ISSN (Print) : 2319-8613
ISSN (Online) : 0975-4024

PREDICTION OF HEART DISEASE USING K-MEANS and ARTIFICIAL NEURAL NETWORK as HYBRID APPROACH to IMPROVE ACCURACY

Amita Malav[#1], Kalyani Kadam[#2], Pooja Kamat[#3]
[#]Computer Science, Symbiosis International University, Pune, India
[1] amita.malav@sitpune.edu.in
[2] kalyanik@sitpune.edu.in
[3] pooja.kamat@sitpune.edu.in

Abstract— The heart is important organ of human body part. Life is completely dependent on efficient working of the heart. What if a heart undergoes a disorder, cardiovascular diseases are the most challenging disease for reducing patient count. According to survey conducted by WHO, about 17 million people die around the globe due to cardiovascular diseases i.e 29.20% among all caused death, mostly in developing countries. Thus there is a need of getting rid of the this complicated task CVD using advanced data mining techniques, in order to discover knowledge of Heart disease prediction. In this paper, we propose an efficient hybrid algorithmic approach for heart disease prediction. This paper serves efficient prediction technique to determine and extract the unknown knowledge of heart disease using hybrid combination of K-means clustering algorithm and artificial neural network.

In our proposed model we considered 14 attribute out of 74 attributes of UCI Heart Disease Data Set [19]. This technique uses medical terms such as age, weight, gender, blood pressure and cholesterol rate etc for prediction. To perform grouping of various attributes it uses k-means algorithm and for predicting it uses Back propagation technique in neural networks. The main objective of this paper is to develop a prototype for predicting heart diseases with higher accuracy rate.

Keyword- Heart disease, K-means, artificial neural network, cardiovascular diseases

I. INTRODUCTION

At the age above 30, the heart attack or CVD is a common problem can be seen in all human beings. Along with changing lifestyle there are many such factors such as smoking, alcohol, cholesterol level, obesity, high blood pressure, diabetes etc. which are responsible factors for the risk of having heart problems. However, resent studies says that, with the introduction of artificial intelligence and medical sciences, we can actually help in preventing any such kind of diseases.

Data mining plays a vital role in healthcare domain. Data Mining and Machine learning comes up as an emerging field of high importance for providing prognosis and a deeper understanding of medical data [9]. In a old survey. The World Health Organization (WHO) has evaluated that 17 million deaths occur in world, every year due to the Heart diseases [12]. Prediction by using data mining techniques gives us accurate result of Heart Diseases. The prediction can solve complicated queries for detecting heart disease and thus assist medical practitioners to make smart clinical decisions.

Researchers are suggesting that applying data mining techniques in identifying effective treatments for patients can improve practitioner performance. Researchers have been investigating and applying different data mining techniques in the diagnosis of heart disease to identify which data mining technique can provide more reliable accuracy. Different data mining techniques have been used to help health care departments in the diagnosis of heart disease [13] [14]. Those most frequently used focus on classification: Naïve Bayes, decision tree, and neural network. In such one of the systems, has used Back-Propagation in neural network which is stated as the best prediction algorithm. The system shows a non-linear relationship between the data and the target output. The characteristics of BP algorithm are that it is adaptive and tolerant towards the noisy data or other outliers present in the medical data.

In our proposed system, we are proposing a hybrid approach to predict or diagnose heart disorders using UCI heart disease dataset [11]. by combining K-means and ANN algorithm. The main goal is to obtain high accuracy rate of prediction. Flow of the paper is given as; after the introduction section, a proper literature survey is made in section II. Section III specifies the proposed system architecture and flow chart of the implementation steps. Next section specifies about the steps required for Kmeans and ANN algorithm. In section IV experiment result is summarized.

II. RELATED WORK

In this section, Data mining techniques used for decision making in heart disease are analysed.

Ankita Dhewan and Meghana Sharma proposed a methodology of hybridizing two data mining techniques like Artificial Neural Network and Genetic Algorithm which was implemented to achieve high accuracy with least error [1].

Limitations: The very big disadvantages of GA are unguided mutations. The mutation operator in GA functions like adding a randomly generated number to a parameter of an individual of the population [10]. This is the only reason of a very slow convergence of genetic algorithm. The time consumed for optimization is much high.

M.Akhil jabbar, B.L Deekshatulua proposed algorithm into two parts i.e first part deals with evaluating attributes using genetic search and second part deals with building classifier and measuring accuracy of classifier. In this paper it compares the accuracy of datasets with and without GA. Results shows that accuracy is increases by 5% when this two are combined.

Limitations: Accuracy is very low with K-nearest neighbour and genetic algorithm takes much more time for optimization [7].

Rovina Dbritto, Aniruddha has given three data mining techniques viz. Naïve Bayes, Support Vector Machine, K-nearest neighbour and Logistic regression. Results shows that Naïve Bayes gives more accuracy compared to other classifier even.

Limitations: The disadvantage is that the Naive Bayes classifier makes a very strong assumption on the shape of your data distribution, i.e. any two features are independent given the output class. Due to this, the result can be very bad. Dependencies among attributes cannot be modelled using Bayesian classifier [2].

Humar Kahramanli, Novruz Allahverdi, used a hybrid neural network that includes artificial neural network and fuzzy neural network. A datasets of 303 samples were taken from patients with heart disease which give 87.4% accuracy on attributes of UCI repository[5] [12].

Limitations: When fuzzy system is combined with neural network, fuzzy systems need to be tuned which is very time consuming and error-prone.

Sudha, Sarath Kumar proposed two algorithms i.e KNN and K-means [15]. Their accuracy was measured which shows that KNN achieve 100% accuracy for different cluster with nearest value while K-means achieves 100% accuracy when value of K number of cluster have is very high.

Limitations: Computation cost is very high as we have to calculate the distance of each query instance to all training samples [8]

Mai Shouman, Tim Turner used a single data mining techniques on different datasets which shows that results can't be compared because of use of different datasets. When single and hybrid data mining techniques on Cleveland datasets in heart disease diagnosis results shows that hybrid techniques shows better results than single techniques. The hybrid technique used was Neural Network ensemble [3].

Limitations: Ensemble training is several times slower than traditional neural network. When solving some rare problems, the ensemble error is greater than error of a traditional neural network.

M. Veera Krishna and S. Prem Kumar proposed different data mining techniques like KNN, K-means, Apriori, PLS-DA. Comparison is done on Performance measures like computational time, positive precision values, negative precision values etc. Results shows that PLS-DA outperforms based on all performance measures.

Limitations: PLS-DA is a complex algorithm which is very difficult to use [4].

III. PROPOSED SYSTEM

In this section we mentioned about the system architecture.fig 1 represents the overview of systems architecture. The core modules of the proposed system consist of :

a) Understanding the input data and selecting the attribute related to heart disease.

b) **Data Preparation:** transformation and pre-processing of missing data is carried out.

c) **Processing Module:** it specifies about the algorithmic approach applied over the system to obtain high accuracy result. Pre-processing modules are separately discussed in upcoming section.

d) **Evaluation and deployment:** Final Analysing modules provide information related to generated output. It compares and conclude about measurable resultant artefacts like sensitivity, accuracy etc.

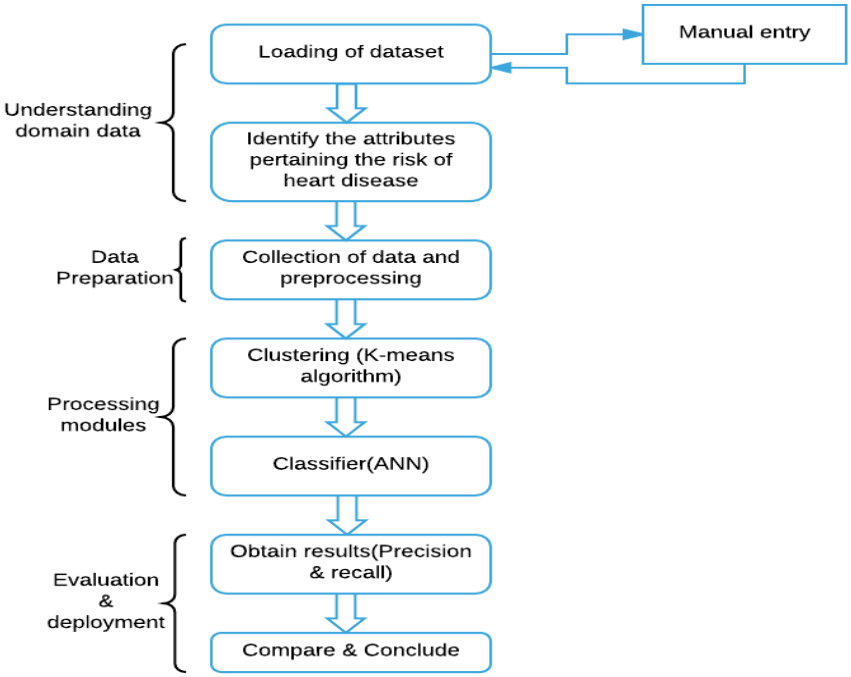

Fig 1: Overview of Proposed System

For diagnostic purpose we have considered these 14 attributes [14]:

Age in years, sex (male, female), chest pain type, resting blood pressure, serum cholesterol in mg/dl , fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, thalassemia, number of major vessels and angiographic disease status, etc.

IV. Algorithmic Description

A. *K-means Algorithm*

The main goal of using Kmeans clustering technique is that it organizes the data into classes such that there is

- high intra-class similarity
- low inter-class similarity

K-means algorithm [15] [16] is famous clustering algorithm widely used in data mining project. The main aim of this clustering is to find the positions μ_i, i=1...k within-cluster to minimize sum of squares distance from the centroid. K-means algorithm depends on k clusters, and it may stuck for different solutions. So to remove such dependency, modified or improved k-means was proposed. Kmeans is accompanied with Lloyd's algorithm to get rid of dependencies. Using this method the results show the quality of clusters is not compromised.

Steps for K-means algorithm are [15]:

1. Initialize the center of the clusters from n data points x_i, i=1...n that have to be partitioned in k clusters
2. Attribute the closest cluster to each data point using Euclidean distance
3. Set the position of each cluster to the mean of all data points belonging to that cluster
4. Repeat steps 2-3 until convergence

In our system Kmeans algorithm plays a crucial role in order to obtain the appropriate number of data groups. Using this algorithm along with Euclidean distance centroids are calculated for different patient attribute. Mean value is taken into account for sample data and henceforth it is judgemental to predicate the patient status. If the mean value of the patient is nearest to the sample mean value, the patient more likely to be affected by heart disease.

B. Artificial Neural Network

An artificial neural network (ANN), usually called neural network (NN). ANN is a mathematical model or computational model that is inspired by the structure and/or functional aspects of biological neural networks [17] [18]. There are three input layers are present in ANN: input layer, hidden layer also called as intermediate layer and output layer. Hidden layers are present in between input and output layer.

Input layer: The input units present in this layer shows the raw information that is fed into the network.

Hidden layer: The activity of each hidden unit is based on the activity of each input unit and weights on the connection between them.

Output layer: The activity of each output unit is based on the activity of each hidden unit and weights on the connection between them.

The ANN algorithm follows:

1. The data from input layer is given to hidden layer.
2. Input values from input layer are used and modified using some weight value and sent to output layer.
3. The value is again modified by some weights from connection between hidden and output layer.
4. This information is processed and output layer gives final output. Finally, this output is processed by activation function.

ANN follows trial and error method in order to get optimal solution. The structure of neural network is shown in Fig 2 [20]

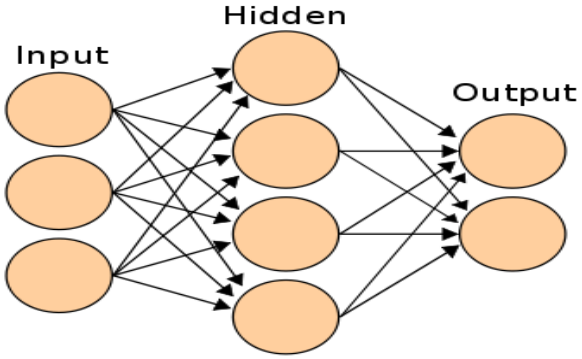

Fig 2: layers of ANN

The output is calculated by below function.

$$y_j = \sum_{f=1}^{k} w_{ij} x_i$$

Where,

y_j represents output neuron.

x_i is input neuron

w_{ij} is the weight connecting x_i and y_j

$\sum$ is sigmoidal function

As mention in figure 2, ANN consist of three layers input layer, hidden layer also called as intermediate layer and output layer. In this system former clustered normalized data groups are feed as input to neuron. The patterns vital to heart attack prediction are selected on basis of the computed significant weightage. Weightage are provided based on the range decided for the selected attribute from the dataset. For example

- sex : (1 = male; 0 = female)
- fbs : (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
- cp: chest pain type (Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic)
- restecg: resting electrocardiographic results (Value 0: normal, Value 1: having ST-T wave abnormality, Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria) etc.

The neural network is trained on dataset. The dataset is divided into two part 70% and 30%. Based on this approach the experimental results are shown below in Table 1[3].

TABLE I ALGORITHMIC PERFORMANCE

Algorithm Used	Accuracy	Time taken (in seconds)
Naïve Bayes	88%	900
KNN	93%	819
Hybrid	97%	778

V. CONCLUSION

As heart disease patients are increasing every year, huge amount of medical data is available. Researchers are applying data mining techniques on this data to diagnosis heart disease. It is analysed that artificial neural network algorithm is best for classification of knowledge data from large amount of medical data.

Population is growing in exponential way. Death rate due to cardiovascular diseases is also increasing. The only solution to control this is to predict the heart disease and medicate it before it gone worse. Our hybrid approach gives higher accuracy rate of 97% of disease detection than earlier proposed method.

REFERENCES

[1] Dewan, A., & Sharma, M. (2015, March). Prediction of heart disease using a hybrid technique in data mining classification. In Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference on (pp. 704-706). IEEE.
[2] Dbritto, Rovina, Anuradha Srinivasaraghavan, and Vincy Joseph. "Comparative Analysis of Accuracy on Heart Disease Prediction using Classification Methods." International Journal of Applied Information Systems 11.2 (2016): 22-25.
[3] Shouman, M., Turner, T., & Stocker, R. (2012, March). Using data mining techniques in heart disease diagnosis and treatment. In Electronics, Communications and Computers (JEC-ECC), 2012 Japan-Egypt Conference on (pp. 173-177). IEEE.
[4] Lakshmi, K. R., Krishna, M. V., & Kumar, S. P. (2013). Performance comparison of data mining techniques for predicting of heart disease survivability. International Journal of Scientific and Research Publications, 3(6), 1-10.
[5] Kahramanli, H., & Allahverdi, N. (2008). Design of a hybrid system for the diabetes and heart diseases. Expert systems with applications, 35(1), 82-89.
[6] Kaur, B., & Singh, W. (2014). Review on heart disease prediction system using data mining techniques. International journal on recent and innovation trends in computing and communication, 2(10), 3003-3008.
[7] Deekshatulu, B. L., & Chandra, P. (2013). Classification of heart disease using k-nearest neighbor and genetic algorithm. Procedia Technology, 10, 85-94.
[8] Dr.Mohanraj, SubhaSuryaa, Sudha, Sarath Kumar, "Heart Disease Prediction using K Nearest Neighbour and K Means Clustering", International Journal of Advanced Engineering Research and Science (IJAERS) 2016.
[9] Shinde, R., Arjun, S., Patil, P., & Waghmare, J. (2015). An Intelligent Heart Disease Prediction System Using K-Means Clustering and Naïve Bayes Algorithm. IJCSIT) International Journal of Computer Science and Information Technologies, 6(1), 637-639.
[10] Kaya, Y., & Pehlivan, H. (2015, November). Feature selection using genetic algorithms for premature ventricular contraction classification. In Electrical and Electronics Engineering (ELECO), 2015 9th International Conference on (pp. 1229-1232). IEEE.
[11] Palaniappan, Sellappan, and Rafiah Awang. "Intelligent heart disease prediction system using data mining techniques." Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on. IEEE, 2008.
[12] Boutayeb, A., & Boutayeb, S. (2005). The burden of non communicable diseases in developing countries. International journal for equity in health, 4(1), 2.
[13] Koh, H. C., & Tan, G. (2011). Data mining applications in healthcare. Journal of healthcare information management, 19(2), 65.
[14] Srinivas, K., Rani, B. K., & Govrdhan, A. (2010). Applications of data mining techniques in healthcare and prediction of heart attacks. International Journal on Computer Science and Engineering (IJCSE), 2(02), 250-255.
[15] Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43-48.
[16] Masethe, H. D., & Masethe, M. A. (2014, October). Prediction of heart disease using classification algorithms. In Proceedings of the world congress on engineering and computer science (Vol. 2, pp. 22-24).
[17] Rani, K. U. (2011). Analysis of heart diseases dataset using neural network approach. arXiv preprint arXiv:1110.2626.
[18] Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. Journal of clinical epidemiology, 49(11), 1225-1231.
[19] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. The American journal of cardiology, 64(5), 304-310.
[20] https://www.google.co.in/search?q=artificial+neural+network&rlz=1C1CHWA_enIN623IN623&source=lnms&tbm=isch&sa=X&ved=0ahUKEwiRvtuehKrVAhXHvrwKHf9tCQIQ_AUICigB&biw=1366&bih=662#imgrc=FWjS2LY7EIFgCM:

AUTHOR PROFILE

Amita Malav, Symbiosis International University, Pune, India.

Kalyani Kadam, Symbiosis International University, Pune, India.

Pooja Kamat, Symbiosis International University, Pune, India.