ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 2, February 2020

The Diagnosis of Coronary Heart Disease Using C4.5 Algorithm

Dr. U.Nilabar Nisha[1] M.E., Ph.D., M.B.A., M.Phil. Gowthamraja S[2], Kalaikannan P[2], Dhivakar M[2]

Assistant Professor, Department of Computer Science and Engineering, Mahendra Institute of Technology, Namakkal, India

Department of Computer Science and Engineering, Mahendra Institute of Technology, Namakkal, India

ABSTRACT: In Machine Learning, Decision tree is the mostly used classifier for predictive Modeling. The C4.5 classifier suffers from over fitting; poor attribute split technique, inability to handle continuous valued and missing valued attributes with high learning cost. Among all, over fitting and split attribute has high impact on the accuracies of prediction. The Efficient Back-track pruning algorithm is introduced here to overcome the drawback of over fitting. The proposed concept is implemented and evaluated with the UCI Machine Learning Hungarian database. This database having 294 records with fourteen attributes were used for forecasting the heart disease and relevant accuracies were measured. This implementation shows that the proposed Back-track pruned algorithm is efficient when compared with existing C4.5 algorithm, which is more suitable for the application of large amounts of healthcare data. Its accuracy has been greatly improved in line with the practical Health care Historical data. The result obtained proves that the performance of Back-track pruned C4.5 algorithm is better than C4.5 algorithm.

KEYWORDS: Data mining, Heart disease, Decision Tree, C4.5 algorithm, Over fitting, Back-track pruning.

I. INTRODUCTION

Data mining is process of extracting hidden knowledge from large volumes of raw data.Datamining is used to discover knowledge out of data and presenting it in a form that is easily understand to humans Disease Prediction plays an important role in data mining [1]. Data Mining is used intensively in the field of medicine to predict diseases such as heart disease, lung cancer, breast cancer etc. [2].

This paper analyzes the heart disease predictions using different classification algorithms. Medicinal data mining has high potential for exploring the unknown patterns in the data sets of medical domain .These patterns can be used for medical analysis in raw medical data [3]. Heart disease was the major cause of casualties in the world. Half of the deaths occur in the countries like India, United States are due to cardiovascular diseases.Medical data mining techniques like Association Rule Mining, Clustering, and Classification Algorithms such as Decision tree, C4.5 Algorithm are implemented to analyze the different kinds of heart based problems [4]. C4.5 Algorithm and Clustering Algorithm like K-Means are the data mining techniques used inmedical field [5].

II. HEART DISEASE IS A TERM COVERING ANY DISORDER OF THE HEART

Unlike cardiovascular disease, which describes problems with the blood vessels and circulatory system as well as the heart, heart disease refers only issues and deformities in the heart itself.

According to the Centers for Disease Control (CDC), heart disease is the leading cause of death in the United Kingdom, United States, Canada, and Australia. One in every four deaths in the U.S. occurs as a result of heart disease.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 2, February 2020

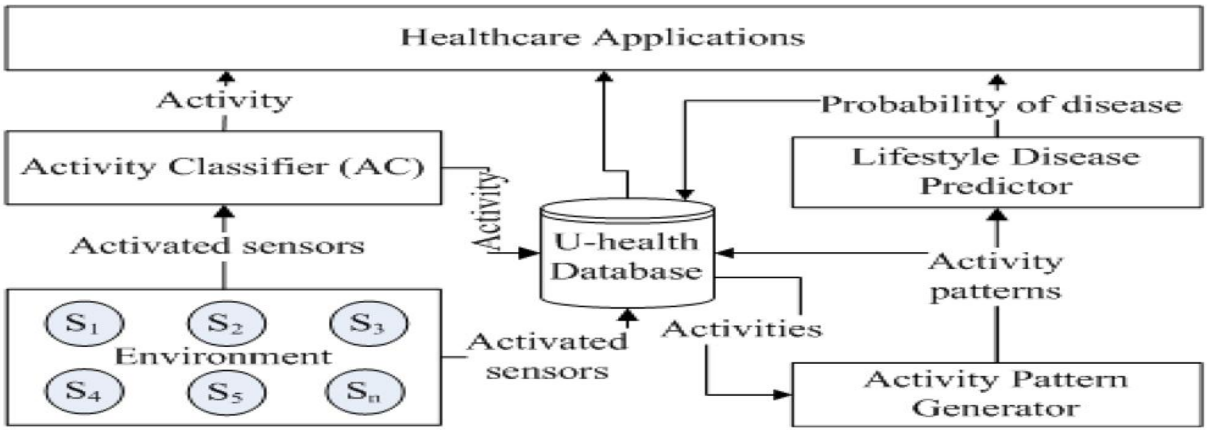

Figure: 1 healthcare applications

Fast facts on heart disease

One in every four deaths in the U.S. is related to heart disease. Coronary heart disease, arrhythmia, and myocardial infarction are some examples of heart disease. Heart disease might be treated with medication or surgery. Quitting smoking and exercising regularly can help prevent heart disease.

The symptoms of heart disease depend on which condition is affecting an individual.

- However, common symptoms include chest pain, breathlessness, and heart palpitations. The chest pain common too many types of heart disease is known as angina, or angina pectoris, and occurs when a part of the heart does not receive enough oxygen.
- Angina can be triggered by stressful events or physical exertion and normally lasts under 10 minutes.
- Heart attacks can also occur as a result of different types of heart disease. The signs of a heart attack are similar to angina except that they can occur during rest and tend to be more severe.
- The symptoms of a heart attack can sometimes resemble indigestion. Heartburn and a stomach ache can occur, as well as a heavy feeling in the chest.

Other symptoms of a heart attack include:

- pain that travels through the body, for example from the chest to the arms, neck, back, abdomen, or jaw
- lightheadedness and dizzy sensations
- profuse sweating
- nausea and vomiting
- Heart failure is also an outcome of heart disease, and breathlessness can occur when the heart becomes too weak to circulate blood.
- Some heart conditions occur with no symptoms at all, especially in older adults and individuals with diabetes.

I. C4.5 ALGORITHM

Decision trees are powerful and popular tools for classification and prediction. Decision trees produce rules, which can be inferred by humans and used in knowledge system such as database. C4.5 is an algorithm for building decision trees. It was designed by Quinlan .It converts the trained trees into sets of if-then rules. It handles discrete and continuous attributes. C4.5 is one of widely-used learning algorithms.C4.5 algorithm builds decision trees from a set of training data using the concept of information entropy. C4.5 is also known as a statistical classifier.

- Check for base cases.
- For each element x, discover the normalized information gain from dividing on x.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 2, February 2020

- Let x_best be the element with the highest normalized information gain.
- Create a decision node that breaks on a best.
- Repeats on the sublists obtained by dividing on x_best, and add those nodes as children of node.

Table 1. Sample data set.

ID	REFERENCE ID	ATTRIBUTE
1	#1	Age
2	#2	Sex
3	#9	painloc: chest pain location
4	#16	Relrest
5	#18	cp: chest pain type
6	#21	trestbps: resting blood pressure
7	#24	chol: serum cholesterol in mg/dl
8	#27	Smoke
9	#28	cigs (cigarettes per day)
10	#31	years (number of years as a smoker)
11	#33	fbs: (fasting blood sugar > 120 mg/dl)
12	#36	dm (1 = history of diabetes; 0 = no such history)
13	#38	famhist: family history of coronary artery disease
14	#42	thalach: maximum heart rate achieved
15	#44	exang: exercise induced angina
16	#45	Sedentary Lifestyle/inactivity
17	#47	ca: number of major vessels (0-3) colored by fluoroscopy
18	#49	Hereditary
19	#51	num: diagnosis of heart disease

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 2, February 2020

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
63	male	typ_angina	145	233	t	left_vent_hyper	150	no	2.3	down	0	fixed_defe	'<50'
67	male	asympt	160	286	f	left_vent_hyper	108	yes	1.5	flat	30	normal	'>50_1'
67	male	asympt	120	229	f	left_vent_hyper	129	yes	2.6	flat	2	reversable	'>50_1'
37	male	non_anginal	130	250	f	normal	187	no	3.5	down	0	normal	'<50'
41	female	atyp_angina	130	204	f	left_vent_hyper	172	no	1.4	up	0	normal	'<50'
56	male	atyp_angina	120	236	f	normal	178	no	0.8	up	0	normal	'<50'
62	female	asympt	140	268	f	left_vent_hyper	160	no	3.6	down	2	normal	'>50_1'
57	female	asympt	120	354	f	normal	163	yes	0.6	up	0	normal	'<50'
63	male	asympt	130	254	f	left_vent_hyper	147	no	1.4	flat	1	reversable	'>50_1'
53	male	asympt	140	203	t	left_vent_hyper	155	yes	3.1	down	0	reversable	'>50_1'

ROC Area	0.64	0.064	0.5	0.5

III. RESULTS AND DISCUSSION

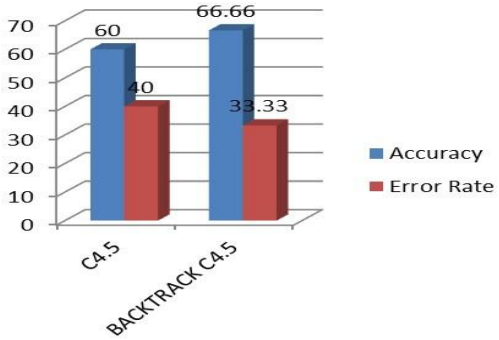

Figure 1. Accuracies and error-rate of back-track pruned C4.5 and C4.5.

The Figure 1 compares the accuracies of C4.5 and Back-Track Pruned C4.5 algorithms. It is proved that the Accuracy of Back-Track Pruned C4.5 is better than C4.5 Classification algorithm. Also the incorrectly

- ***C4.5 classification algorithm***

classified instances were minimum in Back-Track Pruned C4.5 than C4.5. The other

Data set name	Approach	Accuracy
Heart disease	C4.5+CDC	67.7
	C4.5	100

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 2, February 2020

Results of C4.5 and back-pruned C4.5 algorithm

Table 2. Result obtained from back-pruned C4.5 and C4.5 algorithm.

Measures	Existing C4.5 algorithm		Proposed Back-Track Pruned C4.5 algorithm	
Correctly Classified	60%		66.66%	
Incorrectly Classified	40%		33.33 %	
Kappa statistic	0.2		0	
Mean absolute error	0.4233		0.4444	
Root mean squared error	0.6144		0.4714	
Relative absolute error	77.6111%		85.7143 %	
Root relative squared error	112.6336%		90.4534 %	
Coverage of cases (0.95 level)	70%		100 %	
Mean rel. region size (0.95 level)	65%		100 %	
Total No. of Instances	10		10	
Class	Val ues		Values	
	<50	>50_1	<50	>50_1
TP Rate	0.4	0.8	1	0
FP Rate	0.2	0.6	1	0
Precision	0.67	0.571	0.67	0
Recall	0.4	0.8	1	0
F-Measure	0.5	0.667	0.8	0

The Training and Test datasets belongs to Hungarian database with 294 records and fourteen attributes were applied on C4.5 and Back-Track Pruned C4.5 Classification algorithm. It is proved that that Back-Track Pruned C4.5 Classification algorithm is performing better accuracy than C4.5. The BackTrack Pruned C4.5 yields the accuracy of 66.66% while the C4.5 gives 60%. The Incorrectly classified Instances of BackTrack Pruned C4.5 is 33.33% and C4.5 is 40%. The other measures of Back-pruned C4.5 and C4.5 Algorithm is

IV. CONCLUSION

The heart disease dataset from UCI Machine learning Hungarian database with 294 records and 14 attributes has been implemented with C4.5 and Back-track Pruned C4.5 Classification algorithm. The Missing attributes are not included for gain and entropy computations, so that reliability is maintained. Back-pruned C4.5 performs better than C4.5 with 66.66% of improved accuracy. Here, an optimal solution has been provided for over fitting of the decision Tree. It supports both continuous and discrete instances. This approach minimizes the overhead, memory required, size of the tree, time taken and results in improved accuracy by removing the branches not useful with minimized learning cost. As the paper focus is completely on prediction of the heart disease further, the same can be extended to predict the survival rate of the heart attack patients.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 2, February 2020

REFERENCES

1. Chakrabarti S. Data mining: know it all. Burlington, MA: Elsevier/Morgan Kaufmann Publishers, 2009.
2. Xiaoliang Z, Jian W, Hongcan Y, Shangzhuo W. Research and application of the improved algorithm C4. 5 on decision tree. InTest and Measurement, 2009. ICTM'09. International Conference on, 184-187.
3. Jothikumar R, Sivabalan RV. Performance Analysis on Accuracies of Heart Disease Prediction System Using Weka by Classification Techniques. AJBAS 2015; 9: 741-749.
4. Jothikumar R, Sivabalan RV, Sivarajan E. Accuracies of j48 weka classifier with different supervised weka filters for predicting heart diseases. ARPN J Eng Applied Sci2015 ; 10: 7788-7793.
5. Jothikumar R, Sivabalan RV, Kumarasen AS. Data Cleaning Using Weka For Effective Data Mining In Health Care Industries. Int J ApplEng Res 2015.
6. Xiaohu W, Lele W, Nianfeng L. An application of decision tree based on id3. Physics Procedia 2012; 25: 1017-1021.
7. Jabbar MA, Deekshatulu BL, Chndra P. Alternating decision trees for early diagnosis of heart disease. In Circuits, Communication, Control and Computing (I4C), 2014 International Conference on, 322-328.