Journal of Innovative Engineering and Research (JIER)
Vol.- 3,Issue - 1, April 2020, pp. 7-11 (5 pages)

K-Nearest Neighbor Based Approach for Early Prediction of Cardiovascular Disease

DHANRAJ VERMA[1], RAMESWAR SINGH SIKARWAR[2]

[1,2]*Department of Computer Science & Engineering, Dr. A. P. J. Abdul Kalam University, Indore*
Corresponding Author Email: dhanrajmtech@yahoo.co.in

Abstract— Classification may be a classic data processing technique supported machine learning. Basically classification is employed to classify each item during a set of knowledge into one among predefined set of classes or groups. Classification method makes use of mathematical techniques like decision trees, applied mathematics , neural network and statistics. Classification divides data samples into target classes. The classification technique predicts the target class for every data points. for instance , patient are often classified as "high risk" or "low risk" patient on the idea of their disease pattern using data classification approach. it's a supervised learning approach having known class categories. Binary and multilevel are the 2 methods of classification. In binary classification, only two possible classes like , "high" or "low" risk patient could also be considered while the multiclass approach has quite two targets for instance , "high", "medium" and "low" risk patient. Classification techniques also are used for predicting the treatment cost of healthcare services which is increases with rapid climb per annum and is becoming a main concern for everybody .There are several algorithms and methods are developed to unravel the matter of classification. But problem are always arises for locating a replacement algorithm and process for extracting knowledge for improving accuracy and efficiency. Our major objective are-Design a classification based algorithm which classify the given data set efficiently and accurately. Design a classification based algorithm which classify the given data set using simple calculation and also reduce complexity. Design a classification based algorithm which consider all the attributes associated with the given dataset. Design a classification based algorithm which work for both categorical also as numerical of the attributes.

Index Terms— Classification, Prediction, Cardiovascular, Disease, Accuracy, Efficiency.

I. INTRODUCTION

Data Mining is one among the foremost vital and motivating area of research with the target of finding meaningful information from huge data sets. In present era, data processing is becoming popular in healthcare field because there's a requirement of efficient analytical methodology for detecting unknown and valuable information in health data. In health industry, data processing provides several benefits like detection of the fraud in insurance , availability of medical solution to the patients at lower cost, detection of causes of diseases and identification of medical treatment methods. It also helps the healthcare researchers for creating efficient healthcare policies, constructing drug recommendation systems, developing health profiles of people etc. The data generated by the health organizations is extremely vast and sophisticated thanks to which it's difficult to research the info so as to form important decision regarding patient health. This data contains details regarding hospitals, patients, medical claims, treatment cost etc. So, there's a requirement to get a strong tool for analyzing and extracting important information from this complex data. The analysis of health data improves the healthcare by enhancing the performance of patient management tasks. The outcome of knowledge Mining technologies are to supply benefits to healthcare organization for grouping the patients having similar sort of diseases or health issues in order that healthcare organization provides them effective treatments. It also can useful for predicting the length of stay of patients in hospital, for diagnosis and making plan for effective data system management. Recent technologies are utilized in medical field to reinforce the medical services in cost effective manner. Data Mining techniques also are wont to analyze the varied factors that are liable for diseases for instance sort of food, different working environment, education level, living conditions, availability of pure water, health care services, cultural ,environmental and agricultural factors.

II. CLASSIFICATION AND PREDICTION

Classification is that the process of finding a model (or function) that describes and distinguishes data classes or concepts, for the aim of having the ability to use the model to predict the class of objects whose class label is unknown. The derived model is predicated on the analysis of a group of coaching data. The derived model could also be represented in various forms, like classification IF-THEN rules, decision trees,mathematical formulae, or neural networks. Classification methods can handle both numerical and categorical attributes. Constructing fast and accurate classifiers for giant data sets is a crucial task in data processing and knowledge discovery. Classification predicts categorical class labels and classifies data based on the training set. Classification is two-stepprocess.

1**. Model construction**: describing a set of predetermined classes. Each tuple/sample is assumed to long to a predefined class, as determined by the category label attribute .The set of tuples used for model construction is training set .The model is represented as classification rules, decision trees, or mathematical formula

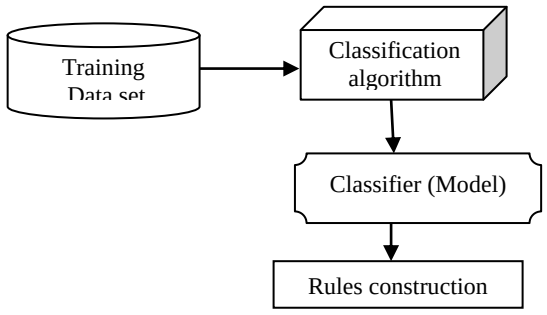

Figure 1. Model construction

2**. Model usage:** for classifying future or unknown objects. Estimate accuracy of the model. The known label of test sample is compared with the classified result from the model .Accuracy rate is that the percentage of test set samples that are correctly classified by the model .Test set is independent of coaching set.

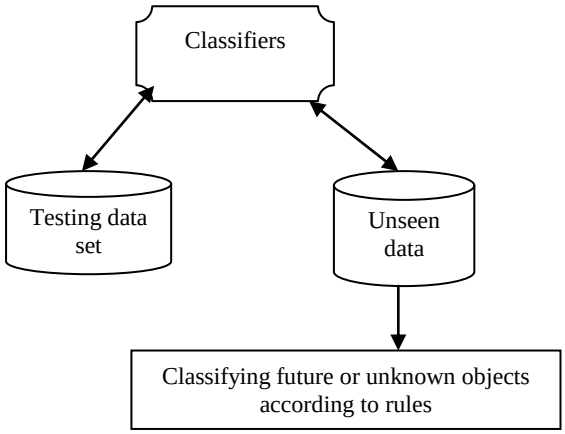

Figure 2. Model usage

III. METHODS OF CLASSIFICATION

Classification is that the task of generalizing known structure to use to new data. The classification task are often seen as a supervised technique where each instance belongs to a category , which is indicated by the worth of a special goal attribute or simply the class attribute. The goal attribute can combat categorical values, each of them like a category . One of the main goals of a Classification algorithm is to maximise the predictive accuracy obtained by the classification model when classifying examples within the test set unseen during training Three are several techniques are used for classification a number of them are.

1. Decision Tree,
2. K-Nearest Neighbor,
3. Support Vector Machines,
4. Naive Bayesian Classifiers,
5. Neural Networks.

1. Decision Trees

A Decision Tree Classifier consists of a choice tree generated on the idea of instances. A decision tree may be a classifier expressed as a recursive partition of the instance space. The decision tree consists of nodes that form a rooted tree, meaning it's a directed tree with a node called "root" that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is named an indoor or test node. All other nodes are called leaves (also referred to as terminal or decision nodes). In a decision tree, each internal node splits the instance space into two or more sub-spaces a certain discrete function of the input attributes values

2. K-Nearest Neighbor Classifiers (KNN)

K-Nearest neighbor classifiers are supported learning by analogy. The training samples are described by n dimensional numeric attributes. Each sample represents some extent in an n-dimensional space. In this way, all of the training samples are stored in an n-dimensional pattern space. When given an unknown sample, a k-nearest neighbor classifier searches the pattern space for the k training samples that are closest to the unknowns sample.

3. Support Vector Machine (SVM)

SVM may be a very effective method for regression, classification and general pattern recognition. it's considered an honest classifier due to its high generalization performance without the necessity to feature a priori knowledge, even when the dimension of the input space is extremely high. it's considered an honest classifier due to its high generalization performance without the necessity to feature a priori knowledge, even when the dimension of the input space is extremely high. The aim of SVM is to seek out the simplest classification function to differentiate between members of the 2 classes within the training data. The metric for the concept of the "best" classification function are often realized geometrically. For a linearly separable dataset, a linear classification function corresponds to a separating hyper plane f(x) that passes through the center of the 2 classes, separating the 2 "Closeness" is defined in terms of Euclidean distance, where the Euclidean distance, where the Euclidean distance between two points, X=(x1,x2,……,xn) and Y=(y1,y2,….,yn) is denoted by d(X,Y).

4. Naive Bayes Classifier

Bayesian classifiers are statistical classifiers. they will predict class membership probabilities, like the probability that a given tuple belongs to a specific class. The Naive Bayes Classifier technique is especially suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. Naïve Bayes model identifies the characteristics of patients with disorder . It shows the probability of every input attribute for the predictable state.

5. Neural Networks.

Neural Network used for classification that uses gradient descent method and supported biological systema nervosum having multiple interrelated processing elements referred to as neurons, functioning in unity to unravel specific problem. Rules are extracted from the trained Neural Network (NN) help to enhance interoperability of the learned network. to unravel a specific problem NN used neurons which are organized processing elements. Neural Network is employed

for classification and pattern recognition. An NN is adaptive in nature because it changes its structure and adjusts its weight so as to attenuate the error. Adjustment of weight is predicated on the knowledge that flows internally and externally through network during learning phase. In NN multiclass, problem could also be addressed by using multilayer feed forward technique, during which Neurons are employed within the output layer rather using one neuron.

IV. LITERATURE SURVEY

In 2011 MaiShouman, Tim Turner, Rob Stocker proposed "Using Decision Tree for Diagnosing heart condition Patients". heart condition is that the leading explanation for death within the world over the past 10 years. Researchers are using several data processing techniques to assist health care professionals within the diagnosis of heart condition . Decision Tree is one among the successful data processing techniques used. However, most research has applied J4.8 Decision Tree, supported Gain Ratio and binary discretization. GiniIndex and knowledge Gain are two other successful sorts of Decision Trees that are less utilized in the diagnosis of heart condition . Also other discretization techniques, voting method, and reduced error pruning are known to supply more accurate Decision Trees. [1].

In 2012 Sunita Soni and O. P. Vyasproposed "Fuzzy Weighted Associative Classifier: Predictive Technique for Health Care Data Mining". They extend the matter of classification using Fuzzy Association. They proposed a replacement Fuzzy Weighted Associative Classifier (FWAC) that generates classification rules using Fuzzy Weighted Support and Confidence framework. They propose a theoretical model to introduce new associative classifier that takes advantage of Fuzzy Weighted Association rule mining. This work presents a replacement foundational approach to Fuzzy Weighted Associative Classifiers where quantitative attributes are discredited to urge transformed binary database [2].

In 2013 V. Krishnaiah, Dr. G. Narsimha, Dr. N. Subhash Chandra "Diagnosis of carcinoma Prediction System using data processing Classification Technique". Cancer is that the most vital explanation for death for both men and ladies . They briefly examine the potential use of classification based data processing techniques like Rule based, Decision tree, Naïve Bays and Artificial Neural Network to massive volume of healthcare data. this is often an extension of Naive ayes to imprecise probabilities that aims at delivering robust classifications also when handling small or incomplete datasets. Discovery of hidden patterns and relationships often goes unexploited. The system extracts hidden knowledge from a historical carcinoma disease database [3].

In 2013 Shamsher Bahadur Patel, Pramod Kumar Yadav, Dr. D. P. Shukla "Predict the Diagnosis of heart condition Patients Using Classification Mining Techniques "They used three Classification function Techniques in data processing are compared for predicting heart condition with reduced number of attributes .They are Naïve Bays, Decision Tree and Classification by Clustering. In our work, Genetic algorithm is employed to work out the attributes which contribute more towards the diagnosis of heart ailments which indirectly reduces the amount of tests which are needed to be taken by a patient. Fourteen attributes are reduced to six attributes using genetic search [4].

In 2014 MariammalD., Jayanthi S., Dr. P. S. K. Patra "Major Disease Diagnosis and Treatment Suggestion Systemusing data processing Techniques". They proposes a model to systematically close those gaps to get if applying single and multiple data processing techniques to all or any disease treatment data can provide as reliable performance as that achieved in diagnosing disease. Using multiple data processing techniques the accuracy alsoimproved.Disease prediction may be a major challenge within the health care industry. rather than going for variety of tests, predicting the main disease with less number of attributes may be a challenging task in data processing . Decision Support in Disease Prediction System is developed using all the five data processing techniques. The Disease diagnosis system extracts hidden knowledge from a historical disease database. this is often the foremost effective model to predict patients with disease [5].

In 2015 Dr. G. RasithaBanuJ. H. BousalJamala"Heart Attack Prediction Using data processing Technique" data processing techniques are wont to analyze this rich collection of knowledge from different perspectives and deriving useful information. They design and develop diagnosis and prediction system for heart diseases supported predictive mining. heart condition may be a term that assigns to an outsized number of medical conditions associated with heart. These medical conditions describe the abnormal health conditions that directly influence the guts and every one its parts. heart condition is major ill health in today's time. They analyzing the varied data processing techniques introduced in recent years for heart condition prediction. disorder remains the most important explanation for deaths worldwide. [6].

In 2016 R. Sabah K. Anandakumar" Study on disorder Classification Using Machine Learning Approaches". The diagnosis of heart condition which depends in most cases on complex grouping of clinical and pathological data. thanks to this complexity, the interest increased during a significant amount between the researchers and clinical professionals about the efficient and accurate heart condition prediction. just in case of heart condition , the right diagnosis in early stage is vital as time is extremely crucial. Numeral number of tests must be requisite from the patient for detecting a disease. Machine learning based method is employed to classify between healthy people and other people with disease. disorder is that the principal source of deaths widespread and therefore the prediction of heart condition is critical at an untimely phase. [7].

In 2017 Sanjay Kumar Sen. "Predicting and Diagnosing of heart condition Using Machine LearningAlgorithms".In order to scale back the massive scale of deaths from heart diseases, a quicksand efficient detection technique is to be discovered. data processing techniques and machine learning algorithms play a really important role during this area. The researchers accelerating their research works to develop a software with the assistance machine learning algorithm which may help doctors to require decision regarding both prediction and diagnosing of heart condition . the most objective of this research paper is predicting the guts disease of a patient using machine learning algorithms. Comparative study of the varied

performances of machine learning algorithms is completed through graphical representation of the results. They administered an experiment to seek out the predictive performance of various classifiers. We select four popular classifiers considering their qualitative performance for the experiment. [8].In 2018 Peoria V, Gladys D "A novel approach for diagnosing heart condition with hybrid classifier."They proposed anOrthogonal Local Preserving Projection (OLPP) method to scale back the function dimension of the input high-dimensional data. The dimension reduction improves the prediction rate with the assistance of hybrid classifier i.e. Group Search Optimization Algorithm (GSO) combine with the Liebenberg-Marquardt (LM) training algorithm within the neural network. The LM training algorithm is employed to unravel the optimization problem and it determines the simplest network parameters like weights and bias that minimizes the error. [9].

V. PROBLEM DEFINITION

The main problem associated with classification techniques are

1. Accuracy: - This include accuracy of the classifier in term of predicting the class label, guessing value of predicted attributes.

2. Speed:- This include the specified time to construct the model (training time) and time to use the model (classification/prediction time)

3. Robustness:- This is that the ability of the classifier or predictor to form correct predictions given noisy data or data with missing values.

4. Scalability:- Efficiency in term of database size.**5. Interpretability:-** Understanding and insight provided by the model. Interpretability is subjective and thus harder to assess.

VI. OBJECTIVE

There are several algorithms and methods are developed to unravel the matter of classification. But problem are always arises for locating a replacement algorithm and process for extracting knowledge for improving accuracy and efficiency. Our major objective are

1. Construct an efficient classification method which classifies the given data set accurately.
2. Construct a classification method which classify the given data set using simple calculation and also reduce complexity.
3. Consturct a classification method which consider all attributes for disorder prediction.
4.Design a classification based algorithm which work for both categorical also as numerical of the attributes.

VII. PROPOSED METHOD

Let (Xi, Ci) where i = 1, 2……., n be data points. Xi denotes feature values &Ci denotes labels for Xi for every i. assuming the amount of classes as 'C'. C ∈ {1, 2, 3, ……, C} for all values of I Let X be some extent that label isn't known, and that we would really like to seek out the label class using proposed approach.

Input:

Given disorder database

Given tuple with condition for disorder

Output:

Cardiovascular Disease yes/no

Method:

1. Calculate "d(X, Xi)" i =1, 2, ….., n; where d denotes the Euclidean distance between the points.
2. Arrange the calculated n Euclidean distances in non-decreasing order.
3. Select minimum distances from this sorted list.
4. Find those distance value which have minimum distance with points like now .
5. Let m denotes the amount of points belonging to the class X points i.e. $m \geq 0$
6. Now find all class level for m point. Calculate
7. percentage values for Yes and No condition.

Number of Records	Bayesian Classifiers	Proposed Approach
1000	0.4337	0.6682
2000	0.4664	0.7622
5000	0.4182	0.7556

VIII. OUTLINE OF PROPOSED APPROACH

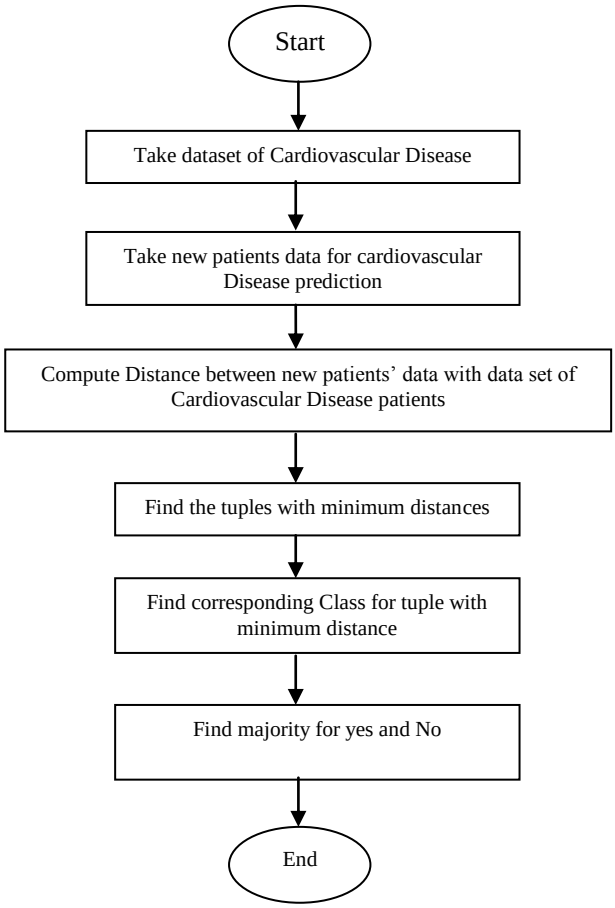

Figure 3. Outline of proposed approach

IX. RESULT AND ANALYSIS

For comparing the performance of the proposed approach we implement the Bayesian Classifiers and proposed approach. Our comparison is based on accuracy and number of tuples

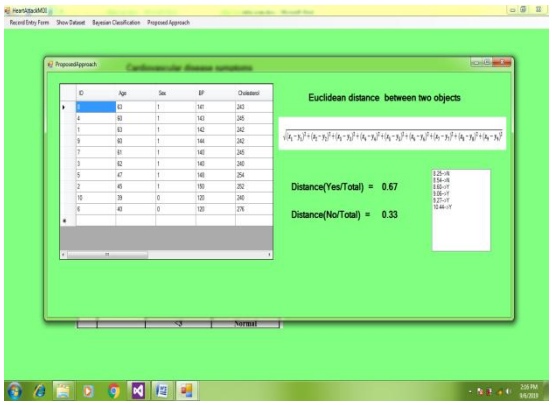

Figure 4. Implementation of proposed work

Table 1. Number of Record and accuracy in percentages

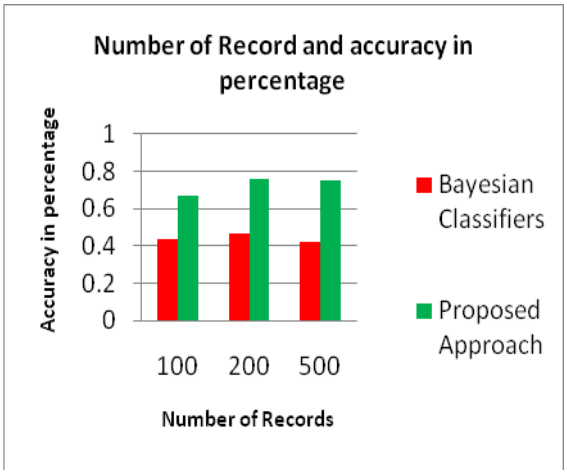

Figure 5. Comparisons Graph Number of Record and accuracy in percentages

X. CONCLUSION

There are several algorithms and methods are developed for classify disorder problem accurately. But problem are always arises for locating a replacement algorithm and process for extracting knowledge for improving accuracy and efficiency the foremost popular classification methods are Artificial neural networks, Decision Tree and Support Vector Machine and Naïve Bayes Classifier. From the experiment it clear that proposed method is more accurately classify the recodes as compared to previous method. Proposed method considers all attribute given to disorder condition. Proposed method is additionally simple to understand and calculation is simple.

REFERENCES

1. Mai Shouman, Tim Turner, Rob Stocker "Using Decision Tree for Diagnosing heart disease Patients" Proceedings of the 9-th Australasian processing Conference (AusDM'11), Ballarat, Australia.
2. SunitaSoni andO.P.Vyas "Fuzzy Weighted Associative Classifier: A Predictive Technique For Health Care Data Mining" International Journal of computing , Engineering and knowledge Technology (IJCSEIT), Vol.2, No.1, February 2012.
3. V. Krishnaiah, Dr. G. Narsimha and Dr. N. Subhash Chandra" Diagnosis of carcinoma Prediction System Using processing Classification Techniques" (IJCSIT) International Journal of computing and knowledge Technologies, Vol. 4 (1), 39 – 45, 2013.
4. ShamsherBahadur Patel, Pramod Kumar Yadav and Dr. D. P.Shukla "Predict the Diagnosis of heart disease Patients Using Classification Mining Techniques" IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS) e-ISSN: 2319-2380, p-ISSN: 2319-2372. Volume 4, Issue 2 (Jul. - Aug. 2013).
5. Mariammal. D, Jayanthi. S, Dr. P. S. K. Patra Major Disease Diagnosis and Treatment Suggestion System Using processing Techniques International Journal of Advanced Research in computing & Technology IJARCST All Rights Reserved 338 Vol. 2 Issue Special 1 ISSN: 2347 - 8446 (Online) ISSN: 2347 – 9817, Jan-March 2014.
6. Dr. G. RasithaBanu, J.H.BousalJamala "Heart Attack Prediction Using processing Technique" International Journal of recent Trends in Engineering and Research (IJMTER) Volume 02, Issue 05, ISSN (Online):2349–9745 ; ISSN (Print):2393-8161, May – 2015.
7. R. SubhaK. AnandakumarStudy on disorder Classification Using Machine Learning Approaches International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 6 (2016) pp 4377-4380 © Research India Publications. http://www.ripublication.com.
8. Sanjay Kumar Sen Predicting and Diagnosing of heart disease Using Machine Learning Algorithms International Journal Of Engineering And computing ISSN:2319-7242 Volume 6 Issue 6 June 2017, Page No. 21623-21631 Index Copernicus value (2015): 58.10 DOI: 10.18535/ijecs/v6i6.14.
9. Poornima V, Gladis D "A novel approach for diagnosing heart disease with hybrid classifier". Biomedical Research 2018; 29 (11 2274-2280 ISSN 0970-938X.