypothesis. By Bayes' theorem
   P(C_i/T)=P(T/C_i)P(C_i) P(T)As P(T) is constant for all classes, only P(T|C_i)P(C_i) needs to be maximized. If the class' prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, P(C_1)=P(C_2)= ... =P(C_n) and we therefore maximize P(T/C_i). Otherwise, we maximize P(T/Ci)P(Ci)
3. In order to calculate P(T/C_i) , the Naïve assumption of class conditional independence is made. There are no dependence relationships amongst the attributes. Thus,
   $P(T/Ci) = \prod_{k=1}^{n} P(\frac{tk}{Ci})$
   =P(t_1/C_i)*P(t_2/C_i)*.....*P(t_m/C_i)
4. In order to predict the class level of T, P(T/C_i)P(C_i) is evaluated for each class C_i. The classifier predicts that the class label of tuple T is the class C_i iff P(T/C_i)P(C_i) > P(T/C_i)P(C_i) for $1 \le j \le m$, $j \ne i$. The predicted class label of tuple T is the class C_i for which P(T/C_i)P(C_i) is the maximum.

B. Example

Let us consider a tuple T.

T=(age>60, sex=female, cp=2, trestbps>130, Chol>200, Fbs=0, Restecg=2, Thalch>150, Exang=0, Oldpeak<3, Slope=2, Ca=2, Thal=6).

The 13 attributes used are as mentioned in Table 1. The risk levels have been rated as 0, 1, 2, 3 and 4. This is the final outcome that we have to find out.C_1, C_2,...C_5 are the corresponding classes. Thus,

C_1 corresponds to Risk-level 0.

C_2 correspond to Risk-level 1.

C_3 correspond to Risk-level 2.

C_4 correspond to Risk-level 3.

C_5 correspond to Risk-level 4.

We need to maximize P(T/C_i)P(C_i) for i=1,2,3,4,5. The prior probability of each class can be calculated using training tuples.

P(C_1)=164/303=0.54
P(C_2)=55/303=0.18
P(C_3)=36/303=0.118
P(C_4)=35/303=0.115
P(C_5)=13/303=0.043

To compute P(T/C_i), the conditional probabilities have been listed in the Table 4.

P(age>60/C_1)	35/164=0.21
P(age>60/C_2)	14/55=0.25
P(age>60/C_3)	14/36=0.38
P(age>60/C_4)	10/35=0.286
P(age>60/C_5)	6/13=0.46
P(sex=female/C_1)	72/164=0.44
P(sex=female /C_2)	9/55=0.16
P(sex=female /C_3)	7/36=1.99
P(sex=female /C_4)	7/35=0.2
P(sex=female /C_5)	2/13=0.15
P(cp=2/C_1)	0.25
P(cp=2/C_2)	0.11
P(cp=2/C_3)	0.027
P(cp=2/C_4)	0.057
P(cp=2/C_5)	0

Similarly, other probabilities could be calculated. Consider all the values required for P(T/C_1) listed below:

A_1	P(age>60/C_1)	0.21
A_2	P(sex=0/C_1)	0.44
A_3	P(cp=2/C_1)	0.25
A_4	P(trestbps>130/C_1)	0.39
A_5	P(Chol>200/C_1)	0.82
A_6	P(fbs=0/C_1)	0.86
A_7	P(Resteccg=2/C_1)	0.41
A_8	P(thalch>150/C_1)	0.725
A_9	P(Exang=0/C_1)	0.86
A_{10}	P(OldPeak<3/C_1)	0.98
A_{11}	P(slope=2/C_1)	0.29
A_{12}	P(Ca=2/C_1)	0.042
A_{13}	P(thal=6/C_1)	0.036

Now,

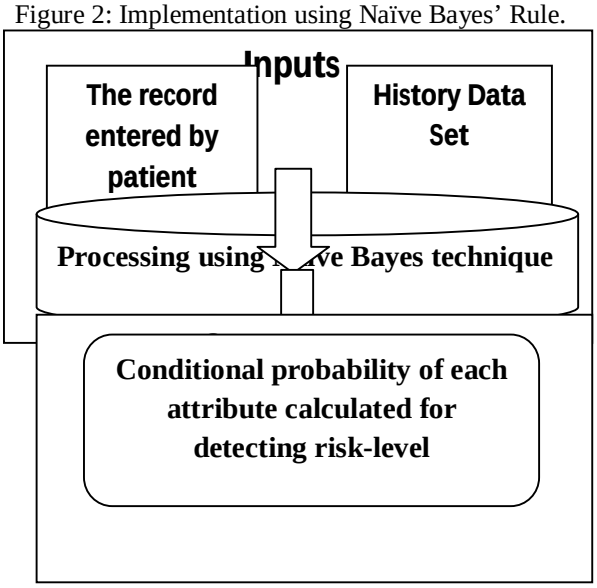
P(T/C_1)=P(A_1/C_1)* P(A_2/C_1) *P(A_3/C_1) *P(A_4/C_1) *P(A_5/C_1) *P(A_6/C_1) *P(A_7/C_1) *P(A_8/C_1) *P(A_9/C_1) *P(A_{10}/C_1) *P(A_{11}/C_1) *P(A_{12}/C_1) *P(A_{13}/C_1)
=0.21*0.44*0.25*0.39*0.82*0.86*0.41*0.725*0.86* 0.98*0.29*0.042*0.036=$6.9*10^{-7}$

P(X/Ci)	**VALUE**
P(X/C_1)	$6.9*10^{-7}$
P(X/C_2)	$8.4*10^{-7}$
P(X/C_3)	$6.6*10^{-8}$
P(X/C_4)	$1.6*10^{-7}$
P(X/C_5)	0

To find a classic C_i, multiply the above mentioned values with respective P(C_i).

$P(T/C_2)P(C_2)$=8.4*10-7*0.18=1.15*10-7
$P(T/C_1)P(C_1)$= $6.9*10^{-7}$* 0.54=$3.7*10^{-7}$
$Max(P(T/C_i)P(C_i))$ =$P(T/C_1)(C_1)$
Thus Naïve Bayes classifier predicts Class C_1 i.e. Risk-level =0 for tuple T.
The model proposed is an effective model to predict risk of heart disease. The Figure 2 describes implementation of Bayes' Rule.

Figure 2: Implementation using Naïve Bayes' Rule.



CONCLUSIONS

Here we have studied both the papers to predict risk of heart disease. In the paper based on classification model, the intensity of risk level of heart disease is not predictable. Improvement is required to increase its consistency and efficiency. The paper based on Bayes' Rule is quite effective but further enhancement can be done in terms of numbers of attributes used. So we plan to develop an Intelligent Heart Disease Decision Support System based on Apriori Algorithm, Genetic Algorithm and Fuzzy Logic to provide an optimal solution to these problems.

REFERENCES

[1] M. Anbarasi and E. Anupriya, "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm", vol.2 no.10, pp. 5370 – 5376, 2010.
[2] G. S. M. Tech, "Decision Support in Heart Disease Prediction System using Naïve Bayes vol.2 no.2, pp. 170 – 176, 2011.
[3] Han, J., Kamber, M.: "Data Mining Concepts and Techniques", Morgan Kaufman Publishers,
[4] http://mlearn.uci.edu/MLSummary. Html
[5] http://en.wikipedia.org/wiki/Genetic_ algorithm

★★★