Recommendation of Attributes for Heart Disease Prediction using Correlation Measure

S.Chellammal, R. Sharmila

ABSTRACT--- Heart diseases are the major cause for human mortality rate. Correct diagnosis and treatment at an early stage will save people from heart disease and will decrease mortality rate due to heart problem. Since ten years various data mining techniques have been used to facilitate the prediction of heart diseases .In general prediction algorithms for trained with huge, known dataset to arrive at a classifier which then predicts the diseases for unknown data with the help of classifying attributes. These attributes also called as features. In this work relevant features are determined for heart disease prediction with known dataset using correlation measures. The results are presented.

Index Terms — Correlation method, relevant features, Prediction

I. INTRODUCTION

In health care industry, predicting heart disease is a challenging issue [1-3]. In early days medical tests such as Electrocardiogram (ECG) and blood tests have been used for predicting heart diseases. In addition to clinical tests, computer aided diagnosis systems, namely, patient information, medical diagnosis and medical images are being used for predicting heart diseases. Machine learning algorithms have significant role in predicting diseases [5]. Nowadays along with machine learning techniques, big data tools and technologies are being employed to handle unstructured data, huge size and speedy data [6]. In[7], we have proposed a conceptual framework for the prediction of diseases using parallel programming models in big data environment. It is pre-requisite to identify features that are relevant to the prediction of diseases. From, literature it is found that 13 attributes, namely, *age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca and thal* are being used for predicting heart diseases. In this work, these features are analysed for their relevancy for prediction of heart disease using correlation technique. Prediction results obtained using three different classifiers namely Naive Bayes (NB) classifier, Multi Layer Perceptron (MLP) and Sequential Minimal Optimization (SMO) with the above attributes are presented. The results show that 10 attributes, namely, *age, sex, cp, restecg, thalach, exang, oldpeak, slope, ca and thal* are found as most relevant attributes in predicting heart diseases. Experiments have been conducted using the thus said classifiers with a typical data set obtained from UCI repository. Accuracy obtained using different classifiers with different sets of attributes are reported in this paper. The paper is organized as follows. Section II gives research works that have their focus on feature selection for prediction of heart disease. Sections III describes the method, tool and dataset used for determining relevant features. Section IV presents results and discussion. Section V concludes the work.

II. LITERATURE SURVEY

In general, filter and wrapper methods [8] are being used for feature selection for predicting heart diseases. In filter methods where feature selection is independent of the prediction algorithm [10], different statistical factors such as Information Gain, Chi-square test, Fisher Score, Correlation, LDA (Linear Discriminant Analysis) and ANOVA (Analysis of Variance) are used for finding relevancy [9]. As wrapper methods are computationally very expensive, filter methods are frequently used in practice [11-12]. Hence we made an investigation on research works which employ filter methods.

From literature, several research works[13-20] have used the thirteen attributes, namely, *age, sex, chest pain type(cp), resting blood pressure(trestbps), serum cholesterol(chol), fasting blood pressure(fbs), resting electrocardiographic results(restecg), maximum heart rate achieved(thalach), exercise induced angina(exang), ST depression induced by exercise relative to rest(oldpeak), the slope of the peak exercise ST segment(slope), number of major vessels colored by flourosopy(ca) and thalassemia(thal)* for prediction of heart diseases. Also, various research works have found relevant features from these 13 attributes using filter techniques. For example in [21], the authors have used three attributes, namely, *cp, ca and thal* for predicting heart diseases. In [22], six attributes namely *cp, thalach, exang, oldpeak, ca and thal* have been used for prediction. In [23], an alternate set of six attributes such as *cp, restecg, ex, thalach, slope and thal* have been used for prediction. In [24], seven attributes namely *cp, restcg, exang, thalach, oldpeak, ca and that have* been used for prediction. Also, in [25], an alternate set of seven attributes, namely, *cp, restecg, exang, thalach, oldpeak, ca and thal* have been used for prediction. In [26], eight attributes namely thalach, *chol, **ca**, exang, slope, oldpeak, restecg and sex* have been used for prediction, In [4] nine attributes namely *thal, ca, exang,*

Revised Manuscript Received on July 10, 2019

S.Chellammal, Bharathidasan University Constituent Arts and Science College, Navalurkuttapattu, Tiruchirappalli, Tamil Nadu, India. (E-mail: chelsganesh@gmail.com)

R. Sharmila, Bharathidasan University Constituent Arts and Science College, Navalurkuttapattu, Tiruchirappalli, Tamil Nadu, India. (E-mail: sharmiparam@gmail.com)

thalach, oldpeak, cp, slope, sex and age have been used for prediction while in [27], eleven attributes, namely, *age, cp, thal, trestbps , chol, ca, slope, restecg, fbs, oldpeak and ca* used for prediction.

Table I highlights the relevant attributes used by different research works. Further for clarity, technique used for feature selection, data set, tool, classifier and its accuracy of prediction are provided in Table 1.

Table 1 Relevant attributes used by different research works

Ref. No	Attribute selection method	Selected attributes	No. of attributes	Dataset & tool	Classification accuracy In %	
					With selected attributes	With 13 attributes
22	*CFS Subset Evaluator* And three search methods namely *Best First Search, Rank search and Genetic Search*	*chest pain type, max heart rate, exercise induced angina, oldpeak, number of major vessels colored and thal*	6	Weka 3.6 and Cleveland database from UCI repository.	*J48* - 83.8284	*J48* - 84.1584
					NB - 84.1584	*NB* - 85.4785
					Logistic Regression - 76.8977	*Logistic Regression* - 77.2277
					Classification via regression - 84.1584	*Classification via regression* -83.1683
					SMO - 84.4884	*SMO* - 82.8383
3	*CFS Subset Evaluator*	*gender, chest pain type, cholesterol, thal, exang and oldpeak.*	6	Weka and Hungarian Heart Disease dataset from UCI repository	*SVM*-89.4	*SVM*-97.9
26	*Binary Artificial Bee Colony*	*chest pain type, resting blood pressure, chol, max heart rate achieved, slope and thal*	6	Cleveland data set from UCI repository.	*BABC-KNN*-92.4	(*Please note*: Results are discussed only for 6 attributes)
21	*CFS and Bayes theorem*	*chest pain type, number of major vessels colored and thal*	3	Weka and Statlog dataset	*NB* - 85.18	*NB* - 83.70
27	*CFS Subset Evaluator with memory based classifier*	*chest pain type, resting electrocardiographic results, exercised induced angina, maximum heart rate achieved, oldpeak, number of major vessels colored and thal*	7	Open source machine learning tool and Statlog Heart Disease dataset from UCI repository	*IBK* - 77.78	*IBK* - 74.75
					K star - 79.18	*K star* - 75.90
					LWL - 69.037	*LW*L - 71.10
4	*Fisher-filtering selection*	*thal, number of major vessels colored, exercise induced angina, max heart rate achieved, oldpeak, chest pain type, slope, sex and age*	9	TANAGRA machine learning tool and UCI Statlog dataset	*BLR*-83.33	*BLR*-82.59
					C4.5-77.41	*C4.5*-74.11
					C-RT-75.56	*C-RT* - 72.96
					SVML- 84.07	*SVML* - 82.59
					SVMP - 55.56	*SVMP* - 55.56
					SVMR- 82.59	*SVMR* - 80.74
					SVMS - 84.07	*SVMS* - 81.85
					ID3 - 70.73	*ID3* - 70.73
					KNN - 70.00	*KNN* - 66.30

					MLP - 82.22	*MLP* - 80.74
					MLR - 83.33	*MLR* - 82.59
					NB - 84.81	*NB* - 82.59
	ReliefF selection	*sex, thal, resting electro cardio graphic results, number of major vessels colored, chest pain type and exercise induced angina.*	6	TANAGRA machine learning tool and UCI Statlog dataset	*BLR* - 83.70	*BLR*-82.59
					C4.5 - 82.96	*C4.5*-74.11
					C-RT - 79.63	*C-RT* - 72.96
					SVML- 84.81	*SVML* - 82.59
					SVMP - 55.56	*SVMP* - 55.56
					SVMR - 80.37	*SVMR* - 80.74
					SVMS - 84.44	*SVMS* - 81.85
					ID3 - 70.37	*ID3* - 70.73
					KNN - 80.00	*KNN* - 66.30
					MLP - 83.33	*MLP* - 80.74
					MLR - 83.70	*MLR* - 82.59
					NB - 83.70	*NB* - 82.59
23	*PSO(Particle Swarm Optimization)*	*Chest pain type, electrocardiographic results, max heart rate achieved, exercise induced angina, oldpeak, number of major vessels colored and thal.*	7	Weka and four different real datasets	*KNN+PSO*-81.4	*KNN*+PSO-78.14
24	*Multi Layer Perceptron*	*maximum heart rate achieved, chol, number of major vessels colored, exercise induced angina, slope, oldpeak, electrocardiographic results and sex*	8	Dataset from medical dataset	*MLP* – 90+	*MLP* – 93+
25	*LSTSVM(Least Square Twin Support Vector Machine)*	*age, chest pain type, thal, blood pressure, chol, number of major vessels colored, slope, electrocardiographic results, fasting blood sugar, oldpeak and maximum heart rate.*	11	MatlabR2012a and Statlog dataset from UCI repository.	*LSTSVM*– 85.18	(*Please note*: Results are discussed only for 11 attributes)

III. PROPOSED APPROACH

In this work it is proposed to find and recommend a list of relevant attributes for different classifiers which yield high accuracy. Relevant features are determined using the steps given in Fig. 1

(i) Step – 1 Rank the attributes according to correlation measure.
(ii) Step - 2 Perform classification of known data using three commonly used classifiers, namely, MLP, SMO and NB and compare the accuracy of different classifier models.
(iii) Step – 3 Recommend relevant features for the chosen classifiers based on accuracy.

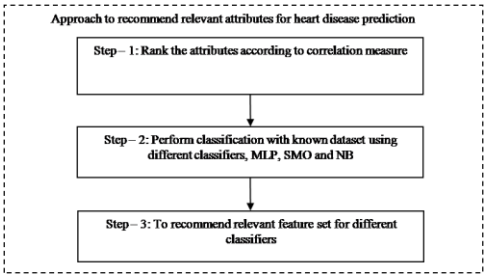

Fig. 1 Proposed Approach

To perform the above steps, three experiments have been conducted. It is proposed to use Cleveland dataset and Weka 3.6.9 tool in Windows 7 operating system. Data are collected from Cleveland database of UCI repository [28]. UCI includes four different databases such as Cleveland (303), Hungarian (294), Switzerland (123), and Long Beach VA (200) for heart disease prediction. This database contains 76 attributes. There class labels are integer, valued from 0 (no presence) to 4(presence). Among these four databases, Cleveland dataset has less number of missing values (only six records contains missing values) than the other datasets. So Cleveland database has been taken up for experiment work. Further the details of attributes of the dataset are given in Table 2.

Table 2 Details of Attributes

S.No	Attribute	Value	Description
1	age	29 – 62	age in years
2	sex	0 – male, 1- female	gender
3	cp	1-typical angina; 2-atypical angina 3-non-anginal pain; 4-asymptomatic	chest pain type
4	trestbps	Numeric value(140mm/Hg)	resting blood pressure in mm/Hg
5	chol	Numeric value(289mg/dl)	serum cholesterol in mg/dl
6	fbs	1-true, 0-false	fasting blood pressure>120mg/dl
7	restecg	0-normal, 1-having ST-T, 2-hypertrophy	resting electrocardiographic results
8	thalach	140,173	maximum heart rate achieved
9	exang	1-yes, 0-no	exercise induced angina
10	oldpeak	Numeric value	ST depression induced by exercise relative to rest
11	slope	1-upsloping, 2-flat, 3-downsloping	the slope of the peak exercise ST segment
12	ca	0-3 vessels	number of major vessels colored by flourosopy
13	thal	3-normal, 6-fixed defect, 7-reversable defect	thalassemia
14	num	0: < 50% diameter narrowing 1: > 50% diameter narrowing	diagnosis of heart disease (angiographic disease status)

EXPERIMENTATION AND RESULTS

There may be many attributes related to a given prediction problem. But not all the attributes have strong association with the prediction. Hence finding the relevant attributes for a given prediction problem is important. In this work, relevant attributes for heart disease prediction are determined using correlation measure. As mentioned above, from literature it is found that the *13 attributes (thal, ca, exang, oldpeak, thalach, cp, slope, sex, age, restecg, trestbps, chol, fbs)* are being used while predicting heart diseases. In order to find the weight or rank of these attributes an experiment has been conducted. In this experiment the correlation between each attribute and class label is found out. Attributes along with their correlation values are given in Table 3.

In order to determine which feature set produces optimal accuracy, second experiment is conducted with three popularly used classifiers, namely, NB, MLP and SMO. While doing the above experiment, attributes are added one by one up to 13 attributes by choosing the attribute with highest weight as the first attribute. Accuracy of these classifiers is computed for different feature sets as given in Table 4.

Table 3 Attributes and their weights

S.No	Attribute	Rank
1	thal	0.4862
2	ca	0.4608
3	exang	0.4368
4	oldpeak	0.4307
5	thalach	0.4217
6	cp	0.3817
7	slope	0.3564
8	sex	0.2809
9	age	0.2254
10	restecg	0.1664
11	trestbps	0.1449
12	chol	0.0852
13	fbs	0.028

Table 4 Accuracy for three different classifiers with different attribute sets

Attribute list	Classifiers Accuracy in %		
	NB	MLP	SMO
thal	76.5677	76.2376	76.5677
thal, ca	79.2079	79.538	75.5776
thal, ca, exang	82.5083	82.5083	78.5479
thal, ca, exang, oldpeak	79.868	80.8581	80.8581
thal, ca, exang, oldpeak, thalach,	82.5083	79.868	84.1584
thal, ca, exang, oldpeak, thalach, cp	84.1584	79.2079	83.8284

thal, ca, exang oldpeak, thalach, cp, slope	84.1584	81.1881	83.8284
thal, ca, exang oldpeak, thalach, cp, slope, sex	84.1584	81.5182	83.4983
thal, ca ,exang oldpeak, thalach, cp, slope, sex, age	83.4983	83.4983	83.4983
thal, ca, exang oldpeak, thalach, cp, slope, sex, age, restecg	84.4884	81.8482	84.8185
thal, ca, exang oldpeak, thalach, cp, slope, sex, age, restecg, trestbps	83.8284	82.8383	84.8185
thal, ca, exang oldpeak, thalach, cp, slope, sex, age, restecg, trestbps, chol	83.8284	80.5281	84.4884
thal, ca, exang oldpeak, thalach, cp, slope, sex, age, restecg, trestbps, chol, fbs	83.4983	80.8581	84.1584

From Table 4, it found that attribute set 9 and attribute set 10 are giving good accuracy. The accuracy values for these attributes sets with NB, MLP and SMO are given in Table 5

Table 5 Accuracy values for different classifiers with attribute set-9 and attribute set-10

Classifier Name	Selected attributes	Number of selected attributes	Accuracy %	
			With selected attributes	With 13 attributes
NB	*thal, cp, ca, oldpeak, exang, thalach, slope, age, sex*	9	83.4983	83.4983
MLP	*thal, cp, ca, oldpeak, exang, thalach, slope, age, sex*	9	83.4983	80.8581
SMO	*thal cp, ca, oldpeak, exang, thalach, slope, age, sex, restecg*	10	84.8185	84.1584

INTER-COMPARISON WITH EXISTING METHODS

Results obtained in this work are inter-compared with feature set represented in literature. From literature, the commonly used feature set are found to be

- Feature set with 6 attributes (*thal, cp, ca, oldpeak, exang, thalach)*
- Feature set with 7 attributes (*thal, cp, ca, oldpeak, exang, thalach, slope)*
- Feature set with 9 attributes (*thal, cp, ca, oldpeak, exang, thalach, slope, age, sex*)
- Feature set with 10 attributes (*thal, cp, ca, oldpeak, exang, thalach, slope, age, sex, restecg*)
- Feature set with 13 attributes (*thal, cp, ca, oldpeak, exang, thalach, slope, age, sex, restecg, trestbps, chol,*
- *fbs*)

In order compare the feature set obtained in the proposed work with that of the above mentioned feature set, third experiment is conducted with above feature set and same classifiers. The accuracy values obtained from the above feature sets for different classifiers are given in Table 6

Table 6 Inter-comparison of the proposed approach with existing methods

Attribute Name	No. of attributes	Classifiers Accuracy in %		
		NB	MLP	SMO
thal, cp, ca, oldpeak, exang, thalach	6	84.1584	79.2079	83.8284
thal, cp, ca, oldpeak, exang, thalach, slope	7	84.1584	81.1881	83.8284
thal, cp, ca, oldpeak, exang, thalach, slope, age, sex	9	83.4983	83.4983	83.4983
thal, cp, ca, oldpeak, exang ,thalach, slope, age, sex, restecg	**10**	**84.4884**	**81.8482**	**84.8185**
thal, cp, ca ,oldpeak exang, ,thalach ,slope, age, sex, restecg, trestbps, chol, fbs	13	83.4983	80.8581	84.1584

CONCLUSION AND FUTURE WORK

In this work, an approach is proposed attributes used for prediction of heart diseases are analyzed using correlation measure. Thirteen attributes for the prediction

of heart disease are identified from literature. These attributes are ranked according to correlation measure.

Then with different classifiers, accuracy values are obtained for all possible feature sets. It is found that feature set consisting of 10 attributes, *thal, cp, ca, oldpeak, exang, thalach, slope, age, sex, restecg* are recommended as relevant feature set(please refer Table 6-feature consisting of 10 attributes are given in bold) for further research. In our further work, it is proposed to use the recommended feature set to study the impact of big data techniques and technologies in enhancing the accuracy of classifiers [29-30].

REFERENCES

1. V. Manikantan & S.Latha,"Predicting the Analysis of Heart Disease Symptoms Using Medicinal Data Mining Methods", International Journal on Advanced Computer Theory and Engineering, Volume-2, Issue-2, pp.5-10, 2013.
2. Dr.A.V.Senthil Kumar, "Heart Disease Prediction Using Data Mining preprocessing and Hierarchical Clustering", International Journal of Advanced Trends in Computer Science and Engineering, Volume-4, No.6, pp.07-18, 2015.
3. Uma.K, M.Hanumathappa, "Heart Disease Prediction Using Classification Techniques with Feature Selection Method", Adarsh Journal of Information Technology, Volume-5, Issue-2, pp.22-29, 2016.
4. Hidayet TAKCI, "Improvement of heart attak prediction by the feature selection methods", Turkish Journal of Electrical & Computer Science, pp.1-10, 2018.
5. Himanshu Sharma, M.A.Rizvi, "Prediction of Heart Disease using Machine Learning Algorithms:A Survey",International Journal on Recent and Innovation Trends in Computing and Communication,Volume-5,Issue-8,pp.99-104, 2017.
6. Ms.S.Suguna, Sakthi Sakunthala.N ,S.Sanjana, S.S.Sanjhana, "A Survey on Prediction of Heart Disease using Big data Algorithms", International Journal of Advanced Research in Computer Engineering & Technology,Volume-6,Issue-3,pp.371-378,2017.
7. R.Sharmila, S.Chellammal, "A Conceptual method to enhance the prediction of heart diseases using big data Techniques", International Journal of Computer Sciences and Engineering, Volume-6, special Issue-4, pp.21-25, 2018.
8. Uma.K, Dr.M.Hanumathappa, "Feature Selection Methods for Heart Disease Prediction with Data mining Techniques", Seventh International Conference on Advanced in Computer Engineering-ACE, 2016.
9. Pinar Yildirim," Filter Based Feature Selection Methods for Prediction of Risks in Hepatitis Disease", International Journal of Machine earning and Computing, Volume. 5, No.4, pp.258-263, 2015.
10. https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example
11. Mark A.Hall, Lloyd A.Smith, "Feature Selection for Machine Learning:Comparing a Correlation-based Filter Approach to the Wrapper". Proceedings of the Twelfth International FLAIRS Conference.
12. Mital Doshi,Dr.Setu K Chaturvedi,"Correlation Based Feature Selection(CFS) Techniques to Predict Student Performance", International Journal of Computer Networks & Communication,Volume-6,No.3,pp.197-206,2014.
13. Chaitrali S.Dangare, Sulabha S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", International Journal of Computer Applications,Volume-47,No.10,pp.44-448,2012.
14. Shamsher Bahadur Patel, Pramod Kumar Yadav,Dr.D.P.Shukla, "Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques",Journal of Agricultural and Veterinary Science,Volume-4,Issue-2,pp.61-64, 2013.
15. Pediredla Praveen Kumar, Sunita A. Yadwad V V D L Tejaswi, "A Data Mining Technique for Prediction of Heart Disease using Hadoop Mapreduce",International Journal of Computer Application,Volume-6,No.6,pp1-8,2016.
16. Aditya Methaila, Prince Kansal, Himanshu Arya, Pankaj Kumar, "Early Heart Disease Prediction using Data Mining Techniques",Computer Science & Information Technology,pp.53-59,2014.
17. Dr.Durairaj.M, Sivagowry.S, "A Pragmatic Approach of PreProcessing the Dataset for Heart Disease Prediction", International Journal of Innovative Research in Computer and Communication Engineering,Volume-2,Issue-11,pp.6457-6465,2014.
18. V.Subha,M.Revathi,D.Murugan, "Comparative Analysis of Support Vector Machine Ensembles for Heart Disease Prediction", International Journal of Computer Science & Communication Networks,Volume-5(6),pp.386-390.
19. T.Revathi, S.Jeevitha,"Comparative Study on Heart Disease Prediction System using Data Mining Techniques",International Journal of Science and Research,Volume-4,Issue-7,pp.2120-2123,2015.
20. Megha Shahi, Er.Rupinder Kaur Gurm, "Heart Disease Prediction System using Data Mining Techniques- A Review",International Journal of Technology and Computing",Volume-3,Issue-4,pp.73-77,2017.
21. T.John Peter, K.Somusundaram, "Study and Development of Novel Feature Selection Framework for Heart Disease Prediction", International journal of Scientific and Research Publications, Volume-2,Issue-10,pp.1-7, 2012.
22. B.Kavitha, R.Naveen Kumar, "Improving Heart Attack Prediction System using Feature Selection and Data Mining Methods", International Journal of Advanced Research in Computer Science,Volume-1,No.v,pp.455-461,2010.
23. Divya Tomar and Sonali Agarwal, "Feature Selection based Least Square Twin Support Vector Machine for Diagnosis of Heart Disease", International Journal of Bio-Science and Bio-Technolgy, Volume-6, No.2, pp.69-82, 2014.
24. Jabbar MA,"Prediction of heart disease using k-nearest neighbour and particle swarm optimization", Biomedical Research,Volume-28,Issue-9,pp.4154-4158.
25. B. Subanya, R R Rajalaxmi, "A Novel Feature Selection Algorithm for Heart Disease Classification", International Journal of Computational Intelligence and Informatics, Volume-4, No.2, pp.117-124, 2014.
26. Kittipol Wisaeng, "Predict the Diagnosis of Heart Disease using Feature Selection and k-Nearest Neighbor Algorithm", Applied Mathematical Science,Volume-8,No.83,pp.4103-4113,2014.
27. Lakshmi Devasena,"Performance Evaluation of Memory based Classifiers with Correlation Based Feature Selection Subset Evaluator for smart Heart Disease Prediction", International Journal of Research in Engineering and Technology, Volume-5, Special issue-7, pp.11-17, 2016.
28. http.//archive.ics.uci.edu/ml/datasets/Heart+Disease.
29. Prediction of Diseases using Big Data Analysis
30. Heart Disease Prediction with MapReduce by using Weighted Association Classifier and K-Means.