International Journal of Software Engineering and Its Applications
Vol.13, No.2 (2019), pp.17-24
http://dx.doi.org/10.21742/ijseia.2019.13.2.03

Prediction of Heart Diseases through Artificial Intelligence and Data Mining

A V S Pavan Kumar

Dept. of Computer Science and Engineering, GIET University, Gunpur, Orissa, India
avspavankumarmca@gmail.com

Abstract

Data mining is the computational procedure of discovering styles in huge statistics sets regarding strategies on the intersection of artificial intelligence, gadget studying, facts, and database systems. It is an interdisciplinary subfield of computer technological know-how. In now a day's lifestyle illnesses are increasing increasingly more. Data mining is one of the solutions for it, it helps us to overcome this problem by exploring old datasets. For any disease if it is identified at early stage treatment can be done easily. A wide range of data is produced in health care institutions, we will use that data to get some useful information. Data mining in medical sector helps doctors for diagnosis and treatment of diseases, this paper makes an effort to study and find interesting patterns from the data of patients.

Keywords: *Data mining, Artificial intelligence, Health care*

1. Introduction

Data Mining is the technique for discovering obscure qualities from huge quantity of information. Because the population will increase, diseases are increasing daily. The examination of this restorative info is hard while not the computer-primarily based investigation [1]. The computer-primarily based on investigate b is that the large regions for the specialists to managing the big life of patient's knowledge sets from multiple points of reading, for instance, perceive complicated indicative tests, translating past outcomes, and consolidating the disparate info along. This prediction system helps the patients and declines the medicative expenses. As per a survey, around 17.5 million deaths are due to heart attacks, which means 31% of global deaths is because of heart attacks. So, if this disease is predicted earlier, it could be very helpful to everyone [2]. Therefore, we can apply data mining on historical data and retrieve some useful information regarding heart diseases.

2. Knowledge discovery process

The phrases Knowhow Discovery in Databases (KDD) and records Mining is typically used interchangeably [3]. KDD is that the technique of fixing the low-level statistics into high-degree data. Therefore, KDD refers to the nontrivial elimination of implicit, antecedently unknown and doubtlessly helpful data from knowledge in databases. Whereas statistics mining Associate in Nursing KDD are often handled as similar words however in real facts mining is an essential step within the KDD system. The Knowhow Discovery in Databases system comprises of a

Article history:
Received (July 25, 2019), Review Result (September 19, 2019), Accepted (November 8, 2019)

Print ISSN: 1738-9984, eISSN: 2208-9802 IJSEIA

couple of steps leading from raw statistics collections to some style of recent records [4]. The repetitive methodology includes the next steps:

a) **Cleaning**: It can also be known as facts cleansing it are a phase whereby noise records and unrelated statistics are eliminated from the gathering.

b) **Integration**: At this degree, various records resources, often heterogeneous, could also be shared in a very commonplace supply.

c) **Data selection**: At this step, the facts associated with the analysis is set on and retrieve from the records assortment.

d) **Data transformation**: to boot said as statistics consolidation, its miles a locality whereby the chosen statistics is remodeled into forms applicable for the mining method.

e**) Data mining**: it's far the important step whereby sensible ways are applied to extract patterns doubtless helpful.

f) **Sample evaluation**: this step, firmly fascinating designs representing ability are recognized supported given measures.

g) **Expertise illustration**: is that the closing phase whereby the determined experience is visually delineated to the user. On this tempo image techniques are accustomed facilitate customers perceive and interpret the info mining consequences.

In today's world health is major issue and medical tests, disease finding is taking longer time as we all that about it [5]. Here day by day we have increased in demand for medical facilities and new technology on medical also lead to increase in complexity in finding the better, most precise treatment, many hospitals are using databases and computers to manage patient's records which made easy to manage data and in the same way we also know that artificial intelligence is also developing now, is used in many prediction and robotic application in this we are also using some of this artificial intelligence technique on a database which is formed on applying data mining techniques on a large patients database. We are using ID3 algorithm Naïve Bayesian classifiers algorithm which can be used for prediction separately by using decision trees and tables respectively here we use both to overcome some limitations and to improve accuracy [6]. We can also use genetic algorithm and KNN algorithm to improve accuracy Another solution for this is to use clustering based on male and female data, we can classify the data based on hierarchical clustering.

3. Prediction of heart disease

Mainly Naive Thomas Bayes algorithm, Decision tree algorithm, Neural Network algorithms area unit normally used.

The Naive Bayesian classifier relies on Thomas Bayes theorem with independence assumptions between predictors. Naive Bayesian model is beneficial for terribly massive datasets with no difficult repetitious parameter estimation. Bayesian classifier outperforms additional subtle classification ways [7].

Where, P(c|x) is that the posterior probability of sophistication given predictor, P(c) is that the previous likelihood of sophistication, P(x|c) is that the chance that is that the likelihood of predictor given category, P(x) is that the previous likelihood of predictor.

Every leaf node holds a category label, each internal node denotes a take a look at on associate degree attribute, each branch four denotes the end result of a take a look at and top node within the tree is that the root node. Associate degree associated call tree is incrementally developed whereas It breaks down a dataset into smaller and smaller subsets. The result is a tree with leaf and call nodes.

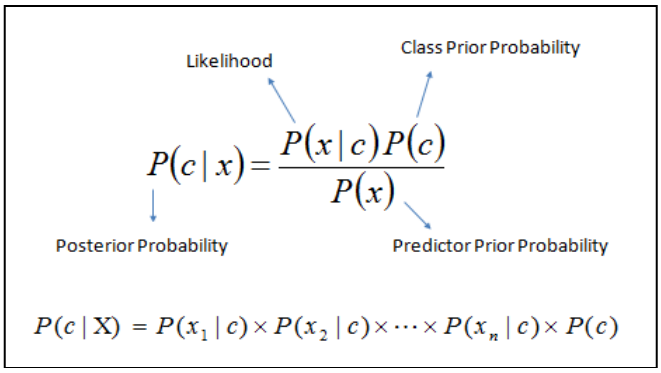

Figure 1. Bayesian classifier model

The decision tree is made top-down ranging from a root node then partition the information into subsets with similar (homogenous) values. For the homogeneity of a sample ID3 formula uses entropy to calculate. If the sample is associate degree equally divided it's entropy of 1 and if the sample is totally homogenized the entropy is zero.

Using frequency table by entropy for one attribute:

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

C4.5 is used for classification, and C4.5 is usually remarked as an applied math classifier. C4.5 chooses the attribute of info} that almost all effectively splits its set of samples into subsets enriched in one category or the opposite at every node of the tree it's the normalized information gain. Any of the attribute with the very best normalized data gain is chosen to create the choice. Then C4.5 formula recurs on the smaller sub-lists.

4. Artificial neural networks

An Artificial Neural network (ANN), often without a doubt referred to as a "neural community" (NN), maybe a mathematical version or procedure version supported biological neural networks, in specific phrases, is associate emulation of biological neural convenience [6][7]. It includes associate interconnected establishment of artificial neurons and processes info the usage of a connectionist technique to computation. In most instances associate ANN is associate accommodative convenience that changes its form all totally on out of doors or internal knowledge that flows via the community within the route of the educational part. In bigger smart phrases neural networks area unit non-linear applied mathematics facts modeling instrumentality. They will be wont to model difficult relationships among inputs and outputs or to find designs in info. A neural network is associate interconnected cluster of nodes, similar to the massive network of neurons within the human brain.

A neural community got to be designed specified the software system of a difficulty and speedy of inputs produces (both 'direct' and through a relaxation approach) the favored set of outputs. Several ways to line the strengths of the connections exist. One manner is to line the weights expressly, employing a priori ability. One another method is to 'train' the neural community with the help of feeding its education patterns and holding it amendment it weights in line with some learning rule. We are capable of categorizing the attending to apprehend things as follows:

Supervised getting to know or associative studying whereby the network is trained by providing it with entering and matching output designs. These enter-output pairs are also provided by an associate external teacher, or via the system that contains the neural network.

5. Theoretical analysis

The coronary heart sickness Prediction systems uses clinical dataset encompass parameters primarily based on threat elements as age, circle of relatives' records, diabetes, high blood pressure, excessive cholesterol tobacco smoking, alcohol consumption, and many others [4][5]. The prognosis time and enhance the diagnosis accuracy, medical Diagnostic choice aid structures. The neural community method is used for studying the heart disease statistics. Making use of feed-forward algorithm with variable gaining knowledge of pace and momentum the heart sickness database is skilled by way of the neural network. The input layer carries 13 neurons to represent thirteen attributes. It includes four magnificence labels particularly regular person, 1st stroke, second stroke and stop of life. The output layer includes 2 neurons to represent these four coaching. The neural network is built with and without hidden layer is mateless and multilayer networks area unit educated. The dataset classifies the man or woman into regular and abnormal individual based on coronary heart diseases.

Some of the causes of coronary heart disorder:

Smoking: The smoking is predominant cause of coronary heart attack, stroke and different peripheral arterial disorder. Nearly forty percent of everybody dies from smoking tobacco accomplish that due to coronary heart and blood liner illnesses. A heart assault is the death of or damage to part of the coronary heart electricity because they deliver of blood to the heart strength is significantly decreased or bunged.

Cholesterol: The strange ranges of lipids (fat) in blood are risk of coronary heart illnesses. Cholesterol is a gentle, waxy substance discovered some of the lipids in bloodstream and it can also include in all the cells of the body. The excessive ranges of LDL (low-density lipoprotein) ldl cholesterol accelerate atherosclerosis increasing the hazard of heart sicknesses.

Weight problems: that is used to explain the health condition of each person substantially above his or her I deal healthful weight. A higher hazard or fitness trouble including heart sickness, stroke, high blood strain, diabetes.

Lack of physical exercise: the shortage of workout is a chance issue used forrising Coronary Artery Disease (CAD) require in bodily work out increases danger of CAD increase for diabetes and excessive blood pressure. Coronary Heart Disorder (CHD) or Ischemic Heart Disease (IHD) is a broad term that can check with any condition that affects the coronary heart. This paper additionally offers the comparison of set of rules on accuracy and data. Normal motives for abnormal heart rhythms (arrhythmias) or situations that can spark off arrhythmias consist of:

- Coronary heart imperfections you're conceived with (intrinsic coronary heart surrenders)
- Coronary deliver path illness
- High blood pressure
- Diabetes
- Smoking
- Inordinate usage of liquor or caffeine
- Drug addiction
- Valvular coronary illness

Heart failure or congestive heart failure, this means that that the heart remains operating, however it isn't pumping blood as nicely because it has to, or getting sufficient oxygen. In addition to these, coronary heart diseases may be hereditary. That means if a person's dad and mom or grandparents has heart diseases, there may be possibility for that character to get coronary heart disorder.

6. Proposed solution

With the existing algorithms the accuracy was now not up to mark and there are issues with every of the algorithm. To conquer this, we will use Genetic set of rules and okay-nearest neighbor set of rules with addition to the present algorithms. Genetic Algorithms attempt to contain thoughts of natural evolution. In most popular, genetic learning starts as follows. An initial populace which has a set of guidelines is created. Each rule represents a string of bits. For instance, allow us to that samples in coaching set area unit delineate by means of Boolean attributes, say B1 and B2, and directions C1 and C2. The rule is "If now not B1 and B2 then C1" is also painted as 011.

6.1. K-nearest-neighbor-classifier

The training tuples are outlined by n attributes. Every tuple represents an element in Associate in nursing n-dimensional space. This means all of the tuples are holding on in n-dimensional pattern house. While an unknown tuple is given, this algorithm searches the sample space for okay tuples towards this unknown tuple. For every tuple, RMS value is calculated among unknown and recognized one Euclidean or Manhattan distances may be used. KNN algorithm is one of the most effective classification set of rules. In spite of such simplicity, it is able to give surprisingly aggressive outcomes. KNN algorithm can also be used for regression troubles. Example for KNN algorithm:

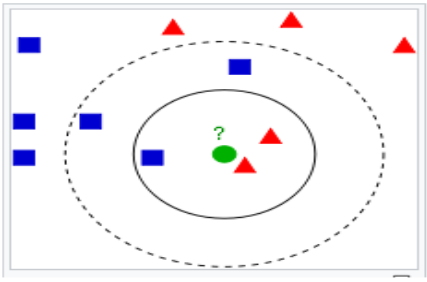

Figure 2. An example of KNN model

6.2. Clustering

Cluster analysis or agglomeration is that the task of clustering a group of objects in such some way that objects within the same group (called a cluster) are a lot of similar (in some sense or another) to every aside from to those in alternative teams (clusters). It may be a main task of exploratory records mining, and a standard technique for applied math statistics analysis, employed in several fields, any other proposed answer for this prediction of coronary heart sicknesses is, Clustering. Data may be divided into clusters initially, i.e., male and female, from that clusters a hierarchy can be created. From which we are able to predict whether the sickness

will come or now not. We also can enhance the accuracy of the usage of each Naive Bayes set of rules and choice tree algorithm.

7. Experimental analysis

The cluster Dendrogram for the currently considered model is calculated and can be represented in the form of a diagram and is shown in [Figure 3] as,

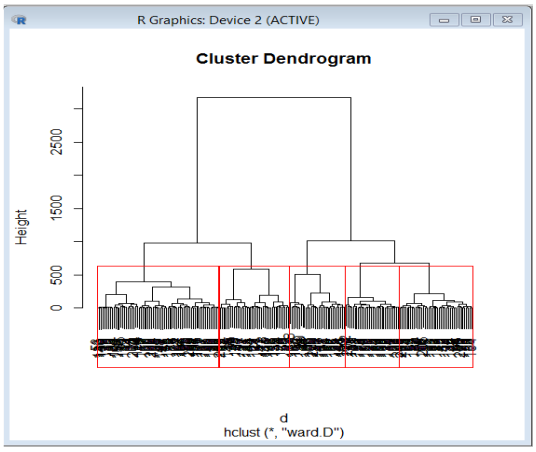

Figure 3. Dendrogram model

Table 1. Disease and symptom details

S.No	Disease	Symptom
1	Fever	Body Pains
2	Cold	Headache
3	Skin Allergy	Headache
4	Fever	Headache
5	Fever	Headache
6	Cold	Headache
7	Skin Allergy	Body Pains
8	Skin Allergy	Body Pains
9	Fever	Headache
10	Skin Allergy	Headache
11	Fever	Body Pains
12	Cold	Headache
13	Cold	Headache

By applying the clustering, Data can be divided into two clusters initially, i.e., male and female, from that clusters a hierarchy can be created. From which we can predict whether the disease will come or not. In the above image, data is divided into 5 clusters using hierarchical clustering (Agglomerative). The detailed details are represented in tabular format and can be observed in detail in [Table 1].

For better understanding the performance of the currently considered model, the frequency table and the likelihood tables are considered and taken for the better modeling and for better analysis. These results are shown in [Table 2] and [Table 3] as,

Table 2. Frequency table for various parameters

S.No	Frequency Table	Body Pain	Headache
1	Cold	0	4
2	Skin Allergy	3	2
3	Fever	2	3
Grand Total		5	9

Table 3. Likelihood table for various parameters

Likelihood Table	Body Pain	Headache	-
Cold	0	4	4/14=1.29
Skin Allergy	3	2	5/14=0.36
Fever	2	3	5/14=0.36
Grand Total	5	9	-
	5/14=0.36	9/14=0.64	-

Problem: Patients gets headache if he/she gets fever.

*P(headache | fever) = (P(fever | headache) * P(headache)) / P(fever)*

P(fever) = 5/14 = 0.36

P(headache) = 9/14 = 0.64

*P(headache/fever) = (0.33*0.64) = 0.60*

It means 60% of people who gets headache if he/she gets fever. After applying Bayes algorithm, if we use decision tree algorithm it can give good results.

8. Conclusion

The rate of diseases is increasing day by day. There are many people doesn't know about the diseases that even existed. Mainly the heart disease is the major death-causing disease around the world. This needs further step of prediction. But with the inaccurate procedures there may be fatal outcomes. So hereby with the help of developed algorithms there is chance of increasing the accuracy. This System provides the idea of prediction of disease with the usage of different algorithms. It may also help to produce a drastic improvement in prediction techniques.

References

[1] Bharti S. and Singh S. N., "Analytical study of heart disease prediction comparing with different algorithms," IEEE International Conference on Computing, Communication & Automation, pp.78-82, **(2015)** DOI: 10.1109/CCAA.2015.7148347

[2] Dewan A. and Sharma M., "Prediction of heart disease using a hybrid technique in data mining classification," In 2nd IEEE International Conference on Computing for Sustainable Global Development (INDIACom), pp.704-706, **(2015)**

[3] Krishnaiah V., Srinivas M., Narsimha G., and Chandra N. S., "Diagnosis of heart disease patients using fuzzy classification technique," In IEEE International Conference on Computing and Communication Technologies, pp.1-7, **(2014)** DOI: 10.1109/ICCCT2.2014.7066746

[4] Shouman M., Turner T., and Stocker R., "Using data mining techniques in heart disease diagnosis and treatment," In IEEE Japan-Egypt Conference on Electronics, Communications and Computers, pp.173-177, **(2012)** DOI: 10.1109/JEC-ECC.2012.6186978

[5] Sudhakar K. and Manimekalai D. M., "Study of heart disease prediction using data mining," International journal of advanced research in computer science and software engineering, vol.4, no.1, pp.1-8, **(2014)**

[6] Thanigaivel R. and Kumar K. R., "Boosted Apriori: An effective data mining association rules for heart disease prediction system," Middle-East Journal of Scientific Research, vol.24, no.1, pp.192-200, **(2016)**

[7] Kumar R. N. and Kumar M. A., "Medical data mining techniques for health care systems," International Journal of Engineering Science, vol.3, no.4, pp.98-105, **(2016)**