International Journal of Pure and Applied Mathematics
Volume 119 No. 12 2018, 13357-13366
ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue
ijpam.eu

A COMPARATIVE STUDY ON HEART DISEASE ANALYSIS USING CLASSIFICATION TECHNIQUES

Hariharan K, Vigneshwar W.S, Sivaramakrishnan N* , Subramaniyaswamy V

School of Computing, SASTRA Deemed University, Thanjavur. India

*Corresponding Author

Abstract

AS it is modern era where people use computers more for work and other purposes physical activities are reduced. Due to work pressure they are not worrying about food habits. This results in introduction of junk food. These junk foods in turn results in many health issues. Major issue is heart disease. It is the major cause of casualty all over the world. Prediction of such heart disease is a tough task. But Countless mining approaches overcome this difficulty. Nowadays data mining techniques play's an important role in many fields such as business application, stock market analysis, e-commerce, medical field and many more. Previously many techniques like Bayesian classification, decision tree and many more are employed for heart disease prediction. In this proposal we are going to do a comparative study on three algorithms.

Keywords Used: heart disease, Support vector machine, decision tree, knn, Supervised Learning, data mining, classification, machine learning, data set

1.Introduction

1.1 Heart Disease:

Heart disease normally refer to abnormal functioning of heart. This usually occurs to aged people, but nowadays it has become common among people of all age groups. Especially the new born babies get affected by this disease [9-20]. This is called as congenital disease.

There are many types of heart disease:

1) Arrhythmia
2) Coronary artery disease
3) Dilated cardiomyopathy
4) Myocardial infarction
5) Cardiac Arrest

Data mining is a process of transforming raw or un useful data into a useful one to extract some information or pattern [1]. Data mining is a five step process they are

1) Identify source information
2) Take the points that need are to be analysed
3) Cull the information which are similar from the data
4) Analyse the important values from derived dataset
5) Clarify and address the result

There are five main popular techniques available for data mining they are

1) Classification
2) Association rule learning
3) Outlier prediction
4) Clustering
5) Regression

Out of these techniques we are going to employ classification in this work. Classification is a supervised learning where the result depends on class label and it will classify data in respective classes [21-33]. Generally classes will be in binary (such as yes/no, present/absent, up/down and more) or more.

There are many classification techniques available some of them are

1) Linear Classifiers like
   1.1) Logistic Regression,
   1.2) Naive Bayes Classifier. And
2) Support Vector Machines (SVM).
3) Decision Trees (D Tree).
4) Boosted Trees.
5) Random Forest.
6) Neural Networks.
7) K Nearest Neighbour (KNN).

We have to give the train set as input and train the model and can predict the class for given test data. We employed SVM, Decision Tree and KNN algorithm to find which yields more accuracy. Sensitivity and Specificity of SVM has already implemented [2].

1.2 Support Vector Machine (SVM):

Support vector machines (SVM) is a supervised (class based) machine learning algorithm generally used for classification. Here each and every attribute will be considered as a dimension. If there is n attributes it is considered to be n-dimension and it is plotted. Here n is the number of attributes we have. Then we enforce the classification by drawing the hyper plane. Hyper plane is the one that differentiate classes. Here the major task is to find the hyper plane. Once we found it then the task is simple. Advantage of this algorithm is it is applicable to bot linear and non-linear dataset.

1.3 Decision Tree (D TREE):

It is another excellent classification techniques used in data mining, statistics and machine learning. From the name it implies that it makes use of tree structure for data classification. Topmost node is considered to be the best predictor for the given data which is considered to be the root node. There are three techniques for attribute selection, which includes the root node. They are :

1) Information Gain
2) Gain Ratio
3) Gini Index

The leaf nodes of the decision tree represents the class label and the branches represents the affiliation of features that leads to the class labels. It is easy to understand and illustrate and is able to handle both analytical and absolute data. It requires less data preprocessing and it makes use of a white box model. If the dataset is large it performs well [6].

1.4 K Nearest Neighbor (KNN):

K Nearest Neighbours algorithm is nonparametric and lazy learning method which is used for classification. The nonparametric implies that it doesn't make any expectation on underlying data distributions. Lazy learning is an approach which doesn't make use of training data points for generalization. That is, explicitly training phase is not involved So all the training data should be loaded during the testing phase. It is the simplest of all machine learning algorithms [3].In training phase we will store the feature as vector and along with it we will store the class labels of training dataset. *k* is a user-defined variable or constant and the query is solved by assigning the label which has highest frequency among the *k* training samples nearby to that query point. Commonly employed distance metric is Euclidean distance for continuous attributes.

2. Related Work:

2.1 Clustering and Bayesian technique:

This paper works on both clustering and classification technique. It employs combination of two algorithms Naïve Bayes classification and K-Means clustering. So at first data's are grouped by clustering technique and then classified using classification technique. [1]

2.2 Risk minimization based SVM technique:

In this paper they have proposed risk minimization based Support vector machine for Heart disease prediction (SSH model). The accuracy, sensitivity and specificity of SRM SSH and the ordinary SVM algorithm are compared. [2]

2.3 Prognostic Data Mining for pharmaceutical Analysis: Critique on Heart Disease Prognosis:

Here they worked on Decision Tree and said it produce more accurate result than KNN and some Neural Networks based classification. Even some time Bayesian classification produces accurate results was the conclusion. After applying genetic algorithms Decision Tree produce more accurate results. [3]

2.4 Competent classification and study of Ischemic Heart Disease using Support Vector Machines based Decision Trees:

Ischemic heart disease is what this paper concentrates about. They have employed tree based proximal support vector machines.it yields more accuracy.it is a nonlinear classifier. Here data of 65 patients have been included which aids for the decision making. [4]

2.5 Perceptive Heart Disease Prediction Using Mining Techniques

Here they employed (IHPD) intelligent heart disease prediction methods such as Decision Tree, Naïve Bayes and Neural Networks. This method supports "What if " query which is not answered by traditional decision support system. [5]

3.Proposed Work:

Methodology emphasised here is a correlative study on accuracy, sensitivity and specificity of three different algorithms mentioned above. Preprocessing is not a part of this work. Generally we have datasets training and test dataset. Training set is used to train or develop a model which is used to predict the query post by the user. For test data class will be predicted for all the three (SVM, DTREE and KNN) algorithms and the accuracy, sensitivity and specificity are compared to find the best algorithm for this prediction problem with the dataset which is selected

We are working on R programming language which is very useful for data mining. Predefined functions are used instead of user defined function. some predefined packages used are caret, rpart, e1071. The library **"rpart"** is used for plotting the decision tree. Generally this package is used to plot any kind of graph. In the process of generating a model, it is suggested to perform all the iterations one by one for better understanding of the underlying concepts. The **"caret"** package in R is specifically developed to handle this issue and also contains various in-built generalized functions that are applicable to all modelling techniques. The package "**e1070"** has Functions for the areas of Statistics, Probability, fuzzy logic clustering, support vector machines, minimum distance computation, naive Bayes classifier and many more.

3.1 Dataset Description:

The Data set is taken from UCI repository (University of California, Irvine).

Different dataset such as Cleveland dataset, Hungarian dataset, Switzerland dataset and Statlog dataset are collected. The most commonly used datasets are Cleveland and statlog datasets because there is no missing values in these dataset.

So no preprocessing work is needed to fill the missing values. The dataset which we have employed here is Statlog dataset. This dataset has 12 attributes with 270 records out of which considerable amount of record is considered to be training dataset and some are test dataset.

1) Age
2) gender
3) type of chest pain
4) blood pressure value
5) serum cholesterol
6) resting electrocardiographic results
7) heart rate per minute
8) exercise induced angina
9) peak value
10) slope value
11) number of major vessels
12) thal

Age: age of the patient

Sex: gender of the patient (male or female) value-0,1

Chest Pain Type:

1) **1-Typical angina**
2) **2-Atypical angina**
3) **3-Non anginal pain**
4) **4-Asymptomatic**

Blood Pressure: Blood Pressure of patient

Serum Cholesterol: Cholesterol level in mg/dl

Electro Cardio Graph: Three values are specified (0,1,2)

Heart rate per minute: Heart rate of patient in one minute

angina: imbalanced flow of blood to heart muscle may result in severe pain in the chest. This is called as angina. Values=0,1

Number of major vessels: (0-3) colored by flourosopy

thal: three values are specified
3 = normal
6 = fixed
7 = reversable

Result: Present or Absent

4. Result:

The test data for which the algorithms have been implemented is displayed in Table-1(only a part of original test data has been displayed):

Table-1

Age	sex	Type	bp	serum	Ecg	heartrate	angina	oldpeak	slope	vessel	thal
65	1	4	120	177	0	140	0	0.4	1	0	7
56	1	3	130	256	2	142	1	0.6	2	1	6
59	1	4	110	239	2	142	1	1.2	2	1	7
60	1	4	140	293	2	170	0	1.2	2	2	7
63	0	4	150	407	2	154	0	4	2	3	7
59	1	4	135	234	0	161	0	0.5	2	0	7
53	1	4	142	226	2	111	1	0	1	0	7
44	1	3	140	235	2	180	0	0	1	0	3
61	1	1	134	234	0	145	0	2.6	2	2	3
57	0	4	128	303	2	159	0	0	1	1	3
71	0	4	112	149	0	125	0	1.6	2	0	3

The algorithms have been implemented and the accuracy of the results were studied. Along with accuracy, two other parameters sensitivity and specificity have been evaluated and they have been plotted as a bar graph. The accuracy, specificity and sensitivity are calculated using confusion matrix.

4.1 Accuracy

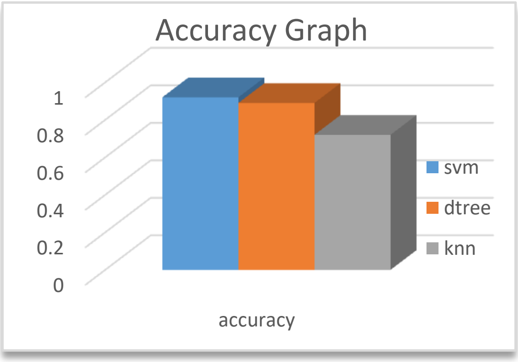
Accuracy produced by the Decision Tree algorithm is nearly 89%, Support Vector Machine algorithm is around 91% and K-Nearest Neighbour algorithm is 72%.

Table-2

Algorithm	Accuracy(%)
SVM	92
Decision Tree	89
K-Nearest Neighbour	72

The accuracy values have been represented in Table-2 and the graph has been plotted in figure 1

Figure 1



4.2 Sensitivity

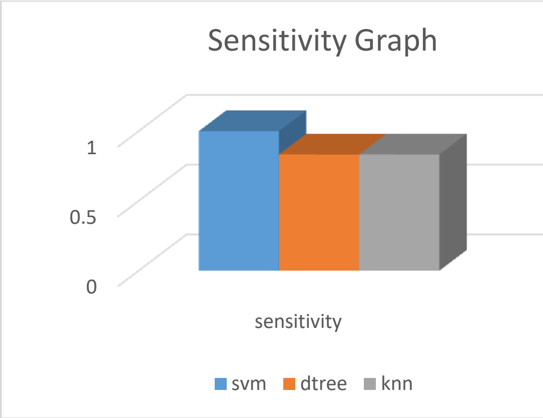
Sensitivity produced by the Decision Tree algorithm is nearly 83%, Support Vector Machine algorithm is around 100% and K-Nearest Neighbour algorithm is 83%.

The sensitivity values have been represented in Table-3 and the graph has been plotted in figure 2

Table 3

Algorithm	Sensitivity(%)
SVM	100
Decision Tree	83
K-Nearest Neighbour	83

Figure 2



4.2 Specificity

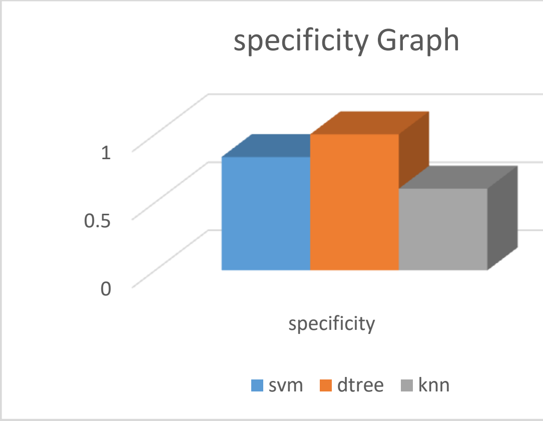
Specificity produced by the Decision Tree algorithm is nearly 100%, Support Vector Machine algorithm is around 83% and K-Nearest Neighbour algorithm is 60%.

The specificity values have been represented in Table-4 and the graph has been plotted in figure 3

Table 4

Algorithm	Specificity(%)
SVM	83
Decision Tree	100
K-Nearest Neighbour	60

Figure 3



5. Result:

With the derived results for this Heart Disease Prediction we conclude that SVM algorithm produces a better result compared to Decision Tree and KNN algorithms. Accuracy has been calculated using Confusion Matrix. Decision Tree and SVM have produced nearly equal results. But when it was tested against different types of data SVM produced a compromising results. So for this dataset SVM works well.

References:

[1] Rucha Shinde, Sandhya Arjun, Priyanka Patil and Prof. Jaishree Waghmare, “An Intelligent Heart Disease Prediction System Using K-Means Clustering and Naïve Bayes Algorithm”, International Journal of Computer Science and Information Technologies, Vol. 6 (1) , 2015, 637-639

[2] P.Selvakumar and Dr.S.P.Rajagopalan, “SSH - Structure risk minimization based Support vector machine for Heart disease prediction”, Proceedings of the 2nd International Conference on Communication and Electronics Systems (ICCES 2017) IEEE Xplore Compliant - Part Number:CFP17AWO-ART, ISBN:978-1-5090-5013-0, 2017

[3] Jyoti Soni, Ujma Ansari, Dipesh Sharma and Sunita Soni, “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

[4] Soman K.P. , Shyam Diwakar M. and Madhavdas P., “Efficient classification and analysis of Ischemic Heart Disease using Proximal Support Vector Machines based Decision Trees”, 0-7803-7651-x/03/$17. ©2003 IEEE, 2003

[5] Sellappan Palaniappan and Rafiah Awang, “Intelligent Heart Disease Prediction System Using Data Mining Techniques”, 978-1-4244-1968-5/08/$25.00 ©2008 IEEE, 2008

[6] Marjia Sultana, Afrin Haider and Mohammad ShorifUddin, “Analysis of Data Mining Techniques for Heart Disease Prediction”, 978-1-5090-2906-8/16/$31.00 ©2016 IEEE, 2016

[7] Milan Kumari and Sunila Godara, “Review of Data Mining Classification Model in Cardio Vascular Disease diagnosis*”,* IJCA, 2011.

[8] Srinivas, Kavitha Rani and Dr. Govarthan, “Application of Data Mining Techniques in Health Care and Prediction of Heart Attack”, IJCSE, Vol 2(2), pp 250-255, 2010.

[9] Logesh, R., Subramaniyaswamy, V., Vijayakumar, V., Gao, X. Z., & Indragandhi, V. (2017). A hybrid quantum-induced swarm intelligence clustering for the urban trip recommendation in smart city. Future Generation Computer Systems, 83, 653-673.

[10] Subramaniyaswamy, V., & Logesh, R. (2017). Adaptive KNN based Recommender System through Mining of User Preferences. Wireless Personal Communications, 97(2), 2229-2247.

[11] Logesh, R., & Subramaniyaswamy, V. (2017). A Reliable Point of Interest Recommendation based on Trust Relevancy between Users. Wireless Personal Communications, 97(2), 2751-2780.

[12] Logesh, R., & Subramaniyaswamy, V. (2017). Learning Recency and Inferring Associations in Location Based Social Network for Emotion Induced Point-of-Interest Recommendation. Journal of Information Science & Engineering, 33(6), 1629–1647.

[13] Subramaniyaswamy, V., Logesh, R., Abejith, M., Umasankar, S., & Umamakeswari, A. (2017). Sentiment Analysis of Tweets for Estimating Criticality and Security of Events. Journal of Organizational and End User Computing (JOEUC), 29(4), 51-71.

[14] Indragandhi, V., Logesh, R., Subramaniyaswamy, V., Vijayakumar, V., Siarry, P., & Uden, L. (2018). Multi-objective optimization and energy management in renewable based AC/DC microgrid. Computers & Electrical Engineering.

[15] Subramaniyaswamy, V., Manogaran, G., Logesh, R., Vijayakumar, V., Chilamkurti, N., Malathi, D., & Senthilselvan, N. (2018). An ontology-driven personalized food recommendation in IoT-based healthcare system. The Journal of Supercomputing, 1-33.

[16] Arunkumar, S., Subramaniyaswamy, V., & Logesh, R. (2018). Hybrid Transform based Adaptive Steganography Scheme using Support Vector Machine for Cloud Storage. Cluster Computing.

[17] Indragandhi, V., Subramaniyaswamy, V., & Logesh, R. (2017). Resources, configurations, and soft computing techniques for power management and control of PV/wind hybrid system. Renewable and Sustainable Energy Reviews, 69, 129-143.

[18] Ravi, L., & Vairavasundaram, S. (2016). A collaborative location based travel recommendation system through enhanced rating prediction for the group of

users. Computational intelligence and neuroscience, 2016, Article ID: 1291358.

[19] Vairavasundaram, S., Varadharajan, V., Vairavasundaram, I., & Ravi, L. (2015). Data mining-based tag recommendation system: an overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 5(3), 87-112.

[20] Logesh, R., Subramaniyaswamy, V., & Vijayakumar, V. (2018). A personalised travel recommender system utilising social network profile and accurate GPS data. Electronic Government, an International Journal, 14(1), 90-113.

[21] Vijayakumar, V., Subramaniyaswamy, V., Logesh, R., & Sivapathi, A. (2018). Effective Knowledge Based Recommeder System for Tailored Multiple Point of Interest Recommendation. International Journal of Web Portals.

[22] Subramaniyaswamy, V., Logesh, R., & Indragandhi, V. (2018). Intelligent sports commentary recommendation system for individual cricket players. International Journal of Advanced Intelligence Paradigms, 10(1-2), 103-117.

[23] Indragandhi, V., Subramaniyaswamy, V., & Logesh, R. (2017). Topological review and analysis of DC-DC boost converters. Journal of Engineering Science and Technology, 12 (6), 1541–1567.

[24] Saravanan, P., Arunkumar, S., Subramaniyaswamy, V., & Logesh, R. (2017). Enhanced web caching using bloom filter for local area networks. International Journal of Mechanical Engineering and Technology, 8(8), 211-217.

[25] Arunkumar, S., Subramaniyaswamy, V., Devika, R., & Logesh, R. (2017). Generating visually meaningful encrypted image using image splitting technique. International Journal of Mechanical Engineering and Technology, 8(8), 361–368.

[26] Subramaniyaswamy, V., Logesh, R., Chandrashekhar, M., Challa, A., & Vijayakumar, V. (2017). A personalised movie recommendation system based on collaborative filtering. International Journal of High Performance Computing and Networking, 10(1-2), 54-63.

[27] Senthilselvan, N., Udaya Sree, N., Medini, T., Subhakari Mounika, G., Subramaniyaswamy, V., Sivaramakrishnan, N., & Logesh, R. (2017). Keyword-aware recommender system based on user demographic attributes. International Journal of Mechanical Engineering and Technology, 8(8), 1466-1476.

[28] Subramaniyaswamy, V., Logesh, R., Vijayakumar, V., & Indragandhi, V. (2015). Automated Message Filtering System in Online Social Network. Procedia Computer Science, 50, 466-475.

[29] Subramaniyaswamy, V., Vijayakumar, V., Logesh, R., & Indragandhi, V. (2015). Unstructured data analysis on big data using map reduce. Procedia Computer Science, 50, 456-465.

[30] Subramaniyaswamy, V., Vijayakumar, V., Logesh, R., & Indragandhi, V. (2015). Intelligent travel recommendation system by mining attributes from community contributed photos. Procedia Computer Science, 50, 447-455.

[31] Vairavasundaram, S., & Logesh, R. (2017). Applying Semantic Relations for

Automatic Topic Ontology Construction. Developments and Trends in Intelligent Technologies and Smart Systems, 48.

[32] Logesh, R., Subramaniyaswamy, V., Malathi, D., Senthilselvan, N., Sasikumar, A., & Saravanan, P. (2017). Dynamic particle swarm optimization for personalized recommender system based on electroencephalography feedback. Biomedical Research, 28(13), 5646-5650.

[33] Arunkumar, S., Subramaniyaswamy, V., Karthikeyan, B., Saravanan, P., & Logesh, R. (2018). Meta-data based secret image sharing application for different sized biomedical images. Biomedical Research,29.