A Comparative Study of Utilization of Single and Hybrid Data Mining Techniques in Heart Disease Diagnosis and Treatment Plan

Rajesh Jagtap
M.E.Scholar
Department of Computer Engg.
DPCOE, Wagholi, Pune

ABSTRACT

In clinical medicine, data mining deals with learning models to predict patient's health. The models can be dedicated to support clinicians in diagnostic and monitoring tasks. Data mining methods are commonly applied in clinical contexts to analyze retrospective data, thus giving healthcare professionals the opportunity to exploit enormous amounts of data routinely collected during their day-by-day activity. Nowadays, clinicians can take advantage of data mining techniques to deal with the huge amount of research results obtained by molecular medicine such as genomic signatures or genetic which may allow transition from population-based to personalized medicine. The different classification and prediction models can be devoted to support medical practitioners in diagnosis and formation of treatment plans. There is need of powerful data analysis tool to extract useful knowledge from huge amount of data available in health care field. Last few years, heart disease is the major cause of death all over the world. In heart disease diagnosis and treatment, single data mining techniques are showing satisfactory level of accuracy. Nowadays, researchers are experimenting the deployment of hybrid data mining techniques showing great level of accuracy. In this paper, single data mining techniques like Naive base, Decision tree, Association rule, Neural network and Regression are studied and compared with hybrid data mining algorithm to achieve an efficient results in heart disease diagnosis and to formulate treatment plan.

General Terms
Healthcare, Diagnosis

Keywords
Hybrid, Data Mining, Genomic, Retrospective, Monitoring

1. INTRODUCTION

The heart is vital organ of human body which pumps blood through the body. Due to inefficient circulation of blood in body organs like brain, suffer and if heart stops working, death occurs within minutes. Human body operation is totally dependent on efficient working of the heart. Heart disease term refers to disease of heart and blood vessel system within it and includes the diverse diseases that affect the heart. Coronary heart disease, cardiovascular disease and Cardiomyopathy are some categories of heart diseases. Today, in the world, Heart disease is the major cause of deaths. The World Health Organization (WHO) has estimated that 13 million deaths occur worldwide, every year due to the cardiac diseases. WHO estimated by 2040, almost 24.6 million people will suffer due to cardiac disease. [1]. The following are number of factors that increases the risk of Heart disease [2]:

(1) Hyper tension
(2) obesity
(3) Smoking
(4) Poor diet
(5) High blood pressure
(6) Physical inactivity
(7) High cholesterol
(8) Family history

The Diagnosis of patient is generally based on symptoms, signs, and physical examination and related to the classification of patients into disease classes or sub-classes on the basis of patient’s data. This activity covers a broad spectrum of clinical cases, including triage at hospital emergency departments, i.e, prioritizing patients based on the severity of their condition. Practically all the doctors are diagnosing heart disease by knowledge and experience. The diagnosis of disease is a challenging and tedious task in medical field. Predicting cardiac disease from numerous
To find out the accuracy of single data mining techniques and explain the data to be interpreted by humans. The current research aims to predict the possibility of getting heart disease according to dataset of patient’s medical record. In practice, Predictions and descriptions are primary goals of data mining [6]. Prediction in data mining comprises attributes or variables in the dataset to detect an unknown or future state values of other attributes [7]. Description highlight on discovering patterns that explains the data to be interpreted by humans.

2. OBJECTIVES OF THE RESEARCH

(1) To find out the accuracy of single data mining techniques and compare it with the accuracy of hybrid data mining techniques to diagnose the heart disease.

(2) To exploit the usefulness of hybridized data mining techniques in heart disease diagnosis and discover the suitable treatment for heart disease patients.

(3) To highlight the importance of computer algorithms in medical applications.

(4) To help medical practitioners to reduce the errors and complexity in diagnosis process and to improve the relationship between patient and medical practitioner.

3. RELATED WORK

Several studies have been done that have attention on diagnosis of heart disease. Researchers have applied different data mining techniques for diagnosis and achieved dissimilar probabilities for different methods.

Data Mining Techniques used in Heart disease diagnosis:
Mai Shouman, Tim Turner, Rob Stocker et al.[1] have shown the comparison of single and hybrid data mining techniques in the diagnosis of heart disease on the CHDD (Cleveland Heart Disease Dataset). The baseline accuracy got in diagnosis using single data mining technique is compared with baseline accuracy got in treatment and The baseline accuracy got in diagnosis using hybrid data mining technique is compared with baseline accuracy got in treatment. These techniques shown different accuracies, where hybrid techniques were more accurate than single techniques accuracy. The best accuracy achieved using single data mining technique was 84.14% by naive bayes. However, the best accuracy achieved using hybrid data mining technique was 89.01% by neural network ensemble. Finally, they observed that hybrid data mining techniques are more accurate and enhanced the accuracy of heart disease diagnosis.

Supervised and Unsupervised learning method:
Mary K. Obenshain et al. [2] focused on data mining product-SAS Enterprise Miner, which often included in data mining application suites for specific application areas such as customer relationship management (CRM), financial management. They have also highlighted the importance of supervised and unsupervised learning methods. Supervised learning methods depend on known values of variables (inputs) are used to make predictions about another variable (target) with known values. The supervised learning methods are used to make prediction about fraudulent claims using different attributes in healthcare organizations. In supervised methods, the models and attributes are known and are applied to the data to predict and find information. Unsupervised learning methods applicable almost in same situations, but are more regularly deployed on data for which a target with known values does not exist.

Machine learning algorithms for data mining tasks:
Rajkumar, A. and G.S. Reena et al. [3] proposed the utilization of Tanagra software used to compare the performance accuracy of data mining algorithms for diagnosis of heart disease dataset. The feature selection in the Tanagra software defined the attribute status of the data present in the heart disease. The authors have compared different supervised machine learning algorithms such as Naive Bayes, k-nn and Decision list. For data mining tasks, Tanagra was proved to be a successful tool which contains collection of machine learning algorithms. In their research, Naive Bayes algorithm shown the best compact time for processing dataset and better performance in accuracy prediction. The time required to run the data for result is faster when compared to other algorithms. It illustrated the enhanced performance according to input attribute. The attributes are entirely classified by this algorithm and it gave 52.33% of accurate result. According to the experimental results the classification accuracy is found to be better using Naive Bayes algorithm as compare to other algorithms. So, it was found that Naive Bayes algorithm plays a crucial role in shaping improved classification accuracy of a dataset.

Verification of clinical data using SOAP:
Razali and Ali et al. [4] surveyed the making of treatment plans for critical upper respiratory infection disease patients using a decision tree. Their study focused on outpatient and was based on data collected from various health centers throughout Malaysia. They have verified clinical data using SOAP (Subjective, Objective, Assessment and Plan) format approach as being practiced in medicine and were recorded electronically via Percuro Clinical Information System (Percuro). Cross-Industry Standard Process for Data Mining (CRISP-DM) model has been applied for the entire research. The data mining analysis is completed through decision trees technique with CS algorithm. The scopes that have been set are patients complaint, age, gender, type of plan and detailed item given to patient. The suggested treatment model gave 94.73% accuracy through giving drugs to patients. The association rules and decision tree technique to treatment plans were shown satisfactory level of performance. They also found that the comparison of decision tree technique with other data mining techniques such as genetic algorithms, naive bayes, and neural network still needs further investigation.

Utilization of decision tree:
Kim et al.[6] evaluated the recent treatments for chronic heart failure using a decision tree and compared the results with those of large-scale clinical trials. They explored the procedures which recommended prescriptions of drugs to increase or decrease plasma level, spontaneous hypertension, fractional shortening and left ventricular end-diastolic diameter in the cardiovascular disease. However, they were unsuccessful to inspect exact data.
mining techniques to identify the suitable treatment for heart disease patients.

Classification based data mining:
Srinivas, K., B.K. Rani, and A. Govrdhan et al. [7] briefly examined the probable use of classification based data mining techniques such as Decision tree, Naive Bayes and Rule based Artificial Neural Network to huge volume of healthcare data. They have provided a well-organized approach for the extraction of significant patterns from the heart disease data warehouses for the efficient prediction of heart attack based on the calculated significant weightage, the frequent patterns having value superior than a predefined threshold were chosen for the valuable prediction of heart attack. For data preprocessing and effective decision making One Dependency Augmented Naive Bayes classifier (ODANB) and naive credal classifier 2 (NCC2) were utilized. That was an extension of naive Bayes to imprecise probabilities that aims at delivering robust classifier 2 (NCC2) were utilized. That was an extension of naive Bayes model detects the characteristics of patients with heart disease. It shows the probability of each input attribute for the predictable state. It uses Bay’s theorem. 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

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

Where,

- \( P(C) \) = Prior probability of hypothesis C
- \( P(E) \) = Prior probability of training data E
- \( P(C|E) \) = Prior probability of C given E
- \( P(E|C) \) = Prior probability of E given C

(2) Neural Network
A neural network (NN) is a parallel, distributed information processing structure consisting of multiple numbers of processing elements called node, they are interconnected via unidirectional signal channels called connections. Each processing element has a single output connection that branches into many connections; each carries the same signal i.e. the processing element output signal.

The NN can be classified in two main groups according to the way they learn:

(a) Supervised learning

\[
x_0 \quad w_0 \quad \text{synapse} \quad w_0x_0 + b \quad \text{activation function} \quad f \left( \sum_{i} w_i x_i + b \right)
\]

Fig. 2. Structure of Multi Layer Perceptron Neural Network, et al. [14]

It is a simple model, in which the networks compute a response to each input and then compare it with target value. If the computed response differs from target value, the weights of the network are adapted according to a learning rule, e.g., Multilayer Perceptron.

(b) Unsupervised learning
These networks learn by identifying special features in the problems they are exposed to, e.g.: Self-organizing feature maps.

(3) Decision Tree
Decision Tree techniques has shown useful accuracy in the diagnosis of heart disease. A decision tree is generated using C4.5 algorithm. It can be used for classification, which builds decision trees from a set of training data, using the concept of information entropy [11]. The training data is a set \( T = \{t_1, t_2, t_3, \ldots \} \) of already classified samples. Each sample \( S_i \) consists of a \( p \)-dimensional vector \( (x_1, i, x_2, i, x_3, i, \ldots, x_p, i) \) where the \( x_j \) represent attributes or features of the sample, as well as the class in which \( S_i \) falls. At each node of the tree, C4.5 selects the attribute of the data that most efficiently splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the maximum normalized information gain is selected to make the decision. The algorithm C4.5 then repeats on the smaller sublists. The typical algorithm for building decision trees is:

(a) Verify all base cases
(b) For each attribute at
   i. Calculate the normalized information gain ratio from splitting on at
(c) Let \( at, best \) be the attribute with the maximum normalized Information gain
(d) Create a decision node that splits on \( at, best \)
(e) Repeat on the sublists obtained by splitting on \( at, best \) and add those nodes as children of node

(4) Linear Regression
Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. It is a data mining function that predicts a number. A regression procedure begins with a data set in which the target values are known [10]. In the regression model relationships between predictors and target are summarized in a model,
which can then be applied to a different data set in which the
target values are unknown. A regression is a statistical analysis
assessing the association between two variables. It is used to
find the relationship between two variables. The relationship
takes the form of an equation for a line that best indicates a
series of data [15]. For example, the line shown in the
figure 3 is the best possible linear representation of the data.
Each data point in the figure has an error associated with its
distance from the regression line. The coefficients a and b in the
regression equation adjust the angle and location of the regression line.
For obtaining the regression equation, the adjustment of a and
b until the sum of the errors that are associated with all the
points reaches its minimum.

5. ASSOCIATION RULE
Association and correlation is usually to find frequent item set
findings among large data sets. Association rule learning is a
well-researched method for discovering interesting relations
between variables in large databases [7]. It is intended to
determine strong rules discovered in databases using different
measures of interestingness. This type of finding helps
businesses to make certain decisions, such as catalog design,
cross marketing and customer shopping behavior analysis.
Association Rule algorithms need to be able to generate rules
with confidence values less than one. However the number
of possible Association Rules for a given dataset is generally
very large and a high proportion of the rules are usually of
little (if any) value.

- Mathematical Model of Association Rule
The problem of association rule mining is defined as: Let
I = \{i_1, i_2, ..., i_n\} be a set of n binary attributes called
items. Let D = \{t_1, t_2, ..., t_m\} be a set of transactions called
the database. Each transaction in D has unique transaction
ID and contains a subset of the items in I. A rule is defined
as an implication of the form X → Y where X, Y ⊆ I
and X ∩ Y = φ. The sets of items (for short item sets) X
and Y and are called antecedent (left-hand-side or LHS) and
consequent (right-hand-side or RHS) of the rule respectively.

6. HYBRID DATA MINING ALGORITHM
Accurate diagnosis and treatment given to patients have been major
issues emphasized in medical services. In recent times, research
is ongoing for investigating data mining techniques to handle the
error and complexity of treatment processes for healthcare service
providers.

The concept of hybridization of data mining techniques is
implemented using following algorithm:

1. Let M be a set of single data mining techniques,
   M={x1, x2, x3, x4, x5} and S=Total support cases available in the
dataset used to calculate the baseline accuracy or probability of
Heart disease diagnosis and to decide treatment.

2. Get the input attributes A1, A2, ..., A13 from patient’s Medical
   Report and Calculate the baseline accuracy or probability by
   applying single data mining techniques to diagnosis dataset D,
   where xεM.

The equation to find baseline accuracy or probability is:

\[ A = \frac{n(S)}{S} \times 100 \quad Where, \ S > n(S) \] (1)
Where, A = Baseline Accuracy or probability,
\( n(S) \) = Selected support cases Or Patient log
S = Total support cases

1. Calculate the output \( O \) of each single data mining technique
   which will come in binary values “0” and “1”. Where,
   \( O = 0 \) (Zero) = Patient does not exist heart disease.
   \( O = 1 \) = Patient does exit heart disease.

2. Calculate the baseline accuracy or probability by applying
   Hybrid (H) data mining techniques to diagnosis dataset D
   using following equation:
   \[ H = (O_1 + O_2 + O_3 + O_4 + O_5) \]  

   Where,
   \( O_1 = \) Output of Naive Base algorithm
   \( O_2 = \) Output of Decision Tree algorithm
   \( O_3 = \) Output of Neural Network algorithm
   \( O_4 = \) Output of Association Rule algorithm
   \( O_5 = \) Output of Linear Regression algorithm

3. Compare the output of Hybrid data mining techniques H with
   Threshold Value (3.0) output of Single data mining techniques M.
   i) Check if \( H = (O_1 + O_2 + O_3 + O_4 + O_5) \geq 3.0 \)
   ii) If ’YES’ goto step 6 and 8, Else goto step 7.

4. Patient does exist heart disease.
5. Patient does not exist heart disease.

7. PATIENT MEDICAL DATASET AND RESULT

   The dataset used in this study is the Cleveland Clinic Foundation Heart disease data set available at
   http://archive.ics.uci.edu/ml/datasets/Heart+Disease. For experiment 13 attributes are considered:

   **Diagnosis:**
   - value 0: \( \leq 50\% \) diameter narrowing (no heart disease);
   - value 1: \( \geq 50\% \) diameter narrowing (has heart disease)

   **Key attribute:**
   - PatientID (Patient’s identification number)
   - Sex (value 1: Male; value 0 : Female)
   - Chest Pain Type (value 1: typical type 1 angina, value 2: typical type angina, value 3: non-angina pain; value 4: asymptomatic)
   - Fasting Blood Sugar (value 1: \( \geq 120 \) mg/dl; value 0: \( \leq 120 \) mg/dl)
   - Restecg - resting electrographic results (value 0: normal; value 1: 1 having ST-T wave abnormality; value 2: showing probable or definite left ventricular hypertrophy)
   - Exang - exercise induced angina (value 1: yes; value 0: no)
   - Slope - the slope of the peak exercise ST segment (value 1: unsloping; value 2: flat; value 3: downsloping)
   - CA - number of major vessels colored by floursopy (value 0 - 3)
   - Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)
   - Trest Blood Pressure (mm Hg on admission to the hospital)
   - Serum Cholesterol (mg/dl)
   - Thalach - maximum heart rate achieved
   - Oldpeak - ST depression induced by exercise relative to rest
   - Age in Year

8. CONCLUSION

   A single mistake in diagnosis leads to incorrect treatment and ultimately patient’s life would be in trouble. The “Trust” is very vital factor between doctor and patient. The different classification and prediction single data mining algorithms namely Naive Base, Decision Tree, Neural network, Association Rule and Linear Regression are implemented. Each algorithm contains certain functions which are helpful to diagnose the heart disease. For perfect analysis of heart disease, the outputs of each algorithm is combined and compared with threshold value 3.0. Here combination of output is considered as “Hybridization”. If the addition of that outputs are greater than 3.0 then the presence of heart disease is finalized and accordingly treatment is recommended.

   Due to advance computer data mining techniques like hybrid data mining technique, doctors are quite relax to treat any patient and especially in case of heart disease, the hybrid technique carried drastic change. All algorithms showing diverse output due to that it is difficult to conclude which one is better and perfect for diagnosis. The accuracy of every algorithm is calculated by analyzing historical dataset. The prediction of heart disease is done using binary values that is “0” and “1”. If the patient does exist heart disease then output shown by web application is “1” and if patient does not exist heart disease then web application shows the output “0”.

   In Future, Additional data mining techniques can be incorporated to provide better results for better life of human being and the hybridization of data mining techniques will be useful in diagnosis and treatment plans of multiple diseases like Cancer prediction, HIV prediction.
9. REFERENCES

[1] Mai Shouman, Tim Turner, Rob Stocker, Using Data Mining Techniques in Heart Disease Diagnosis and Treatment. 978-1-4673-0484-9/12/ 2013 IEEE


