DATA MINING CLASSIFIERS IN THE PREDICTION OF HEART DISEASE

1S.THARANI, 2P.KARTHIK
1M.Phil. Scholar, 2Assistant Professor
Department of Computer Science,
Prist University
Trichirappalli, Tamilnadu, India

Abstract: The Healthcare industry is generally information rich but unluckily not all the data are mined which is required for discover hidden patterns & effective decision making. Data mining techniques are used to notice knowledge in database and for medical research, mainly in Heart disease prediction. Cardiovascular disease connecting high death rates Angiography is, more frequently than not, regarded as the best system for the examination of coronary artery disease; on the other hand, it was connected with significant side effects and high costs. Much investigation has, consequently, been conveyed using data mining and machine learning to attempt alternative modalities cardiovascular disease includes coronary heart disease (CHD), cerebrovascular disease (stroke), Hypertensive heart disease, congenital heart disease peripheral artery disease , rheumatic heart disease, inflammatory heart disease. The main cause of cardiovascular disease is tobacco use, physical inactivity, an unhealthy diet and harmful use of alcohol. Complex data mining benefits from the past experience and algorithm defined with existing software and packages, with certain tools gaining a greater affinity or reputation with different techniques. In this project, the various supervised machine learning classifiers like K-Nearest neighbor and support vector machine is used to identify the heart disease.

Keywords: Data Mining, Heart Disease, coronary heart disease (CHD), K-Nearest neighbor, support vector machine

I. INTRODUCTION

Data mining is a process of extracting interesting knowledge or patterns from large databases. There are several techniques that have been used to discover such kind of knowledge, most of them resulting from machine learning and statistics. The greater part of these approaches focus on the discovery of accurate knowledge. Though this knowledge may be useless if it does not offer some kind of surprisingness to the end user. The tasks performed in the data mining depend on what sort of knowledge someone needs to mine. Data mining technique are the result of a time-consuming process of study and product development.

The heart is an important organ of human body part. It is nothing more than a pump, which pumps blood through the body. If circulation of blood in body is inefficient the organs like brain suffer and if heart stops working altogether, death occurs within minutes. Life is completely dependent on efficient working of the heart. The term Heart disease refers to disease of heart & blood vessel system within it.

A number of factors have been shown that increases the risk of Heart disease: Family history, Smoking, Poor diet, High blood pressure. High blood cholesterol, Obesity, Physical inactivity, Hyper tension. Factors like these are used to analyze the Heart disease. In many cases, diagnosis is generally based on patient’s current test results & doctor’s experience. Thus the diagnosis is a complex task that requires much experience & high skill.

2. LITERATURE SURVEY

In [1] discuss the show analysis of classification data mining technique for heart disease prediction. The three algorithms used in this work are naïve bayes, WAC and apriori. The performance valuation is based on classification matrix, it display the frequency of correct and incorrect prediction. The analyze model is view in Lift charts, Bar charts and Pie charts. This struggle can be further enhanced and expanded by using other data mining techniques like Time series, Clustering and Association rules. Instead of categorical data, the continuous data can be used.

In [4] the fuzzy specialist system is planned for heart disease diagnosis with reduced number of attribute. The author find that how genetic algorithm and fuzzy logic combine together for efficient and cost effectual diagnosis of heart disease. The GA and two models of fuzzy system Mamdani and Takagi-sugeno were used to find the cost. The dataset were taken in Cleveland clinic establishment dataset. The input field is a set of all the certain kind and the output of the system is to get a value ‘1’ or ‘0’ that indicates the presence or absence of disease. It is additional improved by using art classifiers like Decision tree, Naïve bayes, Classification via clustering and SVM classifier.

In [5] data mining classification technique namely RIPPER classifier, decision tree, Artificial Neural Network and Support vector machine are used for predicting cardiovascular heart disease. The arrangement factor used for compare these technique are sensitivity, accuracy, specificity, error rate, true positive rate and false positive rate. To gauge the unbiased estimate of prediction models the author used 10 fold cross validation method. This model was industrial by using data mining classification tool weka version 3.6. It contains 14 attributes and 303 instances. Error rates for RIPPER, Artificial Neural Networks, Support vector machine and Decision tree are 0.2756, 0.2248, 0.1588 and 0.2755 respectively. The accuracy of RIPPER, Artificial Neural Networks, Support Vector Machine and Decision tree are 81.10%, 80.07%, 84.13% and 79.15% respectively. Four classification models, the Support Vector Machine has given least error rate and highest accuracy. The writer concluded that the Support Vector Machine is the best...
technique for predicting the cardiovascular disease. In future in order to improve the efficiency of the classification techniques by creating metamodels.

In [6] heart attack symptom is predicted using biomedical data mining technique. The creator used data classification which is based on supervised machine learning algorithms. For data classification the Tanagra tool is used. Using entropy based cross validations and partitioned techniques, the data is evaluated and the results are compared. The algorithms used in these techniques are Knearest neighbours, K-means and Mean Clustering Algorithm (EMC) is the extension of the K-mean algorithm for clustering process which reduces the number of iterations. In this paper result the author analyzed that the mean clustering algorithm perform well when compare to other algorithm. To run the data the time taken is very fast and it gives the result of accuracy about 83.25%. An enhanced by applying unsupervised machine learning algorithm.

In [8] association rule mining method is used for predicting heart attack. In this paper creator future a novel method CBARBSN for association rule mining based on sequence numbers and clustering the transactional database for predict heart attack. The two important step of this process first the medical data is transformed into binary and the planned method is applied to the binary transactional data. The data is collected from Cleveland database. The medical data contains 14 attributes. From the results author concluded that the proposed algorithm performs better than the existing ARNBSN ( ) algorithm. The performance of the algorithms is compared based on the execution time.

In [9] the coronary artery disease was efficiently diagnosed by rotation forest algorithm in order to support clinical decision-making process. It utilize the Artificial Neural Networks with Levenberg-Marquardt back propagation algorithm of rotation forest ensemble method as base classifier. The algorithm is implemented in matlab. From an experiment, the author diagnosed the disease by comparing the performance of base classifiers in terms of sensitivity, accuracy, AUC and specificity on two things i) without rotation forest classifier, the greatest performance of classifiers and ii) with rotation forest algorithm it actually improve the performance of classifier. Result it is experiential that Levenberg-Marquardt was the best classifier with or without random forest. The accuracy is improved to 94.2% of original classification accuracy which is an improvement of 8%. In prospect the proposed work may be used to develop efficient expert systems for the diagnosis of heart disease.

3. METHODOLOGY
3.1 EXISTING SYSTEM

The beginning of the idea of reversible watermarking in the Barton patent, some methods have been proposed. Among these solutions, most recent schemes use Expansion Embedding modulation, Histogram Shifting modulation or, more recently, their combination. One of the major concerns with these modulations is to avoid underflow and overflow. In the watermarked image while minimize at the similar occasion image distortion. Basically, EE modulation introduced by Thodiet al. is a generalization of Difference Expansion modulation proposed by Tian et al. which expand the difference between two neighboring pixels by shifting to the left its binary representation, thus creating a new virtual least significant bit (LSB) that can be used for data insertion.

DISADVANTAGE:

- The modulation is overflow and underflow.
- The image distortions are more.
- Not efficient.
- Image is not protected in correct way.
- Allows discontinuity in data protection.

3.2 PROPOSED ALGORITHMS

In this project, propose to adapt dynamically the carrier-classes by considering the local specificities of the image. Another refinement we propose is based on the selection of the most locally adapted lossless modulation. This is especially the case for medical images where large black areas exist (i.e., the background area). In these region, in a straight line applying HS on pixels may be more well-organized and of smaller complexity than applying it on prediction-error. Because, the histogram maxima corresponds to the null gray value; capacity is maximized and underflows simply avoided by shifting pixel value to the right, i.e. by adding a positive gray value. This is why propose considering the local content of the image in order to select the most locally adapted lossless modulation. This should allow us to optimize the compromise capacity/image distortion.

ADVANTAGES

- Directly applying HS on pixels may be more efficient and of smaller complexity than applying it on prediction-errors.
- The watermark embedded and extractor remain synchronized because the extractor will retrieve the same reference image. This process to select the most locally appropriate watermarking modulation. This should allow us to optimize the compromise capacity/image distortion.

3.3 K-NN CLASSIFIERS

KNN can be used for both classification and regression predictive problems. The reason of a bias towards classification models is that most analytical problem involves making a decision. These analyses are more insightful and directly link to an
In this article, we will talk about another widely used classification technique called K-nearest neighbors (KNN).

Begin
1. m=1;
2. initialize k prototypes; //arbitrarily chooses k objects as the initial centers.
3. Repeat
   for i=1 to n do
   begin
   for j=1 to k do
   Compute D(X_j, Z_j) = |X_j - Z_j| //Z_j is the center of cluster j;
   if D(X_j, Z_j) = min{D(X_j, Z_j)} then
   X_j ∈ C_j;
   end; // (re)assign each object to the cluster based on the mean
   if m=1 then
   J_m(m) = ∑_{j=1}^{k} ∑_{X_j ∈ C_j} |X_j - Z_j|^2
   m=m+1;
   for j=1 to k do
   Z_j = \frac{1}{n_j} ∑_{X_j ∈ C_j} X_j; // (re)calculate the mean value of
   the objects for each cluster
   J_m(m) = ∑_{j=1}^{k} ∑_{X_j ∈ C_j} |X_j - Z_j|^2; //compute the error
   function
   4. Until J_m(m) - J_m(m-1) < \zeta
   End

Our focus will be primarily on how does the algorithm work and how does the input parameter effect the output/prediction. In pattern recognition, the k-nearest neighbors’ algorithm (k-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depend on whether k-NN is used for categorization or regression.

3.4 SVM

SVM, a decently brand new sort of learning calculation, initially presented. Actually, SVM go for hyper plane incredible isolates the classes of information. SVMs has affirmed the ability not just too precisely isolate elements into right classes, additionally to recognize case whose build up arrangement is not upheld by information. In spite of the fact that SVM are nearly harsh characterize dispersion of preparing cases of every class. It is just reached out numerical figuring’s. Two such expansions, the first is to stretch out SVM relapse examination, where the objective to deliver a direct capacity that can genuinely exact that objective capacity. An additional expansion is to figure out how to rank components as opposed to creating a characterization for individual components. Positioning can be decreased to looking at sets of case and delivering a +1 assess if the combine is in the right positioning request not withstanding −1 generally.

Algorithm for SVM:
Step1: Select candidate = {closest pair of opp class}
Step2: while there are violating points do
Step3: Find a violator
Step4: Candidate = Candidate U violator
Step5: if any x < 0 due to addition of c to s then
Step6: Candidate = candidate/p
Step7: repeat till all such points are pruned
Step8: End of if
Step9: End of while

IV. EXPERIMENT AND RESULTS

In this study mainly classification algorithms such as KNN, SVM algorithm is used for predicting the Heart Disease from the given data set instances and the proposed algorithms are applied on type Heart Disease dataset in the VB.NET and the performance is measured. The heart is main organ of human body part. It is nothing more than a pump, which pump blood through the body. If circulation of blood in body is incompetent the organ like brain suffer and if heart stops working altogether, death occurs within minutes. The term Heart disease refers to disease of heart & blood vessel system within it. A number of factors have been shown that increases the risk of Heart disease: Family history, Smoking , Poor diet , High blood pressure , High blood cholesterol , Obesity , Physical inactivity , Hyper tension etc..Factors like these are used to analyze the Heart disease. In many cases, diagnosis is generally based on patient’s current test results & doctor’s experience. Thus the diagnosis is a complex task that requires much experience & high skill.
4.1 USE CASE DIAGRAM

![Use Case Diagram](image)

Figure 4.1 Use Case Diagram

4.2 RESULT

4.2.1 Admin Login

![Admin Login](image)

4.2.1 Admin Login
Fig: 4.2.2 Admin Home Page

Fig: 4.2.3 Image Upload

Fig: 4.2.4 Image Processing Histrom
V. CONCLUSION

Data mining is the process of extracting hidden interesting patterns from massive database. Medical domain contains heterogeneous data that can be mined properly to provide a variety of useful information for the physicians to detect a disease, predict the survivability of the patients after disease, severity of diseases etc. In the health area Heart disease can be very well predicted using data mining techniques. The application of data mining techniques for predictive analysis is very important in the health care field because it gives us the power to face diseases earlier and therefore save people’s lives through the anticipation of cure. This project was aimed to analyze the application of data mining in Heart Disease prediction. It is observed that data mining provides good results in Heart Disease diagnosis when appropriate tools and techniques applied. Anticipating diseases still remains a major challenge in medical field and pushes us to increase our efforts in developing more data mining algorithms to exploit information intelligently and extract the best knowledge from it. The disease prediction from the risk factors is done by the neural networks. The main objective of this work is to identify heart disease of K-Nearest neighbor in all iterations. The error signal is validated with the threshold value less than 0.001 in each step. Based on the value with the aid of back propagation algorithm the weights of the previous nodes are adjusted. Due to this error adjustment on each node the accuracy of the network is increased.

FUTURE ENHANCEMENT

The disease prediction through the risk factors can be hosted online and hence any internet users can access the system through a web browser and understand the risk of heart disease. Different heart diseases such as rheumatic heart disease, hypertensive heart disease, ischemic heart disease, cerebrovascular disease and inflammatory heart disease can be identified. Other health care systems can be formulated using this proposed model in order to identify the diseases in the early stage. The proposed model requires an efficient processor with good memory configuration to implement it in real time. The proposed model has wide area of application like grid computing, cloud computing, robotic modeling, etc.
REFERENCES
[15] Esra Mahsereci Karabulut & Turgay İbrikçi “Effective Diagnosis of Coronary Artery Disease Using The Rotation Forest Ensemble Method” June 2011 / Accepted: 30 August 2011 / Published online: 13 September 2011 # Springer Science+Business Media, LLC 2011
[16] V.V.Jaya Rama krishniah, D.V.Chandra Sekar, Dr.K.Ramchand H Rao, “Predicting the Heart Attack Symptoms using Biomedical Data Mining Techniques” Volume 1, No. 3, May 2012 ISSN – 2278-1080 The International Journal of Computer Science & Applications (TIJCSA)