ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 4, February 2018

COMPARITIVE STUDY ON HEART DISEASE DIAGONISIS USING DATAMINING TECHNIQUES

D.LOUISA MARY,
Assistant Professor,
Francis Xavier Engineering College, Tirunelveli.

S.MANIMALA,
PG Student
Francis Xavier Engineering College, Tirunelveli.

R.SUGANNA,
PG Student,
Francis Xavier Engineering College, Tirunelveli.

Abstract

Data mining techniques have been applied magnificently in many fields including business, science, the Web, cheminformatics, bioinformatics, and on different types of data such as textual, visual, spatial, real-time and sensor data. This paper presents a research on heart disease diagnosis and prediction. It present an overview of the current research being carried out using the data mining techniques to enhance heart disease diagnosis and prediction using Support Vector Machine(SVM), Naïve bayes classifier and sequential forward floating selection algorithm(SFFS).

I. Introduction

Heart disease refers to various types of conditions which can affect heart function.Few types are:Coronary artery (atherosclerotic) heart disease which affects the arteries to the heart .Valvular heart disease that affects how the valves function to regulate blood flow in and out of the heart.Cardiomyopathy that affects how the heart muscle squeezes.

Data mining in heart disease

Data mining tools and techniques in health care domain that can be used in prediction of heart disease system and their efficient diagnosis. A heart disease prediction model, which implements data mining technique, that can help the medical practitioners in detecting the heart disease status based on the patient's clinical data. Data mining classification techniques for good decision making in the field of health care addressed are namely Naive Bayes, sequential forward floating selection and Support Vector Machines. Hybridizing or combining any of these algorithms helps to make decisions quicker and more accurate. Data mining is a powerful new technology for the removal of hidden information and also extract predictive and actionable information from large databases that can be used to gain deep and novel insight. Using advanced data mining techniques to dig out valuable information, has been considered as an activist approach to improve the quality and accuracy of healthcare service while reduce the healthcare cost and analysis time. Using this technique presence of heart disease can be predicted exactly. Using more input attributes such as controllable and uncontrollable risk factors, more accurate results could be achieve. It can use many of input attributes. Other data mining techniques are also be used for predication such as Clustering, Time series, Association rules. The unstructured data available in healthcare industry database can also be mined using text mining.

Characteristics of Datamining:

- Ease Of Use
- Large Scale Deployment
- Power Consumption

Coronary Heart Disease

Coronary heart disease (CHD) is a disease that builds a waxy substance called plaque inside the coronary arteries. These arteries supply oxygen-rich blood to your heart muscle.When plaque appears in the arteries, then it is called atherosclerosis. The buildup of plaque occurs over many years.

II. Literature Survey

The authors, Ximeng Liu, Rongxing Lu, Jianfeng Ma in [1] proposed a privacy-preserving patient-centric clinical decision support system using naïve Bayesian classifier. By taking the advantage of rising cloud computing technique, processing unit can use big medical dataset stored in cloud platform to prepare naïve Bayesian classifier. And then apply the classifier for disease diagnosis without compromising the privacy of data provider. But in this system, the patient can securely retrieve the diagnosis results according to their own preference entered in the system. For the security mechanism authors provide all the data are processed in the encrypted form, that helps to achieve patientcentric diagnose result retrieval in privacy preserving way. But as the data is growing on increasing in much more faster way that reaches

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 4, February 2018

to cloud, one has to use more efficient data mining technique that helps in privacy preserving patient-centric clinical decision support systems.

The authors V. Krishnaiah, G. Narsimha, N. Subhash Chandra [4] in this paper, gives study of different data mining techniques that can be in use in robotic heart disease prediction systems. The analysis shows that different technologies are used in all the papers with taking different number of attributes reached their results different accuracy depends on tools used for execution. Even though applying data mining techniques to assist health care professionals in the diagnosis of heart disease is having various successes. The symbolic Fuzzy K-NN classifier can be tested with the unstructured data available in health care industry data base by modifying into fuzzified structured data with increased attributes and with a collection of more number of accounts to provide better accuracy to the system in predicting and diagnosing the patients of heart disease. Authors here provides a fast and simple thoughtful of diverse prediction models in data mining and helps to find greatest model for further work. But at the same time, this work can be enhanced by increasing the number of attributes for the existing system of our previous work.

The authors Luis Tabares, Jhonatan Hernandez, Ivan Cabezas [5] in this paper, said that Cloud computing has proved to be a possible solution to currently growing healthcare area. As all the needy once required healthcare services it should be cost-effective, everywhere and elastic model, enabling shared computing wealth between healthcare providers and patients and that is possible by cloud computing platform. They suggested that three main mechanism such as a knowledge base, an inference engine or an artificial intelligence component, are required to representation the knowledge by means of a service would be a suitable approach to implement alerts and reminders, knowledge service and diagnostic/treatment CDSS, and apply them properly. Findings in this paper said that several authors may be not appropriately using the terms cloud-based and cloud computing, since they are focusing on service-based or webbased architectures and also they are not detailing the conducted architectural design procedure. This will raises concerns and worries, since such lack of rigidity on the software engineering process does not allow identifying considered sources and used methods for gathering quality scenarios. As there was not any application of architectural evaluation methods based on scenarios. Consequently, well known concerns such as security and privacy may not be being well validated in practice.

The authors Shreya Anand, Ravindra B Patil, Krishnamoorthy P [6] in this paper [], provides some insights on the different risk models available for assessment of Cardiovascular disease (CVD) risk. Here, the strength and limitations of each of these models are found that there is no India specific CVD risk. To assist the primary care physicians for early diagnosis and management of chronic diseases such as CVDs, author develop WHO/ISH based risk stratification model based android based application. The application develop can be installed in the hospital team or on top of EMR solution to provide risk stratification as well as life style and medication recommendations to the subjects. This solution helps less well trained primary care physicians in the diagnosis and management of chronic diseases such as CVDs, assisted by clinical decision support systems (CDSSs). But as this solution is derived mostly for chronic diseases so it is necessary to develop the similar approach for other rising countries for better delivery of healthcare. [7] proposed a system, in which a predicate is defined for measuring the evidence for a boundary between two regions using Geodesic Graph-based representation of the image. The algorithm is applied to image segmentation using two different kinds of local neighborhoods in constructing the graph. Liver and hepatic tumor segmentation can be automatically processed by the Geodesic graph-cut based method. This system has concentrated on finding a fast and interactive segmentation method for liver and tumor segmentation. In the preprocessing stage, the CT image process is carried over with mean shift filter and statistical thresholding method for reducing processing area with improving detections rate. Second stage is liver segmentation; the liver region has been segmented using the algorithm of the proposed method. The next stage tumor segmentation also followed the same steps. Finally the liver and tumor regions are separately segmented from the computer tomography image.

III. PROPOSED SYSTEM

A.RECORD SET

The publicly available database on internet that is mostly preferred by many researchers is used for prediction. The input database contains CSV (Comma Separated Vector). The database contains 303 records.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 4, February 2018

B.CLUSTER DATA

In this module input attributes are normalized. For this normalization process we use the SFFS algorithm highly depends on the selection of initial cluster centers. SFFS does not guarantee to provide same result for different runs on same data set. Therefore improved algorithm is selected for clustering the dataset that does not require selecting initial clusters as input.

C.TRAINED DATASET

There are three input layers in ANN, they are: input layer, intermediate (called the hidden layer) and output. Several hidden layers can be placed between the input and output layers.

Input Layer – In this layer raw information is fed in to the network.
Hidden Layer - The activity of each hidden unit is determined by the activities of the input units and the weights on the connections between the input and the hidden units.
Output Layer - The behavior of the output units depends on the activity of the hidden units and the weights between the hidden and output units.

D.TESTING DATASET

In this module we are taking the entries from client as attributes specified. Comparing the attributes with trained dataset the system predicts output as per attributes. We can test multiple as well as single entry. Multiple entries used by Admin while single entry used by admin as well as client.

E. METHODS

1.SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) is a class of supervised learning algorithms and is a learning technique which trades off accuracy for generalization error. SVMs build a hyperplane which divides examples such that examples of one class are all on one side of the hyperplane, and examples of the other class are all on the other side. Let thedata are of the form (xi, yi) where the vectors xi are in a dot product space H, and yi are the class labels. Formally, any hyperplane in H is defined as

where w is a vector orthogonal to the hyperplane and represents the dot product. In an SVM, the idea is to find the hyperplane that maximizes the minimum distance from any training data point. The following constraint problem describes the optimal hyperplane:

N	FEATURE SUBSET	CLASSIFICATION ACCURACY
5	{3,7,8,12,13}	69.37%
6	{4,3,7,8,12,13}	72.55%
7	{1,4,3,7,8,12,13}	70.36%
13	{1,2,3...,12,13}	61.93%

Subject to $([xi, w] + b) \geq 1$

For i = 1, 2, …, m where m is the number of training examples. The above problem can be solved by introducing the Lagrange multipliers (ai ≥ 0, i = 1,....,m) and maximizing the following dual problem.

Thus the optimal margin hyperplane is represented as a linear combination of training points. Consequently, the decision function of classifying points only involves dot products between points. The algorithm that finds a separating hyperplane in the feature space can be stated entirely in terms of vectors in the input space and dot products in the feature space.

A. DATASET USED

In order to test our approach we use the Cleveland Heart Database taken from UCI learning data set

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

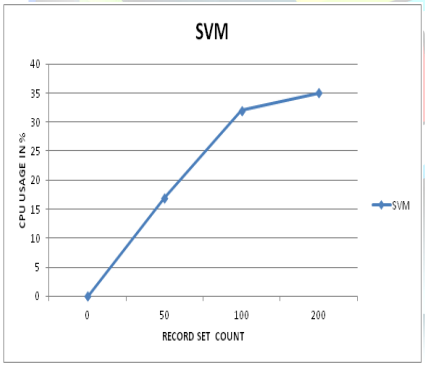
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 4, February 2018

repository which was donated by Detrano. The data set consists of 13 numeric attributes which include age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting ECG, maximum heart rate, exercise induced angina, oldpeak, slope, number of vessels coloured and thal respectively. The classes include integers valued from 0 (no presence) to 4 (types of heart diseases). Total number of patients instances is 303 and 250 of them are used for training and rest are used for testing the SVM.

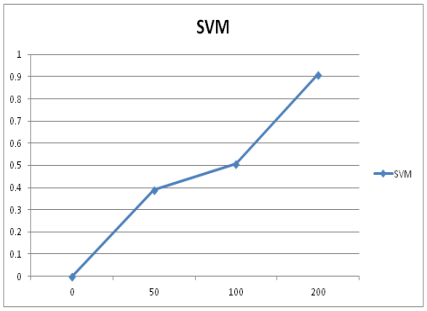
B. EXPERIMENTAL RESULTS

The algorithm was used to select top N features out of a total of 13 features of the Cleveland Heart database. The results for different values of N are summarized in Table below. It is to be noted that as the value of N is increased, the new optimal feature subset is obtained by addition of new features in the previous subset. This shows that addition of a feature affects the capability of that particular subset for classification either in a positive or negative manner and thus, there exists an optimal feature subset at which the accuracy is maximized.

CPU USAGE



CALULATION DELAY



2. SEQUENTIAL FORWARD FLOATING SELECTION ALGORITHM

The feature selection algorithm used in this project is developed based on sequential floating forward selection algorithm. A key codification introduced to the original SFFS algorithm is that, for the groups to be classified, the classification rate for each group is taken into account, instead of only considering the overall classification rate. This is important as the relative occurrence frequency of the groups could be substantially different.

In the modified SFFS algorithm, after each forward step, if the resulted performance is improved, the algorithm will perform a number of backward steps to exclude the least significant feature. On the contrary, if the performance is not improved after a forward step, no backward steps will be carried out. The modified SFFS procedure terminates when the optimization space reaches the required number of features.

Filters find the subsets by their information content.

Wrappers: use a classifier to evaluate subsets by their predictive accuracy (on test data) by statistical resampling or cross-validation.

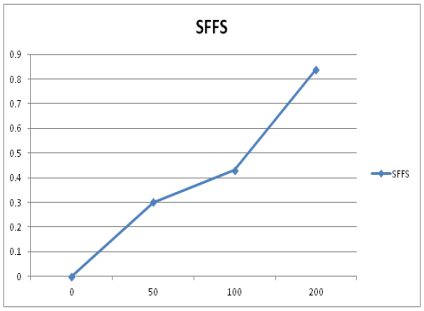
Accuracy: wrappers generally achieve better recognition rates than filters Because they identify the specific interactions between the classifier and the dataset

Ability to generalize: wrappersuse cross-validation measures of predictive accuracy

Algorithm:

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

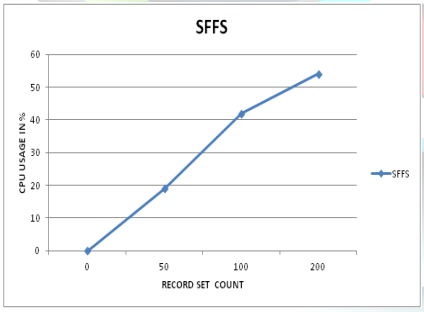
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 4, February 2018



B EXPERIMENTAL RESULTS

The algorithm usedthe 13 features of the Cleveland Heart database out of top n features. The results for different values of N are summarized in Table below. It is to be noted that as the value of N is increased, the new optimal feature subset is obtained by addition of new features in the previous subset. This shows that addition of a feature affects the capability of that particular subset for classification either in a positive or negative manner and thus, there exists an optimal feature subset at which the accuracy is maximized.

CPU USAGE:



CALULATION DELAY

3. NAÏVE BAYES CLASSIFIER.

Naive Bayes classifiers is a probabilistic classifiers based on applying Bayes theorem with strong (naive) independence assumptions between the features. A Naive Bayesian model is very useful in the field of medical science for diagnosing heart diseases. Naive Bayesian classifier outperforms more sophisticated classification methods.

Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c). Naive Bayes classifier considers the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.

N	FEATURE SUBSET	CLASSIFICATION ACCURACY
5	{3,7,8,12,13}	58.66%
6	{4,3,7,8,12,13}	52.68%
7	{1,4,3,7,8,12,13}	59.57%
13	{1,2,3...,12,13}	60.77%

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 4, February 2018

P(c|x) is the posterior probability of class (target) given predictor (attribute).
P(c) is the prior probability of class.
P(x|c) is the likelihood which is the probability of predictor given class.
P(x) is the prior probability of predictor.

DATASET

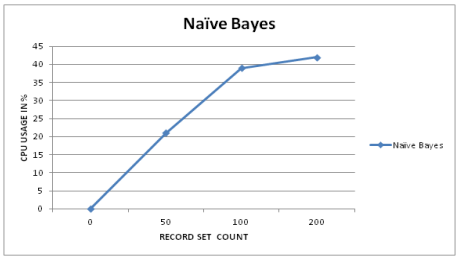
Theclinical data sets are collected from one of the leading diabetic research institute in Chennai contain records of about 500 patients. The diabetes data set is used to the diabetes people to know their risk factors, current management, treatment target achievements and arrangements and outcomes of regular surveillance for complications. This helps them to monitor their care and make the informed choices about their management.

EXPERIMENTAL RESULTS

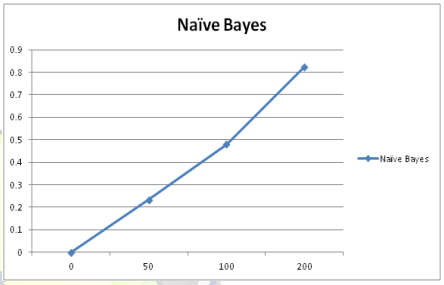
The algorithm selected 13 features out of top N features of the Cleveland Heart database. The results for different values of N are summarized in Table below. It is to be noted that as the value of N is increased, the new optimal feature subset is obtained by addition of new features in the previous subset. This shows that addition of a feature affects the capability of that particular subset for classification either in a positive or negative manner and thus, there exists an optimal feature subset at which the accuracy is maximized.

N	FEATURE SUBSET	CLASSIFICATION ACCURACY
5	{3,7,8,12,13}	63.87%
6	{4,3,7,8,12,13}	69.05%
7	{1,4,3,7,8,12,13}	65.87%
13	{1,2,3...,12,13}	57.06%

CPU USAGE

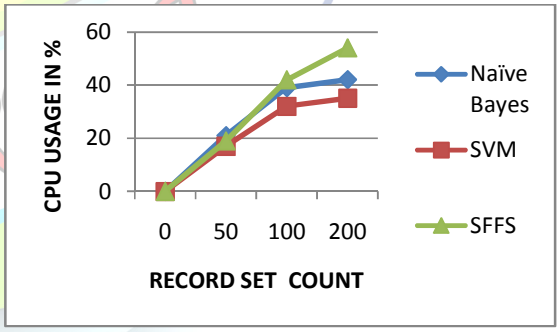


CALULATION DELAY



IV.COMPARISION

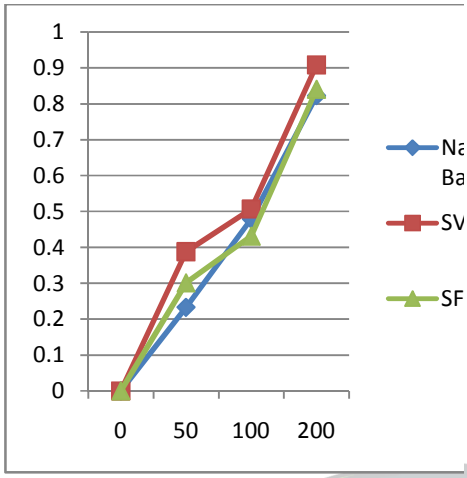
CPU USAGE



CALULATION DELAY

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 4, February 2018



V.CONCLUSION

In this survey, we have reviewed the data mining classification algorithms – Support Vector Machine, Naive Bayes Classifier, Sequential Floating Forward Selection algorithm against detection of heart diseases. The dataset used for this comparison survey are Cleveland, Hungary, Switzerland, and the VA Long Beach. The survey show the Naive Bayes will more efficient in mining result with less delay and SVM shows the better performance in the computation.

VI. 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 scientificResearch vol. 31 no. 4 pp. 642-656 2009 ISSN 1450-216-X.

2. UCI machine learning repository.

3.ClevelandDatabase [online] Available: http://archive.ics.uci.edu/ml/datasets/Heart+Disease.

4. C Kalaiselvi Dr G. M. Nasira "Classification and Prediction of heart disease from diabetes patients using hybrid particle swarm optimization and library support vector machine algorithm" International Journal of Computing Algorithm (IJCOA) vol. 4 pp. 1403-1407 March 2015 ISSN 2278-2397.

5.Aqueel Ahmed Abdul Hannan Shaikh "Data Mining Techniques to Find Out Heart Diseases: An Overview" International Journal of Innovative Technology and Exploring Engineering (IJITEE) vol. I no. 4 pp. 18-23 September 2012 ISSN 2278-3075.

6. SellappanPalaniappanRafiahAwang "Intelligent Heart Disease Prediction System Using Data Mining Techniques" International Journal of Computer Science and Network Security (IJCSNS) vol. 8 no. 8 pp. 343-350 August 2008.

7. Christo Ananth, D.L.Roshni Bai, K.Renuka, A.Vidhya, C.Savithra, "Liver and Hepatic Tumor Segmentation in 3D CT Images", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 3,Issue-2, February 2014,pp 496-503.

8. Niti Guru Anil DahiyaNavinRajpal "Decision Support System for Heart Disease Diagnosis using Neural Network" Delhi Business Review vol. 8 no. 1 January - June 2007.

9."Bakris GL. Preclinical diabetic cardiomyopathy: prevalence screening and outcome" Internal Medicine University Department pp. 1-27 2009.

10.Carlos Ordonez "Improving Heart Disease Prediction Using Constrained Association Rules" Seminar Presentation at University of Tokyo 2004.

11. M. K. Ali et al. "Diabetes and coronary heart disease: Current perspectives" Indian journal of medical research vol. 132 no. 5 pp. 584-597 Nov 2010.

12.LathaParthiban R. Subramanian "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm" International Journal of Biological and Medical Sciences vol. 3 no. 3 pp. 157-160 2008.

13. KiyongNohyHeonGyu Lee Ho-Sun Shon Bum Ju Lee KeunHoRyu "Associative Classification Approach for Diagnosing Cardiovascular Disease" in Springer vol. 345 pp. 721-727 2006.

14.K Collins MS RD "The cancer diabetes and heart disease Link" Todays Dietitian vol. 15 no. 3 pp. 46 Mar 2013.

15.C Kalaiselvi G M Nasira "Prediction of Heart Diseases and Cancer in Diabetic Patients Using Data Mining Techniques" Indian Journal of Science and Technology vol. 8 no. 14 July 2015.

16. M. Harris M "The role of primary health care in preventing the onset of chronic disease with a particular focus on the lifestyle risk factors of obesity tobacco and alcohol" Centre for Primary Health Care and Equity UNSW Canberra: National Preventive health taskforce pp. 1-21 2008.

17.K M. Anderson "Correlation of regional cardiovascular disease mortality in India with lifestyle and nutritional factors" International J Cardiol vol. 108 pp. 291-300 2006.

18.T SanthanamEphzibah "Heart disease prediction using hybrid genetic fuzzy model" Indian Journal of Science and Technology vol. 8 no. 9 May 2015.

19. C Kalaiselvi G M. Nasira "A New Approach for the diagnosis of diabetes and prediction of cancer using ANFIS" WCCCT-14. IEEE Proceedings International conference publications Feb 2014.

20.OlaniyiEbenezer ObaloluwaOyebadeKayodeOyedotunKhashman Adnan "Heart Diseases Diagnosis Using Neural Networks Arbitration" International Journal of Intelligent Systems and Applications vol. 7 no. 12 pp. 75-82 Nov 2015.

21. S. Sivagowry M. Durairaj A. Persia "An empirical study on applying data mining techniques for the analysis and prediction of heart disease" International Conference on Information Communication and Embedded Systems 2013.