ISSN:0975-9646

A stratified sampling technique based on correlation feature selection method for heart disease risk prediction system

Lalita Sharma , Vineet Khanna

Computer Science,
Rajasthan College of Engineering for Women, Jaipur,
Rajasthan Technical University

***Abstract*- In medical, data mining method can be utilized by the physicians to improve clinical diagnosis. In this paper a stratified approach named Correlation Feature Selection Stratified Sampling (CFS-SS) has been introduced. This method is applied to medical diagnosis heart disease risk prediction system. By using this proposed system 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 filter Correlation based feature selection (CFS) system. The output of the system will be the efficiently 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 by using CFS-SS. The efficiency of proposed system (CFS-SS) is better than existing system (CFS).**

***Key words*- CFS, CFS-SS, SVM, NB & DT.**

I. INTRODUCTION

Heart diseases encompass diverse diseases that affect the heart. Coronary heart diseases, cardio myopathy and cardiovascular diseases are some of the categories of heart diseases. The term "cardiovascular diseases" includes a wide range of conditions that affects the heart and blood vessels and the manner in which blood is pumped. Narrowing of the coronary arteries results in the reduction of blood and oxygen supply to the heart and leads to the Coronary heart diseases. Chest pain arises when the blood received by heart muscles is inadequate. Data mining technology provides a user- oriented approach to novel and hidden patterns in the data. The discovered knowledge can be used by healthcare administrators to improve the quality of service. Medical data mining is the search for relationships and patterns within the medical data that could provide useful knowledge for effective medical diagnosis. Extracting useful information from these data bases can lead to discovery of rules for later diagnosis tools. Generally medical data bases are highly voluminous in nature. If a training data set contains irrelevant and redundant features classification may produce less accurate results and search for an optimal subset would be highly expensive especially when the number of data classes increases [1]. Feature selection as a pre-processing step in used to reduce dimensionality, removing irrelevant data and increasing accuracy and improves comprehensibility. Feature selection is categorise in parts first is filter and another is wrapper. 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. Filter is faster than wrapper approach though results of filter method are less accurate than wrapper [2]. In this paper method is used is combination of filter and wrappers methods. CFS is filter and CFS-SS is wrapper. Correlation-based feature selection (CFS) is an effective feature selection method, and the set of features mostly related to some class can be selected from the gene expression data. However, in CFS the features are selected only by calculating the correlation between features and classes, features and features. It does not take into account the characters of various classifiers. When the selected features are applied for heart diseases diagnosis it cannot achieve a satisfying performance. So a Stratified Sampling feature selection method (CFS-SS) based on CFS is used in this study.

II. LITERATURE SURVEY

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. [3]. 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 [4], CMAR (Classification based on Multiple Association Rules) [4], C4.5 (Decision Tree) [5] and SVM (Support Vector Machine) [6]. 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. [7]. 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 such as patterns, relationships amid medical factors connected with

heart disease. IHDPS subsists well-being web-based, user-friendly, scalable, reliable and expandable.

The prediction of Heart disease, Blood Pressure and Sugar with the aid of neural networks was proposed by Niti Guru et al. [8]. Experiments were carried out on a sample database of patients' records. The Neural Network is tested and trained with 13 input variables such as Age, Blood Pressure, Angiography's report and the like. The supervised network has been recommended for diagnosis of heart diseases. Training was carried out with the aid of back propagation algorithm. Whenever unknown data was fed by the doctor, the system identified the unknown data from comparisons with the trained data and generated a list of probable diseases that the patient is vulnerable to.

III. PROPOSED SYSTEM

In this paper, the classification results on the gene set which selected by CFS-SS were compared with the results on the feature set selected by information gain(IG), principal component analysis(PCA) and CFS. native bayes(NB), was used to recognize the samples in the experimental.

A. Problem definition

In the proposed system we use a feature selection algorithm which is slightly different than existing CFS. Here the output of CFS is given as an input to proposed system and then by creating samples sets we get the output for proposed system. Stratified sampling is a method that separates the subsets with N features into L groups according to some strategies. There are N1, N2, ,…,Nn subsets in L groups respectively, and the total subsets of the L groups is 2N. The subsets with same features size will put into one group, and features with different size are put in different groups .Groups formed are homogenous. In our case N=3 as three attributes were selected by existing system .By creating 2^3 subsets (L) and finding which subsets gives the maximum gain ratio we get our final set, and the attributes of this sets will the minimum a features required to predict the output efficiently. These features are given to classifier Naïve Bayes.

B. Data set

Original data set of large records with 13 attributes used by Sellapanetal (2008) is used for consistency. [2] For simplicity, categorical attributes were used for all models. The 13 attributes along with their descriptions are as follows[9]:

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)
3. **Fasting Blood Sugar-** (value 1: > 120 mg/dl; value 0: < 120 mg/dl)
4. **RestEcg-resting electrographic results-** (value 0: normal; Value1: having ST-T wave abnormality; value 2: showing probable.)
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: down sloping)
7. **CA –** number of major vessels colored by fluoroscopy (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.**

C. Correlation feature stratified sampling (CFS-SS)

Stratified sampling is a method that separates the subsets with N features into L groups according to some strategies. There are N1, N2, ..., NL subsets in L groups respectively, and the total subsets of the L groups is 2N [9].In this paper, the subsets with same features size will put into a group. CFS-SS is a feature selection method based on CFS. In CFS the features are selected only by calculating the correlation between features –classes and features- features. When the selected features are applied for heart diagnosis it cannot achieve a satisfying performance. So a Stratified Sampling feature selection method based on CFS (CFS-SS) is used. as output of CFS will be given to CFS-SS and it will generate final output.

The Steps of CFS-SS Select Features Are Listed As Follows:

Step 1:Pre-process the inputting data: Firstly, pre-process the data include adding missing value, normalizing the data, ranking the gene by variance in descend. Then select the top S genes to next step.

Step 2:Pre-select features by CFS: From the top S genes, select the best gene subset (Scfs) containing features highly correlated with the class, yet uncorrelated with other genes in the subset by CFS. After this step, the number of genes in this subset will be reduced to less than 10% of the inputting data.

Step 3: Stratified Sampling from the Scfs. Stratified sampling the all subsets of the Scfs and put the subsets with the same gene size into a group. Select k groups (Sss).

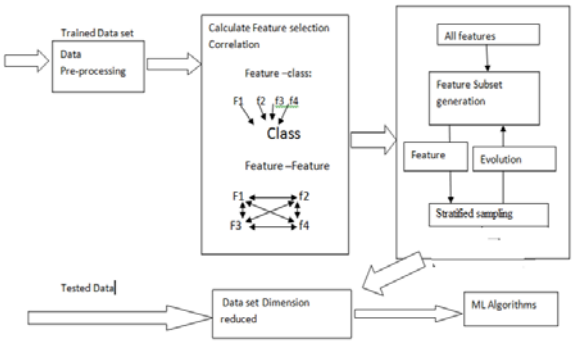
Step 4: Acquire the Best Feature Subset Scfs-ss: Select a classifier to test each element of the Sss using ten-fold cross-validation test. Return the element with the best performance.

The detail of the CFS-SS algorithm is as follows:

Algorithm CFS-SS (S, k, s, classifier)
Inputting S, k, s, classifier
Outputting subset Scfs−ss steps
(1) Data pre- pre-processing
(2) Stemp=filter(S , s)
(3) Scfs=CFS (Stemp)
(4) Sss=SS (Scfs , k)
(5) fori=0 to length(Sss)-1 do
(6) Acc[i]= evaluation (Sss[i],classifier).pctCorrect()
(7) Max pctCorrect=Acc[i]>Max pctCorrect? Acc[i]:Max pctCorrect

(8) endfor
(9) fori=0 to length(Sss)-1 do
(10) ifAcc[i] ==Max pctCorrect then
(11) Scfs−ss.add(Sss[i])
(12) endif
(13) endfor
(14) return Scfs−ss

The output will be reduced dataset .Attributes which is not found to be very efficient of determining the characteristics of heart diseases will be removed and only those parameters which are important will be given to classifiers. The advantage of using reduced dataset is that less parameter are considered for effective performance and hence the system is faster and more appropriate.



Component of CFS-SS

D. Classifiers

KNN, SVM, NB, DT were used to recognize the samples.

1) KNN: K-nearest neighbor (KNN) is one of the most common methods among memory based induction. Given an input vector, KNN extracts k closest vectors in the reference set based on similarity measures, and makes decision for the label of input vector using the labels of the k nearest neighbors. Pearsons coefficient correlation and Euclidean distance have been used as the similarity measure. When we have an input X and a reference set D = $d_1, d_2,\ldots, d_N$, the probability that X may belong to class c_j, $P(X, c_j)$ is defined as follows:

$$P\left(X, c_j\right) = \sum_{di}.Sim\left(X, d_i\right), P\left(d_i, c_j\right) - b_j \quad \text{--- (1)}$$

where $Sim(X, d_i)$ is the similarity between X and d_i and b_j is a bias term.

2) SVM: Support vector machine (SVM) estimates the function classifying the data into two classes. SVM builds up a hyperplane as the decision surface in such a way to maximize the margin of separation between positive and negative examples. SVM achieves this by the structural risk minimization principle that the error rate of a learning machine on the test data is bounded by the sum of the training-error rate and a term that depends on the Vapnik-Chervonenkis (VC) dimension. Given a labeled set of M training samples (Xi, Y i), where Xi ∈ RN and Yi is the associated label, Yi ∈ -1, 1, the discriminant hyperplane is defined by:

$$f(X) = \sum_{i=1}^{M} Y_i\, \alpha_i k(X_i X_i) + b \qquad \text{---- (2)}$$

where k(Xi.Xi) is a kernel function and the sign of f(X) determines the membership of X. Constructing an optimal hyper plane is equivalent to finding all the nonzero (support vectors) and a bias b.

3) DT: J48 is a kind of decision tree (DT), each attribute in the tree is completely independent. A DT model was developed using a variant of the classification and regression tree (CART) method, which consists of two steps tree construction and tree pruning. In the process of the tree construction, the

Algorithm identifies the best predictor variables that divide the sample in the parent node into two child nodes. The split maximizes the homogeneity of the sample population in each child node (e.g., one node is dominated by the cancer samples, and the other is populated with the noncancer samples). Then, the child nodes become parent nodes for further splits, and splitting continues until samples in each node are either in one classification category or cannot be split further to improve the quality of the DT model. To avoid over fitting the training data, the tree is then cut down to a desired size using tree cost-complexity pruning. In the end of the process, each terminal node contains a certain percentage of cancer samples. This percentage specifies the probability of a sample to be the cancer sample.

4) NB: NB is optimal when the features are conditionally independent. i.e. when the probability density function for class, denoted, can be decomposed as. In this case, the densities can be estimated separately for each feature which simplifies the training and makes NB feasible for very large feature sets. NB has been deemed surprisingly accurate. Even when the independence assumption is clearly false. NB may produce linear boundaries between the classes. This will happen if the individual densities are assumed to be Gaussian with the same variance (called Gaussian NB with shared variance). Only the means for the c classes need be estimated for each feature. Alternatively, variances for the classes can be estimated together with the means (Gaussian NB with distinct variance).

5) Naive Bayes (NB) with CFS-SS:

Naive Bayes algorithm has been proved to be better than other classifications in CFS as well as CFSS-SS .The naive Bayes algorithm employs a simplified version of Bayes formula to decide which class a novel instance belongs to. As it has been proved in existing system that Naive Bayes is a better classifier than C4.5 hence we are using it for our proposed system. After applying CFS-SS on training dataset the output achieved is given to NB, it is observed that it gives better efficiency as number of attributes have been reduced. The efficiency achieved is improved by 1.5% as compared to existing system. The efficiency comes out to be 86.45% whereas existing system was 85%.

E. 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.

CONCLUSION AND FUTURE WORK

The objective of our work is to predict more accurately the presence of heart diseases with reduced number of attributes. Originally thirteen attributes were involved in prediction of heart disease. In our work, stratified sampling is used to predict the attributes which contribute towards the prediction of diseases ailments which indirectly reduced the number of tests which are needed to be taken by the patient. Thirteen attributes are reduced to 2. Classifiers Naive Bayes is used to predict the diagnosis of the patient with the better accuracy obtained before reduction of attributes. Reduced dataset will perform better than the full dataset. Inconsistent and missing values were resolved before model construction but in real life that is not the case. Also the intensity of the diseases based on the result was unpredictable .we intend to extend our work applying fuzzy learning models to evaluate the intensity of heart diseases.

REFERENCES

[1]Tohn 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 514-518 (ICAESM -2012) March 30, 31, 2012.

[2] 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.

[3] 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.

[4] 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.

[5] 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

[6] 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.

[7] 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.

[8]Niti Guru, Anil Dahiya and NavinRajpal "Decision support system for heart diseases prediction using neural networks" Delhi Business Review Vol. 8, No. 1,PP 1-6 (January - June 2007)

[9] 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