Intelligent Heart Disease Prediction System Using Probabilistic Neural Network

Indira S. Fal Dessai

B V B College of Engineering & Technology (VTU University), Hubli, Karnataka, India
E-mail : indirafaldessai@gmail.com

***Abstract* – The diagnosis of diseases is a crucial and difficult job in medicine. The recognition of heart disease from diverse features or signs is a multi-layered problem that is not free from false assumptions and is frequently accompanied by hasty effects [1]. Thus an attempt to exploit knowledge and experience of several specialists and clinical screening data of patients composed in databases to assist the diagnosis procedure is regarded as a great challenge. The healthcare industry gathers enormous amounts of heart disease data that unfortunately, are not mined to determine concealed information for effective diagnosing [2]. In this paper, an efficient approach for the intelligent heart disease prediction based on Probabilistic Neural Network (PNN) technique is proposed.**

Initially, the data set containing 13 medical attributes were obtained from the Cleveland heart disease database. The dataset is clustered with the aid of k-means clustering algorithm. The PNN with radial basis function is trained using the selected data sets. The comparison between existing approaches and proposed approach with respect to accuracy of prediction, sensitivity and specificity is recorded in this paper. The results from experiments returned with diminishing fact that there is considerable improvement in classification and prediction. The proposed PNN works as promising tool for prediction of heart disease.

Keywords – Heart disease, Probabilistic neural network (PNN), artificial neural network (ANN), K-means clustering, Naïve Bayes, Data mining.

I. INTRODUCTION

Nowadays provision of quality services at affordable costs is a great challenge faced by hospitals, or medical centers. Quality service implies diagnosing patients correctly and administering treatments that are effective. Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable. Hospitals must also minimize the cost of clinical tests. This can be achieved by employing appropriate computer-based information and/or decision support systems [3].

Most hospitals today employ some sort of patient Information systems to manage their healthcare or patient data [3]. When patient is suffering from heart disease minimum of 13 data are collected by the system [4]. These systems typically generate huge amounts of data which take the form of numbers, text, charts and images. Unfortunately, these data are rarely used to support clinical decision making. There is a wealth of hidden information in these data that is largely untapped. This raises an important research problem as how can we turn these data into useful information that can enable heart disease prediction. This is the main objective for this paper work.

Clinical databases are elements of the domain where the procedure of data mining has developed into an inevitable aspect due to the gradual incline of medical and clinical research data [5]. It is possible for the healthcare industries to gain advantage of Data mining by employing the same as an intelligent diagnostic tool. Therefore, data mining has developed into a vital domain in healthcare [6]. Data mining can deliver an assessment of effectiveness, from the available courses of action [7] by comparing and evaluating causes, symptoms, and courses of treatments. Working on heart disease patients databases is one kind of a real-life application. The detection of a disease from several factors or symptoms is a multi-layered problem and might lead to false assumptions frequently associated with erratic effects. Therefore it appears reasonable to try utilizing the knowledge and experience of several specialists collected in databases towards assisting the diagnosis process [8], [9].

II. REVIEW OF EXISTING APPROACHES

Numerous works for heart disease prediction is achieved by artificial neural network and data mining techniques. Recently a model of Intelligent Heart Disease Prediction System (IHDPS) was built with the aid of neural network and data mining techniques like

decision trees, naïve bayes [3]. The existing approach based on neural network utilizes a multi-layer perceptron (MLP) with back-propagation (BP) algorithm to train the selected significant patterns [1].

A. Multi-Layer Perceptron Neural Network (MLPNN)

Literature analysis unveils a persistent application of feed forward neural networks, from amidst the various categories of connections for artificial neurons [10]. A kind of feedforward neural network mechanism is the Multi-layer Perceptron Neural Networks (MLPNN). The structure of MLPNN is shown in Fig. 1.

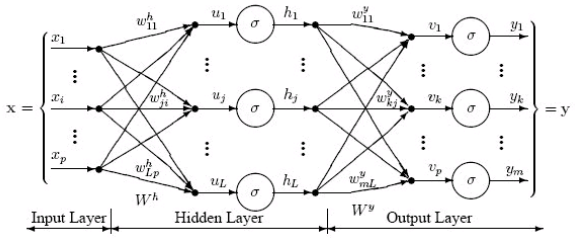

Fig. 1. Structure of MLPNN

In MLPNN the lone and primary task of the neurons in the input layer is the division of the input signal x_i among neurons in the hidden layer. Every neuron j in the hidden layer adds up its input signals x_i once it weights them with the strengths of the respective connections wji from the input layer and determines its output yj as a function f of the sum, given as

$$y_j = f(\sum w_{ji} x_i) \qquad (1)$$

At this instant it is possible for f *to* be a simple threshold function such as a sigmoid, or a hyperbolic tangent function. The output of neurons in the output layer is determined in an identical fashion.

B. Back-Propagation Training

The back-propagation algorithm can be employed effectively to train neural networks; it is widely recognized for applications to layered feed-forward networks, or multi-layer perceptrons [11]. The back-propagation learning algorithm can be divided into two phases: propagation and weight update [12].

Phase 1: Propagation

1. Forward propagation of a training pattern's input through the neural network in order to generate the propagation's output activations.
2. Back propagation of the propagation's output activations through the neural network using the training pattern's target in order to generate the deltas of all output and hidden neurons.

Phase 2: Weight update

For each weight-synapse:

1. Multiply its output delta and input activation to get the gradient of the weight.
2. Bring the weight in the opposite direction of the gradient by subtracting a ratio of it from the weight.

Repeat the phase 1 and 2 until the performance of the network is good enough.

The results illustrated the peculiar strength of each of the methodologies in comprehending the objectives of the specified mining objectives. IHDPS was capable of answering queries that the conventional decision support systems were not able to. It facilitated the establishment of vital knowledge, e.g. patterns, relationships amid medical factors connected with heart disease. IHDPS subsists well being web-based, user-friendly, scalable, reliable and expandable.

C. Data mining

The existing heart disease prediction system with aid of data mining techniques used supervised or unsupervised learning methods. Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases [13]. In supervised learning, a training set is used to learn model parameters whereas in unsupervised learning no training set is used (e.g., k-means clustering is unsupervised) [14].

Naive Bayes or Bayes' Rule is the basis for many machine-learning and data mining methods [15]. The rule (algorithm) is used to create models with predictive capabilities. It provides new ways of exploring and understanding data. It learns from the "evidence" by calculating the correlation between the target (i.e., dependent) and other (i.e., independent) variables. Naïve Bayes (NB) is stated as

$$P\left(h/d = \frac{P(d/h)P(h)}{P(d)}\right) \qquad \ldots\ldots. (2)$$

Where,

P (h): Prior belief (probability of hypothesis h before seeing any data)

P (d/h): likelihood (probability of the data if the hypothesis h is true)

P (d) = ∑ P (d/h). P (h): data evidence (marginal probability of the data)

P (h/d): Posterior (Probability of hypothesis h after having seen the data d)

In the experiments, it is observed that the Naïve Bayes classifier performs almost at par with the other classifiers in most of the cases [16]. Of the different experiments carried out on various datasets, the Naïve Bayes classifier shows a drop in performance in only 3-4 cases, when compared with J48 and MLPNN. This proves the widely held belief that though simple in concept, the Naïve Bayes classifier works well in most data classification problems.

D. Decision Tree

It is a predictive machine-learning model that decides the target value (dependent variable) of a new sample based on various attribute values of the available data. Decision tree (DT) algorithms include CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3) and C4.5. These algorithms differ in selection of splits, when to stop a node from splitting, and assignment of class to a non-split node [17]. J48 is a version of an earlier algorithm developed by J. Ross Quinlan, the very popular C4.5.

In pseudocode the algorithm for C4.5 is [18]:

1. Check for base cases
2. For each attribute *a*
    1. Find the normalized information gain from splitting on *a*
3. Let *a_best* be the attribute with the highest normalized information gain
4. Create a decision *node* that splits on *a_best*
5. Recur on the sublists obtained by splitting on *a_best*, and add those nodes as children of *node*

II. PROPOSED TECHNIQUE FOR INTELLIGENT HEART DISEASE PREDICTION SYSTEM

The existing approaches based on neural network could predict different risk levels [3]. While techniques based on data mining could able to extract patterns in response to the predictable state [1]. To address limitations of existing approaches a new technique based on PNN for heart disease prediction is proposed in this paper.

Probabilistic Neural Network which is a class of radial basis function (RBF) network is useful for automatic pattern recognition, nonlinear mapping and estimation of probabilities of class membership and likelihood ratios [19]. The Fig. 2 displays the architecture for a PNN that recognizes K = 2 classes, but it can be extended to any number K of classes. The input layer (on top) contains N nodes: one for each of the N input features of a feature vector. These are fan-out nodes that branch at each feature input node to all nodes in the pattern (or middle) layer so that each hidden node receives the complete input feature vector x. The hidden nodes are collected into groups: one group for each of the K classes as shown in the fig.. Each hidden node in the group for Class k corresponds to a Gaussian function centered on its associated feature vector in the k^{th} class (there is a Gaussian for each exemplar feature vector). All of the Gaussians in a class group feed their functional values to the same output layer node for that class, so there are K output nodes.

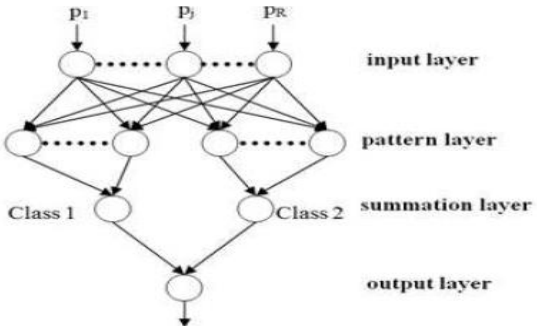

Fig.2. PNN Architecture

For this work, RBF is used as the activation function in the pattern layer. Fig. 3 shows the pattern layer of the PNN..

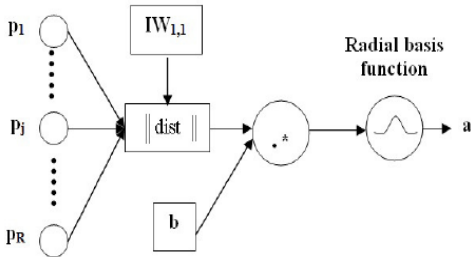

Fig.3. PNN pattern layer

The ||dist|| box shown in Fig. 3 subtracts the input weights, IW1, 1, from the input vector, p, and sums the squares of the differences to find the Euclidean distance [20]. The differences indicate how close the input is to the vectors of the training set. These elements are multiplied element by element, with the bias, b, using the dot product (.*) function and sent to the radial basis transfer function

The output a is given as:

$$a = radbas(\| IW_{1,1} - P \| b) \quad (3)$$

Where radbas is the radial basis activation function.

A. Data Acquisition & Pre-Processing

Cleaning and filtering of the data must be necessarily carried out so as to avoid the creation of deceptive or inappropriate rules or patterns. In our experiment, the heart disease data is refined by removing duplicate records and supplying missing values. A historical data is considered for the experimentation. The various 13 parameters are considered as input to the model.

B. Data Source

A total of 576 records with 13 medical attributes (factors) were obtained from the Cleveland Heart Disease database [4]. For the sake of consistency, only categorical attributes were used for all the four models.

Attribute Information:

1. Age: age in years

2. Sex: sex (1 = male; 0 = female)

3. cp: chest pain type

   -- Value 1: typical angina

   -- Value 2: atypical angina

   -- Value 3: non-anginal pain

   --Value 4: asymptomatic

4. restbps: resting blood pressure (in mm Hg on admission to the hospital)

5. chol: serum cholestoral in mg/dl #10 (trestbps)

6. fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

7. restecg: resting electrocardiographic results

   -- Value 0: normal

   -- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)

   -- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

8. thalach: maximum heart rate achieved

9. exang: exercise induced angina (1 = yes; 0 = no)

10. Oldpeak: ST depression induced by exercise relative to rest

11 slope: the slope of the peak exercise ST segment

   -- Value 1: upsloping

   -- Value 2: flat

   -- Value 3: downsloping

12. ca: number of major vessels (0-3) colored by flourosopy

13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

14. num: diagnosis of heart disease

   -- Value 1: present

   -- Value 0: not_present

Attribute Types

- Real : 1,4,5,8,10,12
- Ordered:11
- Binary: 2,6,9
- Nominal: 7,3
- Missing Values : none

IV. EVALUATION SETUP

As we need to select the most suitable methods for intelligent heart disease prediction, we have to perform an evaluation of existing and proposed approach with respect to cost of false positives and false negatives. The most suitable method for such an environment is ROCCH method. The confusion matrix serves as input for ROCCH method.

A. Confusion Matrix

In predictive analytics, a table of confusion, also known as a confusion matrix, is very useful for evaluating classifiers, as they provide an efficient snapshot of its performance displaying the distribution of correct and incorrect instances.

Table 1. Confusion matrix

		Actual value		
		p	*n*	Total
Prediction	*p'*	True Positive	False Positive	P'
	n'	False Negative	True Negative	N'
	Total	P	N	

The performance of all algorithms is evaluated by computing the percentages of Sensitivity (SE), Specificity (SP) and Accuracy (AC), the respective definitions are as follows:

$$SE = \frac{TP}{(TP + FN)} * 100 \quad (4)$$

$$SP = \frac{TN}{(TN + FN)} * 100 \quad (5)$$

$$AC = \frac{(TP + TN)}{(TN + TP + FN + FP)} * 100 \quad (6)$$

Where TP is the number of true positives,

TN is the number of true negatives,

FN is the number of false negatives, and

FP is the number of false positives

B. The ROCCH Method

The most suitable evaluation method for such an imprecise environment is the Receiver Operating Characteristic Convex Hull (ROCCH) method [21].The Receiver Operating Characteristics (ROC) analysis is a method for evaluating and comparing a classifiers performance. It has been extensively used in signal detection, and it was introduced and extended in for the Machine Learning community. In ROC analysis, instead of a single value of accuracy, a pair of values is recorded for different class and cost conditions a classifier is learned. The values recorded are the False Positive rate (FP) and the True Positive rate (TP), defined in terms of the confusion matrix as:

$$FP = \frac{fp}{(fp + tn)} \quad (7)$$

$$TP = \frac{tp}{(tp + fn)} \quad (8)$$

In this formula, fp is the number of false positives; tp is the number of true positives. Each (FP, TP) pair is plotted as a point in the ROC space.

V. RESULTS AND DISCUSSION

The Proposed PNN along with existing three approaches is trained with different number of data cases. The resulting TP rate and FP rate on training of each set of data cases are recorded. The Table 2 shows the record when system is trained with 500 data cases & tested for heart disease prediction.

The Table 2 presents the prediction and true positive, true negative, false positive, false negative values. The table summarizes the result of all three existing and proposed approach. The proposed technique Probabilistic Neural Network appears to be most effective as it has the highest number of correct prediction (94.6%) as compared to other technique.

In the Table

+WHD: Patients with heart disease

-WHD: Patients with no heart disease

+PHD: Patients Predicted as having heart disease

+PHD: Patients Predicted as having no heart disease

.

Table 2. Analysis with 500 data cases

Model Type	Prediction attribute	No. of Cases	Prediction	TP Rate	FP Rate	Correct identification (%)	Incorrect identification (%)
DT	+WHD,+PHD	186	CORRECT	.827	.145	84.20	15.80
	-WHD,+PHD	39	INCORRECT				
	-WHD,-PHD	235	CORRECT				
	+WHD,-PHD	40	INCORRECT				
NB	+WHD,+PHD	182	CORRECT	.798	.125	84.00	16.00
	-WHD,+PHD	46	INCORRECT				
	-WHD,-PHD	238	CORRECT				
	+WHD,-PHD	34	INCORRECT				
BNN	+WHD,+PHD	174	CORRECT	.763	.162	80.40	19.60
	-WHD,+PHD	54	INCORRECT				
	-WHD,-PHD	228	CORRECT				
	+WHD,-PHD	44	INCORRECT				
PNN	+WHD,+PHD	212	CORRECT	.934	.044	94.60	5.40
	-WHD,+PHD	15	INCORRECT				
	-WHD,-PHD	261	CORRECT				
	+WHD,-PHD	12	INCORRECT				

.

Fig. 4 shows the ROC curve for all four approaches for different sets of training data cases. It's a graph obtained by plotting TP versus FP rate. One point on a ROC diagram dominates another if it is above and to the left, i.e. it has a higher TP and a lower FP. Dominance implies superior performance for a variety of commonly performance measures, including Expected Cost, Weighted Accuracy, Weighted Error, recall and others.

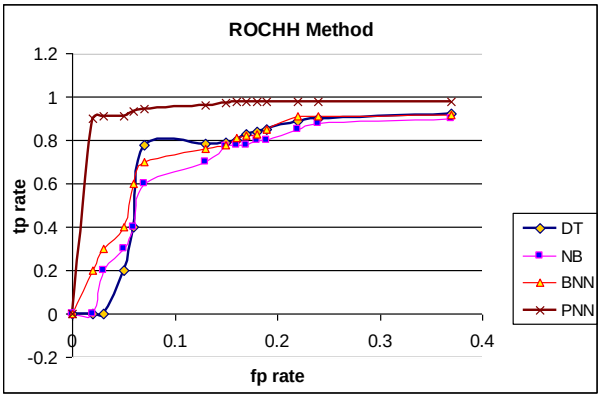

Fig.4. The ROC curve for training data

From fig. 4, we can clearly record that points of PNN dominates the points of other approaches. It's worth observing that TP rate remains constant in case of PNN when trained with enough large number of dataset. From the graph we shall conclude that PNN is better technique for heart disease prediction.

One more evaluation method is by comparing accuracy of prediction. This is done to evaluate effectiveness of the proposed network. Table 3 presents the accuracy in-terms of correctly classified instances and incorrectly classified instances.

Table 3. Classification instances of proposed & existing approaches.

Score for 76 Data cases	DT	NB	BNN	PNN
Correctly Classified Instances	83.57%	83.26%	80.57%	92.10%
Incorrectly Classified Instances	16.43%	16.74%	19.43%	7.89%

From the table we can infer that proposed Probabilistic Neural Network gives better accuracy as compared to other technique. Decision Tree and Naïve Bayes gives equal number of supporting cases followed by Back-propagation Neural Network (with difference less than 3%).

The performance of the algorithm was evaluated by computing the percentages of Sensitivity (SE), Specificity (SP). Table 4 shows the resulted SE, SP for testing data of the proposed networks.

Table 4. Model Analysis Results

No. Of Test Cases :76	Sensitivity	Specificity
DT	78.05%	76.92%
NB	76.29%	77.50%
BNN	75.61%	88.57%
PNN	92.68%	91.42%

From table it is clear that the proposed Probabilistic Neural Network has better sensitivity and specificity (92.68% and 91.42% respectively) compared to other approaches. The PNN system gives the accurate classification when compared to other methods.

VI. CONCLUSION

A prototype heart disease prediction system is developed using neural network and data mining classification modeling techniques. The system extracts hidden knowledge from a historical heart disease database. The models are trained and validated against a test dataset. The most effective model to predict patients with heart disease appeared to be the new proposed technique Probabilistic Neural Network.

During training of system, as the numbers of input data sets are increased all four approaches leads to more accuracy. It is observed that, when Decision tree trained with 100 data cases it gives more number of incorrect cases and result in more correct prediction when trained with 500 data cases. Naïve Bayes and Neural Network also result in better prediction when trained with more number of data sets. Probabilistic Neural Network varies its accuracy with less difference when trained with few or large number of data cases.

The existing approach using Back-propagation neural network for testing is less accurate compared to other existing and new proposed techniques. BNN concludes with good Specificity. The approaches based on data mining technique results in better SE and SP.

From table 3 we can clearly state that proposed method based on PNN has performed better at when patients are with no heart disease. In table 2 it shows large number (261) of supporting cases in support of the result. In table 4, for testing PNN prediction to be more accurate compared with other technique. Sensitivity and Specificity of PNN for heart disease prediction for 76 test cases is far better compared to other technique.

HDPS can be further enhanced and expanded. For example, it can incorporate other medical attribute besides the 15 listed in 3.3.1. As future scope one can also incorporate other data mining techniques, e.g., Time Series, Clustering and Association Rules. Continuous data can also be used instead of just categorical data. Another area is to use Text Mining to mine the vast amount of unstructured data available in healthcare databases. Another challenge would be to integrate text mining and data mining

VII. REFERENCES

[1] Shantakumar B. Patil, Y.S. Kumaraswamy, "Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network", European Journal of Scientific Research, ISSN 1450-216X Vol.31 No.4, pp.642-656, 2009

[2] Shantakumar B. Patil, Y.S. Kumaraswamy, "Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network", IJCSNS International Journal of Computer Science and Network Security, Vol.9 No.2, pp.228-235, 2009

[3] Sellappan Palaniappan, Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", IJCSNS International Journal of Computer Science and Network Security, Vol.8 No.8, August 2008

[4] Blake, C.L., Mertz, C.J.: "UCI Machine Learning Databases", http://mlearn.ics.uci.edu/databases/heart-disease/, 2004

[5] Anamika Gupta, Naveen Kumar, and Vasudha Bhatnagar, "Analysis of Medical Data using Data Mining and Formal Concept Analysis", Proceedings Of World Academy Of Science, Engineering And Technology,Vol. 6, June 2005

[6] S Stilou, P D Bamidis, N Maglaveras, C Pappas, "Mining association rules from clinical databases: an intelligent diagnostic process in healthcare", Stud Health Technol Inform 84: Pt 2. 1399-1403, 2001.

[7] Hian Chye Koh and Gerald Tan, "Data Mining Applications in Healthcare", Journal of healthcare information management, Vol. 19, No. 2, pp. 64-72, 2005.

[8] Frank Lemke and Johann-Adolf Mueller, "Medical data analysis using self-organizing data mining technologies," Systems Analysis Modelling Simulation, Vol. 43, No. 10, pp: 1399 - 1408, 2003.

[9] Andreeva P., M. Dimitrova and A. Gegov, "Information Representation in Cardiological Knowledge Based System", SAER'06, pp: 23-25 Sept, 2006.

[10] Ö. Galip Saracoglu, "Artificial Neural Network Approach for Prediction of Absorption Measurements of an Evanescent Field Fiber Sensor", Sensors, Vol. 8, pp. 1585-1594, 2008.

[11] Savkovic-Stevanovic, "Neural networks for process analysis and optimization: modeling and applications", Computers & chemical engineering, Vol. 18, No 11-12 (14 ref.), pp. 1149-1155, 1994.

[12] http://en.wikipedia.org/wiki/Backpropagation

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

[14] Obenshain, M.K: "Application of Data Mining Techniques to Healthcare Data", Infection Control and Hospital Epidemiology, 25(8), 690–695, 2004

[15] An approach of the Naive Bayes classifier for the document classification by Ioan Pop in General Mathematics Vol. 14, No. 4 (2006), 135–138.

[16] http://www.d.umn.edu/~padhy005/Chapter5.html

[17] Ho, T. J.: "Data Mining and Data Warehousing", Prentice Hall, 2005

[18] S.B. Kotsiantis, Supervised Machine Learning: A Review of Classification Techniques, Informatica 31(2007) 249-268, 2007

[19] An Introduction to Probabilistic Neural Networks by Vincent Cheung ,Kevin Cannons ,Signal & Data Compression Laboratory Electrical & Computer Engineering, University of Manitoba, Winnipeg, Manitoba, Canada.

[20] A Probabilistic Neural Network Approach For Protein Superfamily Classification by PV Nageswara Rao, T Uma Devi, DSVGK Kaladhar, GR Sridhar and Allam Appa Rao in Journal of Theoretical and Applied Information Technology.

[21] Content Based SMS Spam Filtering by José María Gómez Hidalgo, Guillermo Cajigas Bringas and Enrique Puertas Sánz, Proceedings of the 2006 ACM symposium on Document engineering, Pages: 107 - 114 , ISBN:1-59593-515-0, 2006.