Prediction of Heart Disease Using Different Classification Techniques

Sonam Nikhar*, A.M. Karandikar
Department of Computer Science & Engineering, Ramdeobaba College of Engineering & Management, Nagpur, India
*Corresponding author, e-mail: sonam.nik3@gmail.com

Abstract

Data mining is one of the essential areas of research that is more popular in health organization. Heart disease is the leading cause of death in the world over the past 10 years. The healthcare industry gathers enormous amount of heart disease data which are not “mined” to discover hidden information for effective decision making. This research intends to provide a detailed description of Naïve Bayes, decision tree classifier and Selective Bayesian classifier that are applied in our research particularly in the prediction of Heart Disease. It is known that Naïve Bayesian classifier (NB) works very well on some domains, and poorly on some. The performance of NB suffers in domains that involve correlated features. C4.5 decision trees, on the other hand, typically perform better than the Naïve Bayesian algorithm on such domains. This paper describes a Selective Bayesian classifier (SBC) that simply uses only those features that C4.5 would use in its decision tree when learning a small example of a training set, a combination of the two different natures of classifiers. Experiments conducted on Cleveland datasets indicate that SBC performs reliably better than NB on all domains, and SBC outperforms C4.5 on this dataset of which C4.5 outperform NB. Some experiment has been conducted to compare the execution of predictive data mining technique on the same dataset, and the consequence reveals that Decision Tree outperforms over Bayesian classifier and experiment also reveals that selective Bayesian classifier has a better accuracy as compared to other classifiers.

Keywords: data mining, naïve bayes classifier, decision tree classifier, selective bayesian classifier

Copyright © 2017 APTIKOM - All rights reserved.

1. Introduction

Data mining is the computer based process of extracting useful information from enormous sets of databases. Data mining is most helpful in an explorative analysis because of nontrivial information from large volumes of evidence. Medical data mining has great potential for exploring the cryptic patterns in the data sets of the clinical domain. These patterns can be utilized for healthcare diagnosis. However, the available raw medical data are widely distributed, voluminous and heterogeneous in nature. These data need to be collected in an organized form. This collected data can be then integrated to form a medical information system. Data mining provides a user-oriented approach to novel and hidden patterns in the data [1-3].

The data mining tools are useful for answering business questions and techniques for predicting the various diseases in the healthcare field. Disease prediction plays a significant role in data mining. This paper analyzes the heart disease predictions using classification algorithms. These invisible patterns can be utilized for health diagnosis in healthcare data. Data mining technology affords an efficient approach to the latest and indefinite patterns in the data [4-7]. The information which is identified can be used by the healthcare administrators to get better services. Heart disease was the most crucial reason for victims in the countries like India, United States [8-9]. Data mining techniques like clustering, Association Rule Mining, Classification algorithms such as Decision Tree [2], C4.5 algorithm, Naïve Bayes [4] are used to explore the different kinds of heart based problems. These algorithms can be used to enhance the data storage for practical and legal purposes.

It has been shown that Naïve Bayesian classifier is extremely effective in practice and difficult to improve upon [8]. In this paper, we show that it is possible to reliably improve this classifier by using a feature selection method. Naïve Bayes can suffer from oversensitivity to redundant and/or irrelevant attributes. If two or more attributes are highly correlated, they receive too much weight in the final decision as to which class an example belongs to. This leads to a decline in accuracy of prediction in domains with correlated features. C4.5 does not suffer from this problem because if two attributes are
correlated, it will not be possible to use both of them to split the training set, since this would lead to exactly the same split, which makes no difference to the existing tree. This is one of the main reasons C4.5 performs better than NB on domains with correlated attributes [10-11]. We conjecture that the performance of NB improves if it uses only those features that C4.5 used in constructing its decision tree. This method of feature selection would also perform well and learn quickly, that is, it would need fewer training examples to reach high classification accuracy.

2. Literature Survey

Numerous works in literature related to the diagnosis of Heart disease using data mining techniques have motivated this work. A brief literature survey is presented here.

A model Intelligent Heart Disease Prediction System built with the assistance of data mining techniques namely, Neural Network, Naïve Bayes, and Decision Tree. Results show that each technique has its infrequent strength in realizing the objectives of the defined mining goals. IHDPS can answer complex “what if” queries which conventional decision support systems cannot be proposed by Sellappan Palaniappan et al. [2]. The results illustrated the uncouth strength of each of the methodologies in comprehending the goal of the specified mining objectives. IHDPS was capable of responding queries that the traditional decision support systems were not able to. It facilitated the installation of crucial knowledge such as patterns, relationships amid medical factors connected with heart disease. IHDPS remains well-being web-based, user-friendly, reliable, scalable and expandable.

The diagnosis of Heart Disease, Blood Pressure and diabetes with the aid of neural networks was introduced by Niti Guru et al. [7]. Experiments were carried out on a sampled data set of patient’s records. The Neural Network is trained and tested with 13 input variables such as Blood Pressure, Age, Angiography's report and the like. The supervised network has been advised for diagnosis of heart diseases. Training was carried out with the help of back propagation algorithm. Whenever unfamiliar data was inserted by the doctor, the system identified the unknown data from comparisons with the trained data and produced a catalog of probable diseases that the patient is vulnerable to.

In 2014, M.A.Nishara BanuB.Gomathy Professor, Department of Computer Science and Engineering has published a research paper “Disease Forecasting System Using Data Mining Methods”[8].In this article, the pre-processed data is clustered using clustering algorithms as K-means to gather relevant data in a database. Maximal Frequent Item set Algorithm (MAFIA) is applied for mining maximal frequent model in heart disease database. The regular patterns can be classified into different classes using the C4.5 algorithm as training algorithm using the concept of information entropy. The result demonstrates that the designed prediction system is capable of predicting the heart attack successfully.

In 2012, T.John Peter and K. Somasundaram Professor, Dept of CSE presented a paper, “An Empirical Study on Prediction of Heart Disease using classification data mining technique” [5]. In this research paper, the use of pattern recognition and data mining techniques are used for prediction of risk in the medical domain of heart disease medicine is proposed here. Some of the limitations of the traditional medical scoring systems are that there is a presence of intrinsic linear combinations of variables in the input set, and hence they are not skilled at modeling nonlinear complex interactions in medical domains. This limitation is handled in this research by use of classification models which can implicitly detect complex nonlinear relationships between independent and dependent variables as well as the ability to identify all possible interactions between predictor variables.

In 2013, Shamsher Bahadur Patel, Pramod Kumar Yadav, and Dr. D. P.Shukla presented a research paper,“Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques “[6].In this research paper, the health care industry, the data mining is mainly utilized for the prediction of heart disease. The objective of our works to predict the diagnosis of heart disease with a reduced number of attributes using Naïve Bayes, Decision Tree.

3. Proposed Architecture

The working of this system is described in a step by step as shown in Figure 1.

1. Dataset collection which contains patient details.
2. Attributes selection process selects the useful attributes for the prediction of heart disease.
3. After identifying the available data resources, they are further selected, cleaned, made into the desired form.
4. Different classification techniques as stated will be applied on preprocessed data to predict the accuracy of heart disease.
5. Accuracy measure compares the accuracy of different classifiers.

![Figure 1. Proposed Architecture](image)

4. Research Methodology
   In this section, we are introducing methods for a new proposed system.

4.1. Naïve Bayesian Classifier
   In data mining, naïve Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong independence assumptions between the features. Naïve Bayes classifiers are intensely scalable, requiring some parameters linear in the number of variables (features/predictors) in a learning problem.

   The Bayesian Classification depicts a supervised learning method as well as a statistical method for classification. Assumes a fundamental probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems.

4.1.1. Bayes Rule
   A conditional probability is the likelihood of some conclusion, C, given some evidence/observation, E, where a dependence relationship exists between C and E. This probability is denoted as P(C |E) where

   \[
   P(C|E) = \frac{P(E|C)P(C)}{P(E)}
   \]

4.1.2. Naïve Bayesian Classification Algorithm
   The naïve Bayesian classifier, or simple Bayesian classifier, works as follows as shown in Figure 2.
   1. Let D be a training set of tuples and their associated class labels. As usual, each row is represented by an n-dimensional attribute vector, X=(x1, x2,…, xn), depicting n measurements made on the tuple from n attributes, respectively, A1, A2,…, An.
   2. Suppose that there are m classes, C1, C2,…, Cm. Given a tuple, X, the classifier will foreshow that X belongs to the class having the highest posterior probability, conditioned on X. That is; the naïve Bayesian classifier predicts that tuple x belongs to the class Ci if and only if;

   \[
   P(C_i|X) > P(C_j|X) \text{ for } 1 \leq j \leq m, j \neq i
   \]

   Thus, we maximize P(Ci|X). The class Ci for which P(Ci|X) is maximized is called the maximum posterior hypothesis. By Bayes’ theorem

   \[
   P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}
   \]
3. As \( P(X) \) is constant for all classes, only \( P(X|C_i) \) \( P(C_i) \) need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, \( P(C_1)=P(C_2)=...=P(C_m) \), and we would, therefore, maximize \( P(X|C_i) \). Otherwise, we maximize \( P(X|C_i)P(C_i) \). Note that the class prior probabilities may be estimated by \( P(C_i)=|C_i,D|/|D| \), where \( |C_i,D| \) is the number of training tuples of class \( C_i \) in \( D \).

4. Given data sets with many attributes, it would be extremely computationally precious to compute \( P(X|C_i) \). To decrease computation in evaluating \( P(X|C_i) \), the naïve assumption of class conditional independence is made. This reckons that the values of the attributes are conditionally independent of one another, given the class label of the tuple (i.e., that there are no dependence relationships among the attributes). Thus,

\[
P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i)
\]

\(=P(x_1|C_i) \times P(x_2|C_i) \times \ldots \times P(x_m|C_i).\)

We can easily estimate the probabilities \( P(x_1|C_i), P(x_2|C_i), \ldots, P(x_m|C_i) \) from the training tuples. Recall that here \( x_k \) refers to the value of attribute \( A_k \) for tuple \( X \).

5. To predict the class label of \( X \), \( P(X|C_i)P(C_i) \) is evaluated for each class \( C_i \). The classifier predicts that the class label of tuple \( X \) is the class \( C_i \) if and only if

\[
P(X|C_i)P(C_i) > P(X|C_j)P(C_j) \text{ for } 1 \leq j \leq m, j \neq i
\]

In other words, the predicted class label is the class \( C_i \) for which \( P(X|C_i)P(C_i) \) is the maximum.

![Figure 2. Implementation of Naïve Bayes algorithm on the patient data.](image)

### 4.2. Decision Tree

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible outcomes, comprising chance event outcomes, resource costs, and utility. It is one way to display an algorithm. Decision trees are generally used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach an objective, but are also a popular tool for machine learning.

A decision tree is a flowchart-like structure in which each internal node depicts a "test" on an attribute, each branch represents the outcome of the test and each leaf node accounts for a class label. The path from the root to leaf depicts classification rules.

The basic algorithm for decision tree induction is a greedy algorithm that builds decision trees in a top-down recursive divide-and-conquer manner. The algorithm starts with the entire set of rows in the Training set, selects the best attribute that yields maximum information for classification, and originates a test node for this attribute. Then, top-down induction of decision trees divides the current set of tuples according to their values of the current test attribute. Classifier generation stops, if all tuples in a subset pertain to the same class, or if it is not worth to proceed with an additional separation into further subsets, i.e. if further attribute tests produce only information for classification below a pre-specified threshold.
The decision tree algorithm commonly uses an entropy-based measure known as “information gain” as a heuristic for selecting the attribute that will best split the training data into separate classes. The algorithm computes the information gain of each attribute, and in each round, the one with the highest information gain will be chosen as the test attribute for the given set of training data. A well-chosen split point should help in dividing the data to the best possible limit. After all, a primary criterion in the greedy decision tree approach is to build shorter trees. The best split point can be quickly evaluated by considering each unique value for that feature in the given data as a possible split point and calculating the associated information gain.

4.2.1. Information Gain

The critical step in decision trees is the selection of the best test attribute. The information gain measure is used to select the test attribute at each node in the tree.

First, another related term called entropy needs to be introduced. In general, entropy is a measure of the purity in an arbitrary collection of examples. Let $S$ be a set consisting of s data samples. Suppose the class label attribute has m distinct values defining m different categories, $C_k$. Let $s_i$ be the number of samples of $S$ in class $C_k$. The expected information needed to classify a given sample is provided by:

$$I(S_1, S_2, \ldots, S_m) = - \sum_{k=1}^{m} p_k \log_2 (p_k)$$

Where, $p_k$ is the probability that an arbitrary sample belongs to class $C_k$ and is estimated by $s_k / s$. Let attribute A have v distinct values, {a1,a2,…,av}. Attribute A can be used to partition $S$ into v subsets, {S1,S2,…,Sv}, where $S_j$ contains those samples in $S$ that have value aj of A. Let $s_{kj}$ be the number of samples of class $C_k$ in a subset $S_j$. The entropy, or expected information based on the partitioning into subsets by A, is given by:

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \cdots + s_{mj}}{s} I(s_{1j}, \ldots, s_{mj})$$

The term $\frac{s_{1j} + \cdots + s_{mj}}{s}$ acts as the weight of the j-th subset and is the number of samples in the subset divided by the total number of samples in S. For a given subset $S_j$:

$$I(S_{1j}, S_{2j}, \ldots, S_{mj}) = - \sum_{k=1}^{m} p_{kj} \log_2 (p_{kj})$$

Where; $p_{kj} = \frac{s_{kj}}{|S_j|}$ and is the probability that a sample in $S_j$ belongs to class $C_k$. The entropy is zero when the sample is pure, i.e. when all the examples in the sample S belong to one class. Entropy has a maximum value of 1 when the sample is maximally impure, i.e. there are same proportions of positive and negative examples in the sample S.

The encoding information would be gained by branching on A is:

$$\text{Information Gain}(A) = I(s_1, s_2, \ldots, s_m) - E(A)$$

The attribute with the highest information gain is chosen as the test attribute for the current node. Such approach minimizes the expected number of tests needed to classify an object and guarantees that a simple (but may not be the simplest) tree is found as shown in Figure 3.
4.3. Selective Bayesian Classifier

Our purpose is to improve the performance of the Naïve Bayesian classifier by removing redundant and/or irrelevant attributes from the dataset, and only choosing those that are most informative in classification task. To achieve this, we use the trees that are constructed by C4.5.

4.3.1. Description

As described in section 3, the features that C4.5 selected in constructing its decision tree are likely to be the ones that are most descriptive in terms of the classifier, in spite of the fact that a tree structure inherently incorporates dependencies among attributes, while Naïve Bayes works on a conditional independence assumption. C4.5 will naturally construct a tree that does not have an overly complicated branching structure if it does not have too many examples that need to be learned. As the number of training examples increases, the attributes that are considered will usually be the ones that are not correlated.

This is mainly because C4.5 will use only one of a set of correlated features for making good splits in training set. However, sometimes many of the branches may reflect noise or outliers (overfitting) in the training data. “Tree pruning” procedure in C4.5 attempts to identify and remove those least reliable branches, with the goal of improving classification accuracy on unseen data. Even after pruning, if the resulted decision tree is still too deep or grown into too many levels, our algorithm only picks attributes contained in the first few levels of the tree as the most representative attributes. This is supported by the fact that by the selection of attributes that split the data in the best possible way at every node, C4.5 will try to ensure that it encounters a leaf at the very earliest possible point, i.e. it prefers to construct shorter trees. And by its algorithm, C4.5 will find trees that have attributes with higher information gain nearer to the root. We conjecture that this simple method of feature selection would help improve Naïve Bayesian classifier’s performance and learn quickly, that is, it would need fewer training examples to reach high classification accuracy.

4.3.2. Algorithm

Figure 1 shows the algorithm for the Selective Bayesian classifier.

1. Take a 10% sample of training data.
2. Run C4.5 on data from step 1.
3. Select a set of attributes that appear only in the first 3 levels of the simplified decision tree as relevant features.
4. Repeat 5 times (step 1-3)
5. Form a union of all the attributes from the 5 rounds.
6. Run Naïve Bayesian classifier on the training and test data using only the final features selected in step

Figure 1. Selective Bayesian Classifier Algorithm: Feature Selection Using C4.5
5. EXPERIMENTAL EVALUATION

5.1. Data Sets

Clinical databases have collected a significant amount of information about patients and their medical conditions. The term Heart disease encompasses the diverse conditions that affect the heart.

Records set with medical attributes were obtained from the Cleveland Heart Disease database. With the help of the dataset, the patterns significant to the heart attack diagnosis are extracted. The records were split equally into two datasets: training dataset and testing dataset.

A total of 303 records with 76 medical attribute were obtained. All attributes are numeric-valued. We are working on a reduced set of attributes, i.e. only 19 attributes. The following table shows the list of attributes on which we are working.

<table>
<thead>
<tr>
<th>No.</th>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Id</td>
<td>Identification number.</td>
</tr>
<tr>
<td>2</td>
<td>Age</td>
<td>It will take age in years as input.</td>
</tr>
<tr>
<td>3</td>
<td>Sex</td>
<td>It will take two values as input i.e. value 1: Male and value 0: Female.</td>
</tr>
<tr>
<td>4</td>
<td>Chest Pain</td>
<td>Value 1: Yes, Value 0: No.</td>
</tr>
<tr>
<td>5</td>
<td>Chest Pain Type</td>
<td>It will take four value as input which shows the chest pain type as value 1: typical type-1 angina, value 2: typical type angina, value 3: non-angina pain; value 4: asymptomatic.</td>
</tr>
<tr>
<td>6</td>
<td>Trest Blood Pressure</td>
<td>resting blood pressure (in mm Hg on admission to the hospital)</td>
</tr>
<tr>
<td>7</td>
<td>chol</td>
<td>serum cholesterol in mg/dl.</td>
</tr>
<tr>
<td>8</td>
<td>Fasting Blood Sugar</td>
<td>It will take two values as input i.e. value 1 for FBS &gt; 120 mg/dl and value 0: for FBS &lt; 120 mg/dl.</td>
</tr>
<tr>
<td>9</td>
<td>Restecg</td>
<td>Resting electrographic results will take 3 value as input i.e. value 0: normal; value 1: 1 having ST-T wave abnormality; value 2: showing probable or definite left ventricular hypertrophy.</td>
</tr>
<tr>
<td>10</td>
<td>thalach</td>
<td>Maximum heart rate achieved by the patient.</td>
</tr>
<tr>
<td>11</td>
<td>Exang</td>
<td>Exercise induced angina will take two values as input i.e. value 1: yes and value 0: no.</td>
</tr>
<tr>
<td>12</td>
<td>Old peak</td>
<td>ST depression induced by exercise relative to rest.</td>
</tr>
<tr>
<td>13</td>
<td>Slope</td>
<td>The slope of the peak exercise ST segment will take three values as input i.e. value 1: unsloping; value 2: flat and value 3: down sloping.</td>
</tr>
<tr>
<td>14</td>
<td>CA</td>
<td>Number of major vessels colored by fluoroscopy will take three value as input (value 0 – 3)</td>
</tr>
<tr>
<td>15</td>
<td>Thal</td>
<td>It will take 3 input values i.e. value 3: normal; value 6: fixed defect and value 7: reversible defect.</td>
</tr>
<tr>
<td>16</td>
<td>Smoke</td>
<td>I believe this is 1=Yes,0=No(is or is not smoker)</td>
</tr>
<tr>
<td>17</td>
<td>Dm</td>
<td>Value 1: History of diabetes; Value 0: No such history of diabetes.</td>
</tr>
<tr>
<td>18</td>
<td>Famhist</td>
<td>Family history of coronary artery disease (1=Yes;0=No)</td>
</tr>
<tr>
<td>19</td>
<td>Num</td>
<td>Diagnosis of heart disease</td>
</tr>
<tr>
<td></td>
<td>Value 0: No Risk; Value 1: Low Risk</td>
<td>Value 2: Risk, Value 3: High risk; Value 4: Higher Risk</td>
</tr>
</tbody>
</table>

5.2. Data Preprocessing

After identifying the available data resources, they are further selected, cleaned, made into the desired form, and also being formatted.

5.3. Experimental Design

The dataset is divided into training and testing dataset.

- Produce disjoint training and test sets as follows.
  - 80% training and 20% test data.
- For each set of training and test data, run.
  - Naïve Bayesian Classifier (NBC)
  - Decision Tree (DT)
  - Selective Bayesian Classifier (SBC)
- Compare the accuracy of each classifier.

The classifier accuracy is determined by percentage calculation method, i.e. the number of correctly classified fields divided by total number of fields.

5.4. Experimental results evaluation

From these results as shown Table 1, it is concluded that although most researchers are using different classifier techniques such as Neural network, SVM, KNN and binary discretization with Gain Ratio Decision Tree in the diagnosis of heart disease, applying Naïve Bayes and Decision tree with
information gain calculations, Selective Bayesian classifier provides better results in the diagnosis of heart disease and provides better accuracy as compared to other classifiers. We surmise that the improvement in accuracy arises from the selective Bayesian classifier with reduced attributes. We have also observed that decision tree outperforms over Naïve Bayes and Selective Bayesian classifier outperforms over decision tree. To increase the accuracy we have used selective Bayesian classifier which is nothing but a combination of both the classifier.

<table>
<thead>
<tr>
<th>Classification Techniques</th>
<th>Accuracy with 18 Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes Classifier</td>
<td>82%</td>
</tr>
<tr>
<td>Decision Tree Classifier</td>
<td>92%</td>
</tr>
<tr>
<td>Selective Bayesian Classifier</td>
<td>97%</td>
</tr>
</tbody>
</table>

6. Result and Analysis

Out of 61 testing fields, Naïve Bayes Classifier was able to correctly classify 50 fields correctly; Decision tree Classifier was able to correctly classify 55 fields. Selective Bayesian classifier with Feature Selection based on Entropy Method was able to correctly classify 56 fields. Figure 4 shows the number of correctly classified fields using all the four classifiers:

![Figure 4. Graphical Representation of Results](image)

Accuracy is calculated using the following formula:

\[
\text{Accuracy} = \frac{\text{Total Correctly Classified fields}}{\text{Total fields}} \times 100
\]

7. Conclusion

In this paper, we introduce about the heart disease prediction system with different classifier techniques for the prediction of heart disease. The techniques are Naïve Bayes classifier, decision tree classifier and Selective Bayesian classifier; we have analyzed that the decision tree has better accuracy as compared to Naïve Bayes classifier and Selective Bayesian classifier has better accuracy against both the classifier. This work suggests that C4.5 decision trees systematically select good features for Naïve Bayesian classifier to use. We believe the reasons are that C4.5 does not use redundant attributes in constructing decision trees, since they cannot generate different splits of training data. When few training examples are available, C4.5 uses the most relevant features it can find. The high accuracy SBC achieves with few training examples is indicative of the fact that using these features for probabilistic induction leads to higher accuracy both in Bayesian classifier and C4.5 itself in each of the domains we have examined.

References


