HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES

S Anitha, N Sridevi

Journal of Analysis and Computation (JAC)
(An International Peer Reviewed Journal), www.ijaconline.com, ISSN 0973-2861
Volume XIII, Issue II, February 2019

HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES

Dr. S. Anitha[1], Dr. N. Sridevi[2]

[1]Assistant Professor, Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India
[2]Assistant Professor, College of Administrative and Financial Sciences, AMA International University, Salmabad, Kingdom of Bahrain.

ABSTRACT

Heart diseases have an abundant pact of attention in medical research due to its impact on human health. Heart diseases are amongst the nation's prominent cause of death. Data mining has developed as a vital approach for computing applications in medical informatics. Numerous algorithms connected with data mining have considerably helped to recognise medical data more evidently .In this work, supervised machine learning algorithms namely SVM, KNN and Naive Bayes are used to predict the heart diseases. The machine learning algorithms are implemented using R programming language. The performances of the algorithms are measured in terms of accuracy. The functionality of the algorithms are examined and the outcomes were deliberated.

Keywords - KNN, SVM, Naïve Bayes, Heart diseases, Data mining

[1] INTRODUCTION

There is an overwhelming progress in the amount of electronic health records being collected by healthcare facilities. Accuracy is particularly important when it comes to patient care and computerizing this enormous amount of data improves the quality of the whole system. But how do healthcare providers examine through all the information efficiently? This is where data mining has recognised to be extremely effective. Data mining combines statistical analysis, machine learning and database technology to mine hidden patterns and relationships from large databases [3].

Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information from a data set and transform the information into an understandable structure for further use. Hence this research work is intended to use data mining techniques in health care data to predict the outcomes.

Cardiovascular diseases (CVDs) have now developed as the primary cause of death in India. Heart disease and stroke are the prime causes and are accountable for >80% of CVD deaths [1]. A foremost challenge facing healthcare organizations is the provision of quality services at reasonable costs. Quality service indicates diagnosing patients correctly and administering treatments that are

effective. Clinical decisions are often made based on doctors' perception and practice rather than on the knowledge-rich data hidden in the database. This practice points to uninvited biases, mistakes and extreme medical expenses which affects the quality of facility delivered to patients [2].

Supervised learning trains a model on known input and output data so that it can predict future outputs, and unsupervised learning, which finds hidden patterns or intrinsic structures in input data. This research work is intended to use supervised machine learning algorithms to predict the heart diseases. Supervised methods are an effort to determine the association between input attributes and a target attribute. The relationship revealed is represented in a structure referred to as a model.

Classification model and regression model are the two main models in supervised learning. Here this work concentrates on classification model. Classification deals with allocating observations into distinct classes, rather than appraising continuous quantities. This research work uses some of the classification algorithms like SVM, Naïve Bayes and KNN to predict the heart diseases and compare their performance.

[2] RELATED WORK

Several research works have been done on diagnosis of heart disease. They used different data mining techniques for diagnosis & achieved different results for different methods.

Chaitrali S. Dangare et.al [4], in their work on "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", used neural networks, Decision tree and Naïve Bayes algorithms. Among the three, neural network algorithm predicts the heart disease with highest accuracy.

Sellappan Palaniappan et.al [5], "Intelligent Heart Disease Prediction System Using Data Mining Techniques" namely Decision Trees, Naïve Bayes and Neural Network. Their experimental results show that each technique has its unique strength in realizing the objectives of the defined mining goals.

Poornima Singh et.al, [6], "Effective heart disease prediction system using data mining techniques", a prediction system is developed using neural network for predicting the risk of heart level. The achieved outcomes have shown that the designed diagnostic system can efficiently forecast the risk level of heart diseases.

Era Singh Kajal and Nishika [7] in their work, "Prediction of Heart Disease using Data Mining Techniques", used K-mean clustering and MAFIA algorithm for Heart disease prediction system and achieved the accuracy of 89%.

Mirpouya Mirmozaffari et.al [8], "Data Mining Classification Algorithms for Heart Disease Prediction", proved that Random tree algorithm gives highest accuracy and lowest errors among the highest performance algorithm.

Aditya Methaila et al [9], "Early Heart Disease Prediction Using Data Mining Techniques", intends to use data mining Classification Techniques, namely Decision Trees, Naïve Bayes and Neural Network, along with weighted association Apriori algorithm and MAFIA algorithm.

A. Sheik Abdullah et al [10], " A Data mining Model for predicting the Coronary Heart Disease using Random Forest Classifier", a Data mining model has been developed using Random Forest classifier to increase the forecast accuracy and to examine several events related to CHD.

[3] DATA SET

The data set used in this research work is from UCI Machine Learning Repository [11]. The dataset is a collection of medical analytical reports with values for 76 attributes, but all published experiments refer to using a subset of 14 of them. Hence this research work also intended to use only the 14 attributes. The various attribute and their description are shown in the table 1.

TABLE 1: Attribute Information

Name	Description
Age	Age in years
Sex	1=Male ,0=Female
Trestbps	Resting blood pressure(in mm Hg)
Cp	Chest pain type
Chol	Serum Cholesterol in mg/dl
Fbs	Fasting blood sugar> 120 mg/dl
Restecg	Resting Electrocardiography results
Thalach	Maximum heart rate achieved
Exang	Exercise induced angina
Old peak ST	Depression induced by exercise relative to rest
Slope	The slope of the peak exercise segment
Ca	Number of major vessels colour by fluoroscopy that ranges between 0 and 3
Thal	3=Normal 6=fixed defect 7=reversible defect

[4] DATA MINING TECHNIQUES USED FOR PREDICTION

Three different data mining classification techniques namely KNN, Naive Bayes and Support Vector Machine are used to analyse the dataset.

K-NEAREST NEIGHBOR (KNN) ALGORITHM

Let (xi, ci) where i = 1, 2……., n be data points. xi denotes feature values & ci denotes labels for xi for each i. Assuming the number of classes as 'c'.

$c_i \in \{1, 2, 3, \ldots\ldots, c\}$ for all values of i

Let x be a fact for which label is not identified, and we would like to discover the label class using k-nearest neighbor algorithms.

PSEUDO CODE

- Calculate "d(x, x_i)" i =1, 2, ….., n; where d denotes the Euclidean distance between the points is calculated as follows
- distance=$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$
- Organise the premeditated n Euclidean distances in increasing order.
- Let k be a +ve number, take the leading k distances from this sorted list.
- Find those k-points corresponding to these k-distances.
- Let k_i denotes the quantity of facts fitting to the i^{th} class among k points i.e. $k \geq 0$
- If $k_i > k_j \; \forall \; i \neq j$ then put x in class i.

NAIVE BAYES ALGORITHM

Bayesian rational is useful to decision making. The representation for Naive Bayes is probabilities. It works on Bayes theorem of probability to predict the class of unknown data set. A list of probabilities is stored to file for a learned naive Bayes model. This includes:

- Class Likelihoods: The likelihoods of each class in the training dataset.
- Conditional Likelihoods: The conditional likelihoods of each input value given each class value.

PSEUDO CODE

Learning Phase: Learning a naive Bayes model from your training data is fast. Given a training set S and F features and L classes,For each target value of c_i(c_i=c_1,....,c_L)

$\hat{P}(c_i)$← estimate P(c_i) with examples in S;

For all feature value x_{jk} of each feature x_j(j=1,...,F;k=1,...,N_j)

$\hat{P}(x_j=x_{jk}|\ c_i)$ ←estimate P(x_{jk}|c_i) with examples in S;

Output: F* L conditional probabilistic models

Testing Phase: Training is fast because only the probability of each class and the probability of each class given different input (x) values need to be calculated. Given an unknown instance x'=(a'$_1$,...,a'$_n$)
Look up tables to assign the label c* to X' if

$[\hat{P}(a'_1|c^*)...\hat{P}(a'_n|c^*)]\hat{P}(c^*)>[\hat{P}(a'_1|c_i)...\hat{P}(a'_n|c_i)]\ \hat{P}(c_i), c_i \neq c^*, c_i = c_1,...,c_L$

SUPPORT VECTOR MACHINE

A Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification purpose. SVMs have been extensively researched in the data mining and machine learning field and applied to applications in various domains. SVMs are more commonly used in classification

problems. Two special properties of SVMs are that SVMs achieve (1) high generalization by maximizing the margin and (2) support an efficient learning of nonlinear functions by kernel trick.

SVM CLASSIFICATION

The primary method of SVMs is a twofold classifier where the output of cultured task is either true or false. Twofold SVMs are classifiers which distinguish data points of two classes. Each data points is represented by an n-dimensional vector. A linear classifier generally separates the two classes with a hyper plane, there are many linear classifiers that correctly classify the two groups of data.

In direction to attain extreme separation between the two classes, SVM picks the hyper plane which has the biggest boundary. The boundary is the summation of the shortest distance from the separating hyper plane to the nearest data point of both categories. Such a hyper plane is possible to simplify well, implication that the hyper plane properly categorizes “invisible” or testing data points.

PSEUDO CODE

Step: 1 The data points D or training set can be expressed mathematically as follows.

$$D = \{(x_1,y_1),(x_2,y_2),...,(x_m,y_m)\}$$

Here, x_i is a n-dimensional real vector, y_i is either 1 or -1 denoting the class to which the point x_i belongs.

Step: 2 The SVM classification function F(x) takes the following form where w is the weight vector and b is the bias, which will be computed by SVM in the training process.

$$F(x) = w.x-b \quad (1)$$

Step: 3 To correctly classify the training set, F(x) must return positive numbers for positive data points and negative numbers otherwise, that is, for every point x_i in D,

$$w.x_i - b > 0 \text{ if } y_i = 1 \text{ , and}$$
$$w.x_i - b < 0 \text{ if } y_i = -1 \quad (2)$$

These conditions can be revised such that

$$y_i (w.x_i - b) > 0, \forall(x_i, y_i) \in D \quad (3)$$

If there exists such a linear function F that correctly classifies every point in D or satisfies Eq.(3) then D is called linearly separable.

Step: 4 F needs to maximize the margin where margin is the distance from the hyperplane to the closest data points. To maximise the margin Eq.(3) is revised into the following Eq.(4).

$$y_i (w.x_i - b) \geq 1, \forall(x_i, y_i) \in \quad (4)$$

The distance from the hyperplane to a vector xi is formulated as $\frac{|F(xi)|}{||w||}$. Thus, the margin becomes

$$\text{margin} = \frac{1}{||w||} \quad (5)$$

Because when x_i are the neighbouring vectors, F(x) will return 1 according to Eq.(4). The neighbouring vectors, that satisfy Eq.(4) with equality sign, are called support vectors.

Thus the objective of the SVM is to find the optimal separating hyper plane which maximizes the margin of the training data.

[5] EXPERIMENTAL RESULTS

R programming is used for implementing the classification techniques. The dataset consists of 302 records in the Heart diseases database. For the experimental purpose the dataset is divided into training dataset and testing dataset in the ratio of 70:30 respectively. The data mining classification algorithms namely KNN, Naïve Bayes and Support Vector Machine are implemented using R programming.

In the arena of machine learning, a confusion matrix, otherwise called an error matrix, is a particular table design that permits perception of the execution of a calculation. Each line of the matrix speaks to the examples in an anticipated class while every section speaks to the cases in a real class. The confusion matrix has the following entries. They are, TP (True Positive): It means the quantity of records delegated genuine while they were in reality evident. FN (False Negative): It signifies the quantity of records delegated false while they were in reality obvious. FP (False Positive): It indicates the quantity of records named genuine while they were in reality false. TN (True Negative): It means the quantity of records delegated false while they were in reality false.

The confusion matrix obtained by three different algorithms is given below. In the table2, Class 0 represents heart diseases and Class 1 represents no heart diseases.

Table 2: Confusion Matrix obtained using the Classification Algorithms

	Class 0	**Class 1**
Class 0	25	7
Class 1	14	44

Confusion matrix of KNN Algorithm

	Class 0	**Class 1**
Class 0	30	5
Class 1	7	48

Confusion matrix of Naïve Bayes Algorithm

	Class 0	**Class 1**
Class 0	26	7
Class 1	13	44

Confusion matrix of SVM Algorithm

From table3, it is evident that among the three algorithms used for prediction of heart diseases using the clinical data, Naïve Bayes algorithm predicts the diseases with the highest accuracy of 86.6% when compared to KNN and SVM.

Table 3: Accuracy of the classification algorithms

Classification Algorithm	Accuracy in %
KNN	76.67
Naïve Bayes	86.6
SVM	77.7

[6] CONCLUSION

Heart disease is the most common diseases in India. Early detection of heart diseases will increase the survival rate hence this research work is intended to predict the whether the patient has heart disease or not with the help of clinical data which will assist the diagnosis process. Three supervised machine learning algorithms namely KNN, Naive Bayes and SVM are compared in terms of accuracy using the heart diseases dataset. From the experimental results it's evident that Naïve Bayes algorithm predicts the heart disease with the accuracy of 86.6%. In future, the performance of Naïve Bayes algorithm can be compared with various classification algorithms like, random forest, decision tree.

REFERENCES

[1] Prabhakaran et al, "Cardiovascular Disease in India", Circulation, Vol 133, No.16, pg.no: 1605 – 1620, 2016.

[2] G.Subbalakshmi et al., "Decision Support in Heart Disease Prediction System using Naive Bayes", Indian Journal of Computer Science and Engineering (IJCSE), Vol.2 , No.2,pg.no:170-176, 2011.

[3] Thuraisingham, B., "A Primer for Understanding and Applying Data Mining", IT Professional, 28-31, 2000.

[4] Chaitrali S. Dangare et.al, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", International Journal of Computer Applications,Vol.47,No.10,pg.no:44 – 48, 2012.

[5] Sellappan Palaniappan et.al, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", IJCSNS International Journal of Computer Science and Network Security, Vol.8, No.8,pg.no: 343 – 350, 2008.

[6] Poornima Singh et.al, "Effective heart disease prediction system using data mining techniques", International Journal of Nano medicine, pg.no:121- 124, 2018.

[7] Era Singh Kajal and Nishika, "Prediction of Heart Disease using Data Mining Techniques", International Journal of Advance Research, Ideas and Innovations in Technology, Vol.2, No.3, pg.no: 1 – 7, 2016.

[8] Mirpouya Mirmozaffari et.al, "Data Mining Classification Algorithms for Heart Disease Prediction", International Journal of Computing, Communications & Instrumentation Engg. (IJCCIE), Vol. 4, No.1, pg.no: 11-15, 2017.

[9] Aditya Methaila et al, "Early Heart Disease Prediction Using Data Mining Techniques", Computer Science and Information Technology, pg.no:53 – 59,2014.

[10] A. Sheik Abdullah et al , " A Data mining Model for predicting the Coronary Heart Disease using Random Forest Classifier", International Conference on Recent Trends in Computational Methods, Communication and Controls, pg.no:22 – 25, 2012.

[11] https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/heart-disease.names