IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 2, Ver. X (Mar-Apr. 2014), PP 32-37
www.iosrjournals.org

A Heart Disease Risk Prediction System Based On Novel Technique Stratified Sampling

Lalita Sharma[1], Vineet Khanna[2]
[1]*Computer Science, RCEW, Rajasthan Technical University, INDIA*
[2]*Computer Science, RCEW, Rajasthan Technical University, INDIA*

Abstract : *Medical decision support systems are designed to support clinicians in their diagnosis. The prediction of heart disease pattern with classification algorithms is proposed in this paper. It is essential to find the best fit feature selection algorithm that has greater accuracy on classification in the case of heart disease classification. By using feature selection method the dimensionality of the data is reduced. In this paper, a novel feature selection method based on correlation based feature selection (CFS) is proposed. In this method, filters and wrappers were combined to eliminate the noise and redundancies in gene expression data. And the experimental results show that the method can gain the better performance in comparison with corresponding approaches.*

Key words: *filter approach, wrapper approach, CFS, CFS-SS, Naïve Bayesian*

I. INTRODUCTION

A novel technique to develop the multi-parametric feature with linear and nonlinear characteristics of HRV (Heart Rate Variability) was proposed by Heon Gyu Lee et al. [2]. Statistical and classification techniques were utilized to develop the multi- parametric feature of HRV. Besides, they have assessed the linear and the non-linear properties of HRV for three recumbent positions, to be precise the supine, left lateral and right lateral position. Numerous experiments were conducted by them on linear and nonlinear characteristics of HRV indices to assess several classifiers such as Bayesian classifiers [3], CMAR (Classification based on Multiple Association Rules) [3], C4.5 (Decision Tree) [4] and SVM (Support Vector Machine) [5]. SVM surmounted the other classifiers.

A model Intelligent Heart Disease Prediction System (IHDPS) built with the aid of data mining techniques like DT, NB and NN was proposed by Sellappan Palaniappan et al. [6]. The results illustrated the peculiar strength of each of the methodologies in comprehending the objectives of the specified mining objectives.

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. A classification task begins with a data set in which the class assignments are known. Classification models are tested by comparing the predicted values to known target values in a set of test data. The historical data for a classification is typically divided into two data sets: one for building the model; the other for testing the model.

By using our proposed system i.e. Correlation Feature Stratified Sampling (CFS-SS) the attributes are grouped together into homogenous sub groups, before sampling the strata will be mutually exclusive, every attribute will be assigned to only one stratum. The original dataset is given to the existing system. The output of the system will be the efficiency achieved without stratified sampling. Stratified sampling of all the sub sets are put together in such a manner that subset of same group size will be in one group. The efficiency of proposed system (CFS)[1] is better than existing system (CFS-SS).

1.1 Data mining techniques:-

Data mining is a process of discovering interesting knowledge such as patters, associations, changes, and significant structure from a large amount of data stored in database, data Warehouses or other information repositories. Due to wide availability of data in electronic forms, and the imminent need to convert this data into useful information and knowledge for broad applications.Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making. The steps identified in extracting knowledge from data are:

Fig.1 the steps of extracting knowledge from data

1.2 Process of feature selection:-

Feature selection algorithms perform search through the space of feature subsets, and, as a consequence, must address four basic issues affecting the nature of the search [8].

Starting point:- Selecting a point in the feature sub set space from which to begin the search can affect the direction of the search. One option is to begin with no features and successively add attributes. In this case, the search is said to proceed forward through the search space. Conversely, the search can begin with all features and successively remove them. In this case, the search proceeds backward through the search space. Another alternative is to begin somewhere in the middle and move outwards from this point.

Search organization: - An exhaustive search of the feature sub space is prohibitive for all but a small initial number of features. With N initial features there exists 2N possible subsets. Heuristic search strategies are more feasible than exhaustive ones and can give good results, although they do not guarantee finding the optimal subset.

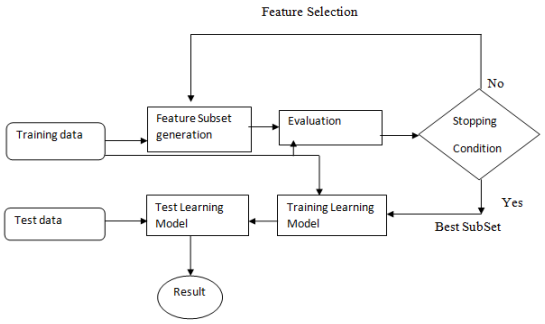

Fig. 2. Feature Selection Process

Evaluation strategy:- How feature subsets are evaluated is the single biggest differentiating factor among feature selection algorithms form chine learning. One paradigm, dubbed the filter operates independent of any learning algorithm—undesirable features are filtered out of the data before learning begins. These algorithms use heuristics based on general characteristics of the data to evaluate the merit of feature subsets. This method, called the wrapper uses an induction algorithm along with a statistical re-sampling technique such as cross-validation to estimate the final accuracy of feature subsets. Figure 3 and 4 illustrates the filter and wrapper approach of feature selection respectively.

Stopping criterion:- A feature selector must decide when to stop searching through the space of feature subsets. Depending on the evaluation strategy, a feature selector might stop adding or removing features when one of the alternatives improves upon the merit of a current feature subset. Alternatively, the algorithm might continue over the feature subset as long as the merit does not degrade. A further option could be to continue generating feature subsets until reaching the opposite end of the search space and then select the best.

1.3 Types of feature selection:-

There are two types of feature selection algorithms one is filter approach and other one is wrapper approach.

The filter approach is faster than wrapper approach though results of filter method are less accurate than wrapper [4].

1.3.1 Filter approach feature selection:-

The earliest approach to feature selection in machine learning algorithm was filter method. The filter approach actually precedes the actual classification process. The filter approach, is independent of the learning induction algorithm, computationally simple fast and scalable. The filer method uses the intrinsic prosperities of data and the target class to be learned for feature selection. Filter methods use a proxy measure instead of the error rate to score a feature subset. Using filter method, feature selection is done one and then can be provided as input to different classifiers. Filters, evaluating the features according to the heuristic function based on general characteristics of the data; and wrappers, evaluating the features using the characteristics of the data joint with the learning algorithm.

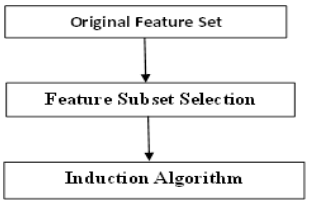

Fig. 3 Filter Approach for feature Selection

1.3.2 Wrapper approach for feature selection:-

Wrapper model approach uses the method of classification itself to measure the importance of features set; hence the feature selected depends on the classifier model used. Wrapper methods use a predictive model to score feature subsets. Each new subset is used to train a model, which is tested on a hold-out set. Counting the number of mistakes made on that hold-out set (the error rate of the model) gives the score for that subset. As wrapper methods train a new model for each subset, they are very computationally intensive, but usually provide the best performing feature set for that particular type of model. Wrapper methods generally result in better performance than filter methods because the feature selection process is optimized for the classification algorithm to be used. However, wrapper methods are too expensive for large dimensional database in terms of computational complexity and time required, and hence they are much slower than filter approach. The working of wrapper approach is shown in figure 4

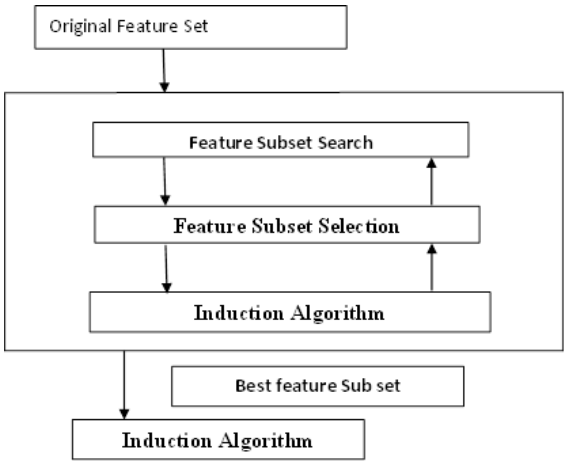

Fig. 4 Wrapper Approach for feature Selection

1.4. Potential applications:-

The data mining applications helps healthcare organization to take right decisions by using hidden knowledge from the huge amount of medical data. The figure No. 5 illustrates how the data from the traditional clinic and web based health care systems can be used to extract knowledge by applying web mining and data mining techniques which further helps the doctors and patients to make decisions.

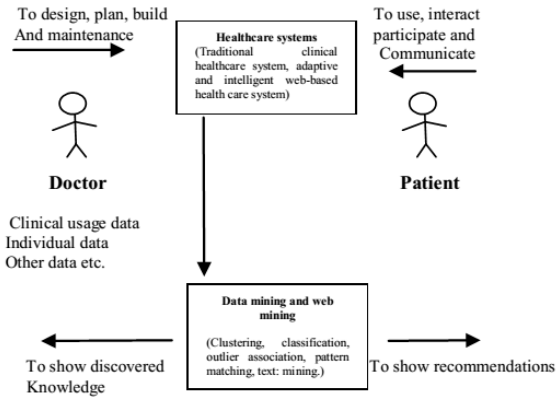

Fig. 5 The cycle of applying web mining or data mining in health care system

II. Proposed Method

2.1 Data source:-

For this study we have used dataset taken from UCI medical site which provides datasets for study purposes. The 13 attributes along with their descriptions are as follows [8]: For simplicity, categorical attributes were used for all models. The attribute ―Diagnosis‖ was identified as the predictable attribute with value ―1‖ for patients with heart disease and value ―0‖ for patients with no heart disease. The attribute ―Patient_ID‖ was used as the key; the rest are input attributes.

Description of attributes :

Predictable attribute

1. Diagnosis (value 0: < 50% diameter narrowing (no heart disease); value 1: > 50% diameter narrowing (has heart disease))

Key attribute

1. PatientID – Patient's identification number

Input attributes

1. Sex (value 1: Male; value 0 : Female)
2. Chest Pain Type (value 1: typical type 1 angina, value 2: typical type angina, value 3: non-angina pain; value 4: asymptomatic)
3. Fasting Blood Sugar (value 1: > 120 mg/dl; value 0: < 120 mg/dl)
4. Resteeg – resting electrographic results (value 0: normal; value 1: 1 having ST-T wave abnormality; value 2: showing probable or definite left ventricular hypertrophy)
5. Exang – exercise induced angina (value 1: yes; value 0: no)
6. Slope – the slope of the peak exercise ST segment (value 1: unsloping; value 2: flat; value 3: downsloping)
7. CA – number of major vessels colored by flourosopy (value 0 – 3)
8. Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)
9. Trest Blood Pressure (mm Hg on admission to the hospital)
10. Serum Cholesterol (mg/dl)
11. Thalach – maximum heart rate achieved
12. Oldpeak – ST depression induced by exercise relative to rest
13. Age in Year

Fig. 6. Description of attributes

2.2 Workflow of CFS :-

Step 1: CFS first calculates the Entropy

Entropy $H(X) = \sum j^{n}=1\ p_j \log_2 p_j$

Step 2 : Calculate Information Gain (IG)

Information Gain = IG(Y|X) = H(Y) – H(Y|X)

H(Y|X)= Conditional entropy

Step 3: Calculate Symmetric Uncertainty

Symmetric Uncertainty = (2*gain(X1,X2))/(H(X1)*H(X2)

Step 4: Step 1 to step 3 will be repeated till a proper sub set achieved. If expanding a subset results in no improvement, the search drops back to the next best unexpanded subset and continues from there

Step 5: This reduced dataset t will be given to classifiers such as C4.5 and naïve bayes

Step 6: Calculation of Naive Bayes Algorithm

Compute p(1): Probability person does not have an heart diseases.

p(2): Probability of person having heart diseases.

Compute p (a1,a2…..an|1): Probability of all attributes of test data for p(1)\

Compute p (a1,a2…..an|2): Probability of all attributes of test data for p(2)

Calculate max {p(a1,a2,…..an|1)*p(1),p(p(a1,a2,…..an|2)*p(2)}

Output is generated.

Step 7: Comparing results of both algorithm we found that the C4.5 result is less accurate than Naïve bayes

2.3 Workflow of proposed System:-

In the proposed System we are using CFS-SS as a feature selection algorithm.
Following are the steps for proposed system:-

Step 1: The input to the CFS-SS will be the output of CFS. As we know from CFS we have got 3 attributes .this 3 attributes will be given as input to CFS-SS.
Step 2: As we have 3 attributes we will have 2^3 subsets.
Step 3: accuracy of each subset is calculated
Step 4: if accuracy < max_accuracy then
Accuracy = max _accuracy
Step 5: Step 4 continues till accuracy of the entire attribute is calculated.
Step 6: The output of CFS-SS is given to Naive Bayes algorithm which calculated the final prediction is similar manner as in CFS.

III. Result analysis

3.1 Graphical comparison of accuracy of CFS and CFS-SS:-

Finally we show the performance of our system using a graphical representation. It is a 2 D graph which compares the accuracy of the various methods which have been used by in this study. From graph we can conclude that CFS is a very effective feature selection algorithm as after using it the accuracy has been increased by 10 % The modification done to the existing CFS which is our proposed system (CFS-SS) is proved to even better than CFS as it has added more to the accuracy.

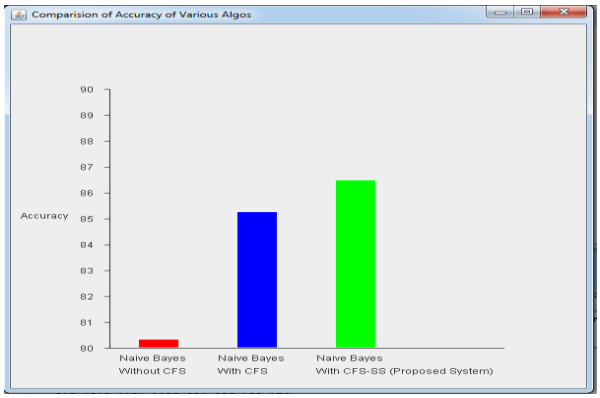

Fig.7 Graphical Representation of Accuracy of our System

3.2 Main GUI:-

The prediction phase is the main phase which decided the output whether a patient has a heart diseases or not. As shown in figure the result shows "risk predict =>low" risk by inserting 13 attributes in the application form of heart disease risk prediction system.

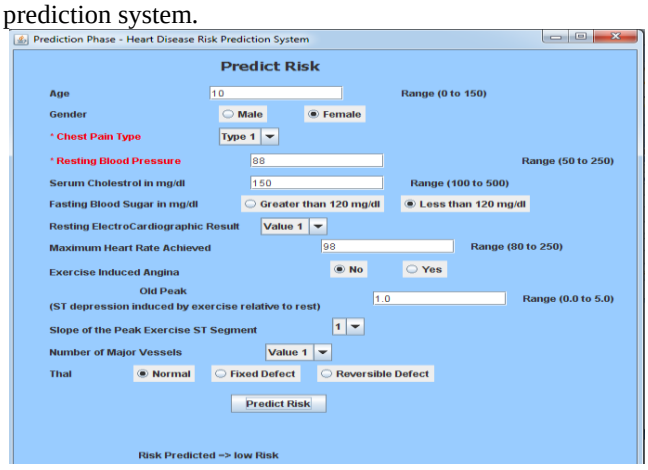

Fig. 8 Prediction Phase with all attributes of System

3.3 Improved GUI:-

As shown in figure the result shows "risk predict =>low" risk by inserting 2 attributes in the application form of heart disease risk prediction system. As the main concept of our project is to reduce the features required to predict the result correctly we will required only two attribute values i.e chest pain type and resting blood pressure. This GUI shows that after giving only this values result is predicted properly.

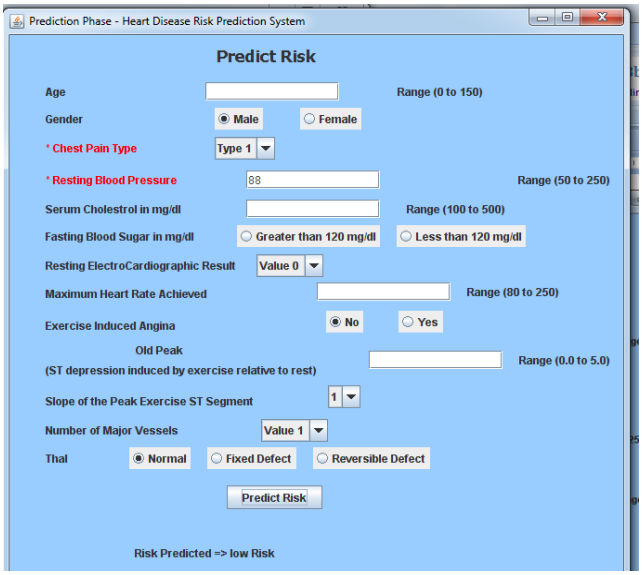

Fig. 9 Prediction Phase with 2 attributes of System.

IV. Conclusion:-

This paper proposes a methods to investigate the performance of different classification algorithm NB on heart disease dataset. The heart disease prediction is useful for cardiovascular clinicians which contains the patient's records. This patient's record is classified and predicted who are having the heart diseases. We have studies various classification algorithms. After that we found that NB gives the better accuracy than other classifiers. The dataset have the large volume of data which consumes more time for classification. Thereby we have reduced the dimensionality of data using the attribute selection methods. Then the reduced data is classified using CFS and CFS-SS feature selection algorithms. We found that NB classifier gives the better accuracy for heart disease prediction after applying the CFS-SS feature selection method.

References:-

[1] John Peter "An Empirical Study On Prediction Of Heart Disease Using Classification Data Mining Techniques" IEEE-International Conference On Advances In Engineering, Science And Management PP. 514-517(ICAESM -2012) March 30, 31, 2012.

[2] Shweta Kharya "Using Data Mining Techniques For Diagnosis And Prognosis Of Cancer Disease Abdelghani Bellaachia, Erhan Guven" International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012

[3] Li, W., Han, 1., Pei, 1.: "CMAR: Accurate and Efficient Classification Based on Multiple Association Rules". In: Proc. of 2001 International Conference on Data Mining, 2001

[4] Jyoti Soni, Ujma Ansari, Dipesh Sharma "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction" International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March 2011.

[5] Abdelghani Bellaachia,Erhan Guven "Predicting Breast Cancer Surviability Using data Mining techniques" Software technology and Engineering (ICSTE),2010 2nd international Conference on 3-5 Oct,2010.

[6] 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.

[7] Feature Selection Algorithms: A Survey and Experimental Evaluation Luis Carlos Molina, Lluís Belanche, Àngela Nebot Jordi Girona 1-3, Campus Nord C6, 08034, Barcelona, Spain. {lcmolina,belanche,angela}@lsi.upc pp1-5

[8] Asha Gowda Karegowda, M.A.Jayaram, A.S .Manjunath" Feature Subset Selection using Cascaded GA & CFS: A Filter Approach in Supervised Learning" international conference june-2011