RESEARCH ARTICLE OPEN ACCESS

Propose a Enhanced Framework for Prediction of Heart Disease

K. Sudhakar*, Dr. M. Manimekalai**
*Research Scholar, Department of Computer Applications, Shrimati Indira Gandhi College, Trichy, Tamilnadu, India
** Director and Head, Department of Computer Applications, Shrimati Indira Gandhi College, Trichy, Tamilnadu, India

ABSTRACT
Heart disease diagnosis requires more experience and it is a complex task. The Heart MRI, ECG and Stress Test etc are the numbers of medical tests are prescribed by the doctor for examining the heart disease and it is the way of tradition in the prediction of heart disease. Today world, the hidden information of the huge amount of health care data is contained by the health care industry. The effective decisions are made by means of this hidden information. For appropriate results, the advanced data mining techniques with the information which is based on the computer are used. In any empirical sciences, for the inference and categorisation, the new mathematical techniques to be used called Artificial neural networks (ANNs) it also be used to the modelling of the real neural networks. Acting, Wanting, knowing, remembering, perceiving, thinking and inferring are the nature of mental phenomena and these can be understand by using the theory of ANN. The problem of probability and induction can be arised for the inference and classification because these are the powerful instruments of ANN. In this paper, the classification techniques like Naive Bayes Classification algorithm and Artificial Neural Networks are used to classify the attributes in the given data set. The attribute filtering techniques like PCA (Principle Component Analysis) filtering and Information Gain Attribute Subset Evaluation technique for feature selection in the given data set to predict the heart disease symptoms. A new framework is proposed which is based on the above techniques, the framework will take the input dataset and fed into the feature selection techniques block, which selects any one techniques that gives the least number of attributes and then classification task is done using two algorithms, the same attributes that are selected by two classification task is taken for the prediction of heart disease. This framework consumes the time for predicting the symptoms of heart disease which make the user to know the important attributes based on the proposed framework.
Keywords **–** ANN, PCA, SVM, CFS.

I. INTRODUCTION

In the past 10 years Heart disease becomes the major cause for death all around the globe (World Health Organization 2007) [1]. To help the professionals of health care, the several data mining techniques are used by the researchers in the findings of heart disease. The European Public Health Alliance stated that heart attacks, strokes and other circulatory diseases is accounted as 41% of all deaths (European Public Health Alliance 2010) [2]. The one fifth lives of the Asian are lost due to the non communicable disease such as chronic respiratory diseases, cardiovascular diseases and cancer etc and this is descripted in the ESCAP (Economical and Social Communication of Asia and Pacific 2010) [2]. ABS (The Australian Bureau of Statistics) described that circulatory system diseases and heart diseases are the primary reason for death in Australia, causing 33.7% all deaths (Australian Bureau of Statistics 2010) [3]. The heart disease patients were motivated around the globe every year. In addition to the availability of large amount of patients' data from which to extort useful knowledge, in the diagnosis of heart disease, for facilitating the professionals of health care, the data mining techniques have been used by the researchers [4]. Nowadays, data mining is the exploration of large datasets to extort hidden and formerly unknown patterns, relationships and knowledge that are complicated to detect with conventional statistical methods. In the emerging field of healthcare data mining plays a major role to extract the details for the deeper understanding of the medical data in the providing of prognosis [5]. Due to the development of modern technology, data mining applications in healthcare consist about the analysis of health care centres for enhancement of health policy-making and prevention of hospital errors, early detection, prevention of diseases and preventable hospital deaths, more value for money and cost savings, and detection of fraudulent insurance claims.

The characteristic selection has been an energetic and productive in the field of research area through pattern recognition, machine learning, statistics and data mining communities. The main intention of attribute selection is to choose a subset of input variables by eradicating features, which are irrelevant or of non prognostic information. Feature selection

[6] has proven in both theory and practice to be valuable in ornamental learning efficiency, escalating analytical accuracy and reducing complexity of well-read results. Feature selection in administered learning has a chief objective in finding a feature subset that fabricates higher classification accuracy. The number of feature N increases because the expansions of the domain dimensionality. Among that finding an optimal feature subset is intractable and exertions associated feature selections have been demonstrated to be NP-hard. At this point, it is crucial to depict the traditional feature selection process, which consists of four basic steps, namely, validation of the subset, stopping criterion, evaluation of subset and subset generation. Subset generation is a investigation process that generates the candidate feature subsets for assessment based on a certain search strategy. Depends on the certain assessment, the comparison with the best prior one and each candidate subset is evaluated. If the new subset revolves to be better, it reinstates best one. Whenever the stopping condition is fulfilled until the process is repeated. There are large number of features that can exceed the number of data themselves often exemplifies the data used in ML [7]. This kind of problem is known as "the curse of dimensionality" generates a challenge for a mixture of ML applications for decision support. This can amplify the risk of taking into account correlated or redundant attributes which can lead to lower classification accuracy. As a result, the process of eliminating irrelevant features is a crucial phase for designing the decision support systems with high accuracy.

In this technical world, data mining is the only consistent source accessible to unravel the intricacy of congregated data. Meanwhile, the two categories of data mining tasks can be generally categorized such as descriptive and predictive. Descriptive mining tasks illustrate the common attributes of the data in the database. Predictive mining tasks execute implication on the present data in order to formulate the predictions. Data available for mining is raw data. The data collects from different source, therefore the format may be different. Moreover, it may consist of noisy data, irrelevant attributes, missing data etc. Discretization – Once the data mining algorithm cannot cope with continuous attributes, discretization [8] needs to be employed. However, this step consists of transforming a continuous attribute into an unconditional attribute, taking only a small number of isolated values. Frequently, Discretization often improves the comprehensibility of the discovered knowledge. Attribute Selection – not all attributes are relevant and so for selecting a subset of attributes relevant for mining, among all original attributes, attribute selection [9] is mandatory.

II. CLASSIFICATION TECHNQIUES

The most commonly used data mining technique is the classification that occupies a set of pre-classified patterns to develop a model that can categorize the population of records at large. The learning and classification is involved by the process called the data classification. By the classification algorithm, the training data are analyzed in the learning [4] [10]. The approximation of the classification rules, the test data are used in the classification. To the new data tuples, the rules can be applied when the accuracy is acceptable. To verify the set of parameters which is needed for the proper discrimination, the pre-classified examples are used in the classifier-training algorithm. The model which is called as a classifier, only after these parameters encoded by the algorithm. Artificial neural network, Naive Bayesian classification algorithm classification techniques are used in this paper.

1. Artificial Neural Network

Moreover, the realistic provisional terms in neural networks are non-linear statistical data modelling tools. For discovering the patterns or modelling the complex relationships between inputs and outputs the neural network can be used. The process of collecting information from datasets is the data warehousing firms also known as data mining by using neural network tool [11]. The more informed decisions are made by users that helping data of the cross-fertilization and there is distinction between these data warehouse and ordinary databases and there is an authentic manipulation.

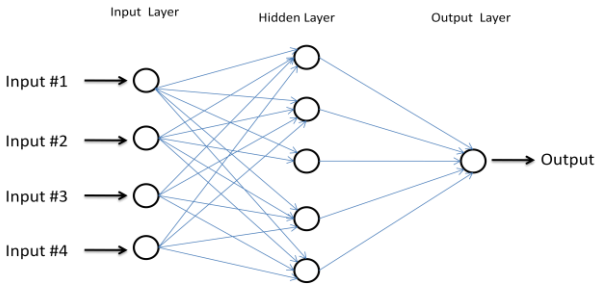

Figure 1: Example ANN

Among the algorithms the most popular neural network algorithms are Hopfield, Multilayer perception, counter propagation networks, radial basis function and self organizing maps etc. In which, the feed forward neural network was the first and simplest type of artificial neural network consists of 3 units input layer, hidden layer and output layer. There are no cycles or loops in this network. A neural network [12] has to be configured to fabricate the required set of outputs. Basically there are three learning conditions for neural network. 1) Supervised Learning, 2) Unsupervised Learning, 3) Reinforcement learning the perception is the basic unit of an artificial neural network used for

classification where patterns are linearly separable. The basic model of neuron used in perception is the Mcculluchpitts model. The learning for artificial neural networks are as follows:

Step 1: Ret D= {{Xi,Yi}/i=1, 2, 3----n}} be the set of training example.

Step 2: Initialize the weight vector with random value,W (o).

Step 3: Repeat.

Step 4:For each training sample (Xi, Yi) ∈ D.

Step 5: Compute the predicted output Yi^ (k)

Step 6: For each weight we does.

Step 7: Update the weight we (k+1) = Wj(k) +(y i - yi^ (k))xij.

Step 8: End for.

Step 9: End for.

Step 10: Until stopping criteria is met.

2. Naïve Bayesian Classification Technique

The Bayesian Classification signifies a supervised learning method and also as a statistical method for classification [11]. In which it assumes a fundamental probabilistic model and it allocates us to incarcerate uncertainty about the model in a ethical way by determining probabilities of the results. And also it can unravel diagnostic and predictive problems. The Bayesian theorem as follows:

Given training data Y, posterior probability of a hypothes32is I, R(I|Y), follows the Bayes theorem R(I|Y)=R(Y|I)R(I)/R(Y).

Algorithm:

The Naive Bayes algorithm is based on Bayesian theorem as given by above equation.

Step 1: Each data sample is represented by an n dimensional feature vector, S = (s1, s2….. sm), depicting n measurements made on the sample from n attributes, respectively S1, S2, Sn.

Step 2: Suppose that there are m classes, T1, T2……Tk. Given an unknown data sample, Y (i.e., having no class label), the classifier will predict that Y belongs to the class having the highest posterior probability, conditioned if and only if:

$R(Tj/Y)>R(T_l/Y)$ for all $1<=l<=k$ and $l!=j$

R(Tj|Y) is maximized then. R(Tj|Y) is maximized for the class Tj is called the maximum posterior hypothesis. By Bayes theorem,

Step 3: Only $R(Y|T_l)R(Tj)$ need be maximized when R(Y) is constant for all classes. It is assumed that the classes are equally likely when the class prior probabilities are not known, i.e. R(T1) = R(T2) = …..= $R(T_k)$, and consequently we would maximize R(Y|Tj). or else, we maximize R(Y|Tj)R(Tj). Note that the class prior probabilities may be estimated by R(Tj) = aj/a , where Aj is the number of training samples of class Tj, and a is the total number of training samples on Y. That is, the naive probability assigns an unknown sample Y to the class Tj.

III. FEATURE SELECTION TECHNIQUES

In general, Feature subset selection is a pre-processing step used in machine learning [13]. It is valuable in reducing dimensionality and eradicates irrelevant data therefore it increases the learning accuracy. It refers to the problem of identifying those features that are useful in predicting class. Features can be discrete, continuous or nominal. On the whole, features are described in three types.

1)Relevant 2) Irrelevant3) Redundant. Feature selection methods wrapper and embedded models. Moreover, Filter model rely on analyzing the general qualities of data and evaluating features and will not involve any learning algorithm, where as wrapper model uses après determined learning algorithm and use learning algorithms performance on the provided features in the evaluation step to identify relevant feature. The Embedded models integrate the feature selection as a part of the model training process.

The collection of datas from medical sources is highly voluminous in nature. The various significant factors distress the success of data mining on medical data. If the data is irrelevant, redundant then knowledge discovery during training phase is more difficult. Figure 2 shows flow of FSS.

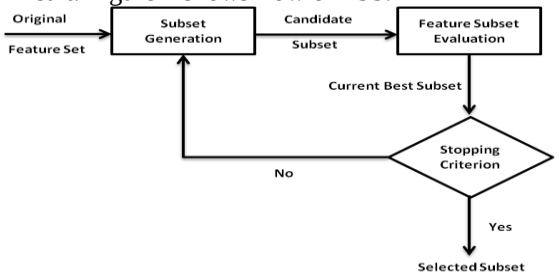

Figure 2: Feature Subset Selection

1. Principle Component Analysis (PCA) Feature Selection Technique

The Principle component analysis (PCA) [14] is a stagnant technique used in many applications like face recognition, pattern recognition, image compression and data mining. Further PCA is used to shrink the dimensionality of the data consisting of a large no. of attributes. PCA can be generalized as multiple factor analysis and as correspondence analysis to handle heterogeneous sets of variables and quantitative variables respectively. The following are the main procedure for the principle component analysis (PCA). Scientifically PCA depends on SVD of rectangular matrices and Eigen decomposition of positive semi definite matrices.

Step 1: Obtain the input matrix

Step 2: Subtract the mean from the data set in all dimensions

Step 3: Calculate covariance matrix of this mean subtracted data set.

Step 4: Calculate the Eigen values and Eigen Vector from covariance matrix
Step 5: Form a feature vector
Step 6: Derive the new data set.

2. Information Gain Attribute Subset Feature Selection Technique

In this method, the discernibility function is used [12]. The discernibility function is given as follows: For an information system (I,H), s discernibility function DF is a boolean function of m Boolean variables e1,e2……..en corresponding to the attributes e1,e2.....en respectively, and defined as follows: DF(e1,e2….en)=S1 ∧ S2 ∧ …..Sm where ej∈ S. The proposed algorithm for the information gain attribute subset evaluation is defined as below:
Step 1: For selected dataset, the discernibility matrix can be computed.
Using P[J,I]={ e ∈ E, where X[J]≠X[I] and Y[J]≠Y[I]} J,I=1,2,….m Eq1 Where X are conditional attributes and Y is a decision attribute. This discernibility matrix P is symmetric. Where P[a,b]=P[b,a] and P[a,a]=0. Due to this, the consideration of the lower triangle or upper triangle of the matrix is sufficient.
Step 2: For the discernibility matrix, the discernibility function is to be computed P[a,b] by using DF(a) = ∧ {∧ P[a,b] / a,b ∈ I; P[a,b] ≠0} Eq2
Step 3: On applying the expansion law on the attribute selection at least two numbers which belongs to the large number of conjunctive sets.
Step 4: For each component, the expansion law cannot be applied until repeat steps 1 to 3.
Step 5: For their corresponding attributes, all strongly equivalent classes are substituted.
Step 6: By using the formula, for the simplified discernibility function that containing the attributes, the information gain is to be calculated.
Gain(Gi) = F(Ri) - F(Gi) Eq.3 Where
F(G) $=\sum_{j=1}^{m} \mathrm{Rj} \log_2 Rj$ ----- (4)
$\sum_{j=1}^{m} \mathrm{Rj} \log_2 Rj$ = $-\frac{r1}{r}\log_2\frac{r1}{r} - \frac{r2}{r}\log_2\frac{r2}{r} \ldots \frac{rm}{r}\log_2\frac{rm}{r}$ ---------- (5)
Where Rj is the ratio of conditional attribute R in dataset. When Gi has | Gi | kinds of attribute values and condition attribute Rj partitions set R using attribute Gi, the value of information F(Gi) is defined as
F(Gi) = $\sum_{i=1}^{Gi} Ki * F(Bi)$ -------- (6)
Step 7: The attribute with the least gain value is to be remove from the discernbility function, where as, the attribute with the highest value is to be added to the reduction set. Until, the discernbility function reaches null set, goto step 6.

IV. PROPOSED FRAMEWORK

The following figure represents the framework of our paper for predicting the heart disease of the patient earlier using some important tests (attributes) by means of using artificial neural network and feature selection techniques.

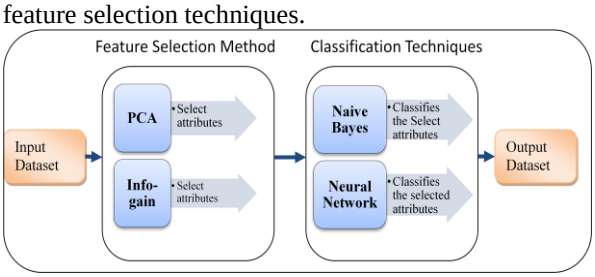

Figure 3: Proposed Framework

In this framework, the input dataset is feed into the feature selection method block where the feature selection is done according to the given data set, and it will takes the method which gives the reduced number of selected attributes from the given number of attributes. And that selected attributes is given to the classification techniques, and ANN is classify the selected attributes. And that resultant dataset is generated for predicting the heart disease in easy manner and in earlier stage. This proposed framework reduces the task of clustering, the grouping of same attributes takes place by using our framework for predicting the heart disease using given heart disease data set.

V. DATASET AND TOOL

The data used in this study is the Hungarian institute of cardiology. The dataset contains 303 instances and 14 attributes of the heart disease patient [15]. The general purpose machine learning and data mining tool is an Orange (http://orange.biolab.si). It features a multilayer architecture suitable for different kinds of users, from inexperienced data mining beginners to programmers who prefer to access the tool through its scripting interface. In the paper we outline the history of Orange's development and present its current state, achievements and the future challenges. The following are the attributes of the given heart disease dataset.

S. No	Attribute
1	Age
2	Gender
3	Chest Pain
4	Rest SBP
5	Cholestrol
6	Fasting Blood
7	Rest ECG
8	Maximum HR
9	Exer Ind
10	ST by exercise
11	Slope peak exc ST
12	Major vessels colored
13	Thal
14	Diameter Narrowing

Table 1: The attributes of the heart disease dataset

VI. EXPERIMENTAL RESULT AND ANLYSIS

The dataset is fed into the orange tool for the execution of the above framework without using cluster technique. The attribute filtering is done for the given heart disease dataset. The dataset contain 14 attributes and 303 instances.

Feature Selection Method	**Total Number of Attributes**	**Number of Filtered Attributes**
Information Gain Method	14	8
PCA	14	6

Table 2: Filtered Attributes using feature selection method

Attribute Filtering Method	**Filtered Attributes**	**Our framework result**
Information gain	8 (5,8,10,13,3,12,4,9)	4 (3,9,12,13)
PCA	6 (3,7,9,11,12,13)	

Table 3: Result of our framework using filtered attributes

S. No	**Classifiers**	**Correctly Classified Instances**	**Incorr-ectly Classi-fied Instances**	**Accuracy of the classify-cation**
1	SVM	232	71	76.45%
2	ANN	254	49	83.70%
3	Classification Tree	228	75	75.25%
4	Naïve Bayes	251	52	82.75%

Table 4: Classifiers accuracy with full dataset

Attribute Filtering Method	**Number of attributes filtered**
CFS	3 (3,5,4)
Information Gain	10 (3,5,21,4,15,19,11,30,29,26)
Gain ratio	9 (3,5,2,21,4,15,19,11,6)
PCA	11 (3,5,21,4,15,19,2,11,30,23,29)

Table 5: Number of filtered attributes by each attribute filtering method From the above table 2, the attribute filtering method of PCA selects the 6 attributes and Information gain filters 8 attributes out of 14 attributes. Table 3 gives the result of our framework which clusters the same attributes from the attribute filtering methods of PCA and information. From the table 4, the classifiers accuracy with the full dataset is more for ANN than SVM, classification tree and Naïve Bayes algorithm. The classification accuracy for ANN is 83.70 % and Naïve Bayesian Classification is 82.75% than the other methods. And from the table 5, the Information gain and PCA attribute algorithm gives more reduced attributes than the CFS and Gain ratio attribute filtering methods.

VII. CONCLUSION

Classification methods are most common and efficient one for predicting the heart disease when the dataset contains the large number of attributes. And feature selection is used to filtered the attributes on their importance for the prediction. In our proposed framework we are using PCA and Information gain attribute filtering techniques, because these two methods gives more number of attributes than the others. And from the experimental result we can conclude that the ANN and Naïve Bayes give more classification accuracy than the others. And our framework clusters the same attributes which are resultant from the filtering methods and automatically it reduces the task of clusters. So, in our framework we are considering the ANN and Naïve Bayes classification methods, PCA and information attribute filtering methods are best for the predicting heart disease using heart disease dataset.

REFERENCES

[1] "*A Safe Future- Global Public Health Security in the 21st Century*", *The World Health Report 2007*.

[2] "Commission Staff Working Document", *European Commission, Implementation of Health Programme 2010*.

[3] "*Drugs in Australia 2010-Tobacco, Alcohol and other drugs*", *Australian Institute of Health and Welfare*.

[4] Vikas Chaurasia, Saurabh Pal, "*Early Prediction of Heart Diseases Using Data Mining Techniques*", *Carib.j. Sci Tech, 2013, Vol.1*, pp.no:208-217.

[5] Hian Chye Koh and Gerald Tan, "Data Mining Applications in Healthcare", *Journal of Healthcare Information Management — Vol. 19, No. 2*, pp.no: 64-72.

[6] S. Saravana kumar, S.Rinesh, "*Effective Heart Disease Prediction using Frequent Feature Selection Method", International Journal of Innovative Research in Computer and Communication Engineering, Vol.2, Special Issue 1, March 2014*, pp.no: 2767-2774.

[7] Jyoti Soni, Uzma Ansari, Dipesh Sharma and Sunita Soni, "*Intelligent and Effective*

Heart Disease Prediction System using Weighted Associative Classifiers", *International Journal on Computer Science and Engineering (IJCSE), Vol. 3 No. 6 June 2011*, pp.no:2385-2392.

[8] Aieman Quadir Siddique, Md. Saddam Hossain, "*Predicting Heart-disease from Medical Data by Applying Naïve Bayes and Apriori Algorithm*", *International Journal of Scientific & Engineering Research, Volume 4, Issue 10, October-2013*, pp.no: 224-231.

[9] Priyanka Palod, Jayesh Gangrade, "*Efficient Model For Chd Using Association Rule With Fds*", *International Journal of Innovative Research in Science, Engineering and Technology, Vol. 2, Issue 6, June 2013*, pp.no:2157-2162.

[10] Shamsher Bahadur Patel, Pramod Kumar Yadav, Dr. D. P.Shukla, "*Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques*", *IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS), Volume 4, Issue 2 (Jul. - Aug. 2013)*, PP 61-64.

[11] D Ratnam, P Hima Bindu, V. Mallik Sai, S.P. Rama Devi, P. Raghavendra Rao, "*Computer-Based Clinical Decision Support System for Prediction of Heart Diseases Using Naïve Bayes Algorithm*", *International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014*, pp.no: 2384-2388.

[12] Selvakumar.P, DR. Rajagopalan. S.P, "*A Survey On Neural Network Models For Heart Disease Prediction*", *Journal of Theoretical and Applied Information Technology, 20th September 2014. Vol. 67 No.2*, pp.no:485-497.

[13] M. Anbarasi, E. Anupriya, N.CH.S. N. Iyengar, "*Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm*", *International Journal of Engineering Science and Technology, Vol. 2(10), 2010*, 5370-5376.

[14] Negar Ziasabounchi and Iman N. Askerzade, "*A Comparative Study of Heart Disease Prediction Based on Principal Component Analysis and Clustering Methods*", *Turkish Journal of Mathematics and Computer Science*, pp.no:1-11.

[15] Heart Disease Dataset source- ftp://ftp.ncbi.nlm.nih.gov/geo/datasets/