VOL. 12, NO. 23, DECEMBER 2017 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

DEVELOPMENT OF AN EXPERT SYSTEM ALGORITHM FOR DIAGNOSING CARDIOVASCULAR DISEASE USING ROUGH SET THEORY IMPLEMENTED IN MATLAB

Aaron Don M. Africa
Department of Electronics and Communications Engineering, Gokongwei College of Engineering, De La Salle University, Manila, Philippines
E-Mail: aaron.africa@dlsu.edu.ph

ABSTRACT

Cardiovascular disease refers to conditions that involve narrow or blocked blood vessels. This disease when remained untreated may lead to a heart attack. When a person has a cardiovascular disease, the heart may not be able to pump enough blood to the body. When there is insufficient blood the brain or other organs may become damaged. Cardiovascular disease is challenging to diagnose because its symptoms may be mistaken for other diseases. Early detection, if a person has cardiovascular disease is a big advantage in combating the ailment. This is because diagnosing the disease early may reduce the complications it may bring. This research will develop an Expert System Algorithm for the diagnosis of cardiovascular disease. This research will guide the person diagnosing the disease to provide the appropriate recommendation. The Rough Set Theory will be used to reduce the rules so it can be easily diagnosed. This research will utilize the Statlog Heart Data Set of the UCI machine learning repository. Matrix Laboratory or MATLAB will be used to implement the system.

Keywords: cardiovascular disease, rough set theory, expert system, matrix laboratory, public health engineering.

1. INTRODUCTION

1.1 Background of the Study

Cardiovascular disease is a dangerous disease. This disease is one of the leading causes of deaths in the world. Cardiovascular disease is a general term that refers to conditions affecting the heart and blood vessels. This disease is associated with the buildup of atherosclerosis or fatty deposits inside the arteries. This disease can cause damage to other organs like heart, brain and kidneys especially if a heart attack occurs. A heart attack happens when blood flow to a part of a heart is blocked by a blood clot. If this clot clots completely the blood flow that part of the heart muscle supplied by the artery begins to die [1]. Many people survive their first heart attack but the problem is once you already had a heart attack you are not the same as before. Some of your organs may already have received damage and you are prone to another heart attack. The next heart attack a person may have can be fatal and may lead to further complications. The best way is to prevent a heart attack before it occurs in the first place [2]. One way to do it is to detect if a person has a cardiovascular disease. This disease is difficult to diagnose because it may be mistaken for other diseases [3].This research will develop an Expert System algorithm that can be used to diagnose cardiovascular disease using Rough Set Theory. This algorithm will be implemented in MATLAB.

1.2 Statement of the Problem

Diagnosing if a person has cardiovascular disease is a difficult task. One way to make diagnosis is to use an Expert System. This system simulates the judgment of a human expert to give the correct diagnosis. The main problem in using an Expert System is the information that you should input to it should be complete. In order to make a diagnosis all of the condition attribute information should be known. The problem is there are instances when the value of the condition attribute is unknown. The data cannot be obtained therefore a diagnosis cannot be provided.

1.3 Significance of the Study

Developing an Expert System algorithm for the diagnosis of cardiovascular disease will help cardiologist in making diagnosis. Knowing if a person has this disease will greatly help because early treatment can be performed before it is too late. This research will also help in solving the problem with Expert Systems about dealing with incomplete information. Integrating the Expert System algorithm with Rough Set Theory will enable the system to handle incomplete information.

1.4 Objectives

1.4.1 Main objective

- To develop an Expert System algorithm for the diagnosis cardiovascular disease.

1.4.2 Specific objective

- To implement the Expert System algorithm in MATLAB
- To use Empirical testing to validate the results
- To utilize the Statlog Heart Data Set of the UCI machine learning repository as the training data

VOL. 12, NO. 23, DECEMBER 2017 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

1.5 Assumptions, scope and limitations

- This study is about the diagnosis of cardiovascular disease only using the Statlog Heart UCI dataset

- The scope of this study is limited only to the attributes of the Statlog Heart UCI dataset

2. REVIEW OF THE RELATED LITERATURE

2.1 Expert system

An Expert System is defined as a computer program that represents knowledge and can reason with it. This program also simulates a human expert in giving a recommendation on the possible cause of a problem [4]. In order to solve expert level problems, an Expert System should have efficient access to a knowledge base. This system should also have a reasoning mechanism in order to apply the knowledge to the problems inputted into it. Generally, these systems are built on the ideas of knowledge representation. These systems also have production rules in order to give the correct possible cause. An Expert System usually has a shell which is an existing knowledge independent framework on which additional information can be inputted. [5]. Expert Systems have no limits on the problems it can solve. Expert systems are distinguished from typical computer systems because they simulate human reasoning, perform reasoning over the database oh human knowledge and solve problems by using approximate methods. The knowledge acquisition component allows the Expert System to be inputted with data from human experts. These data can be refined if required. Expert Systems currently are used in fields like artificial intelligence, digital processing and biomedical engineering [6].

2.2 Cardiovascular disease

Cardiovascular disease is generally known as damage to the heart and blood vessels [7]. The term cardiovascular disease includes stroke, heart disease and other diseases of the heart. Major blood vessels consist of arteries. These arteries carry blood away from the heart and veins which returns it. Capillaries are tiny vessels where the exchange of carbon dioxide and oxygen occurs. There are instances when fatty materials like cholesterol deposit on the walls of vessels. These deposits are known as plaque. These plaques clogs the artery, thus reduce the space for blood flow. This phenomenon is called arteriosclerosis [8]. If for example the plaque ruptures in the artery walls, platelets try to repair the damage. The problem here is more clots might form and overtime the walls of the blood vessels may lose their elasticity. The loss of elasticity will contribute to hypertension or high blood pressure. The blockage of an artery leads to a part of the body being starved by oxygen. This in turn will result to a heart attack or stroke [9]. It is better to detect if a cardiovascular disease is present early on so heart attack can be prevented.

2.3 Statlog heart data set of the UCI machine learning repository

The UCI machine learning repository are datasets compiled by University of California Irvine. They have different fields like annealing datasets and automobile datasets. Their datasets have various forms like categorical, numerical and mixed. Research subjects can be in the field of statistics, mathematics and artificial intelligence. Their repository is open source and free to download for scientific uses. The Statlog Heart dataset is a multivariate dataset which is real and categorical in form. This dataset slightly varies from the Hungarian UCI database which is also included in the UCI machine learning repository. It consist 13 attributes where the 13th attribute which is the thal , is the decision attribute [10].

3. THEORETICAL FRAMEWORK

In the early 1980's Rough Set Theory Zdzislaw I. Pawlak developed the Rough Set Theory [11]. This theory deals with the classification analysis of data tables. The primary goal of the Rough Set analysis is to synthesize concepts of approximation from the acquired data [12]. The general modelling process typically consist of sequence of various steps that all requires several degrees of tuning and fine adjustments. In order to make this function, an interactive management process is required. Rough Sets consist of information systems. An information system is where the data is represented in a table. Every column represents an attribute and the rows represent the number of cases. The intersection of the rows and columns are called objects. In an information system *IS = (U,A)* where *U* is a non-empty finite set of objects called the universe and *A* is a non-empty finite set of attributes. In an information system $a: U \to V_a$ for every $a \in A$. The set V_a is referred to as the value set of *a*[13].

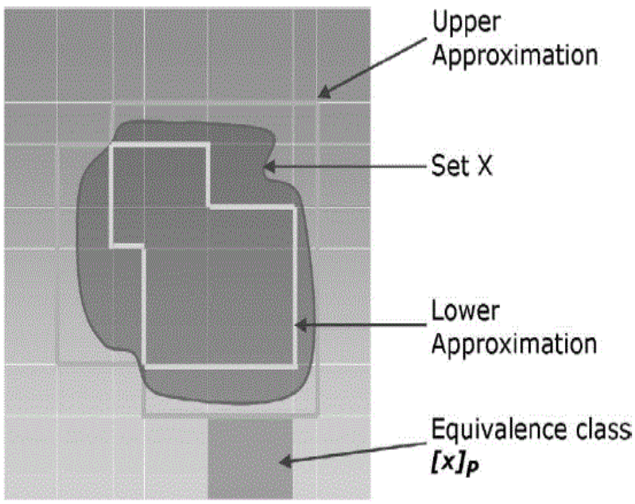

Figure-1. Rough set theory lower and upper approximations [14].

ARPN Journal of Engineering and Applied Sciences

Rough Set Theory also has a concept of lower and upper approximation. Figure-1 shows its representation. The lower approximation is also known as the positive region it is the union of all equivalence classes in *[x]p* and contains the target set. The upper approximation is the union of all the equivalence classes of *[x]p* and should have a non-empty intersection with the target set [14].

4. METHODOLOGY

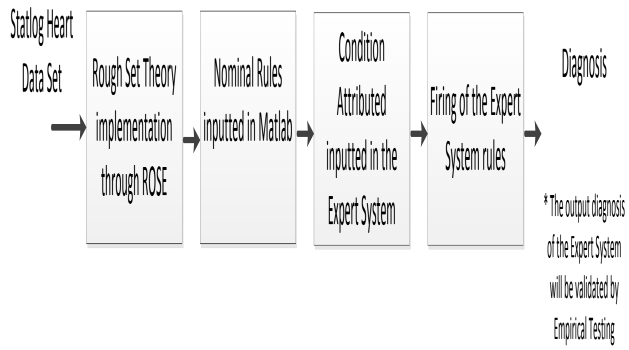

Figure-2. Block diagram of the methodology.

The Statlog Heart Data Set of the UCI machine learning repository will be used as the data. This database consists of 13 attributes which are the following:

A) age
B) sex
C) chest pain type (4 values)
D) resting blood pressure
E) serum cholesterol in mg/dl
F) fasting blood sugar > 120 mg/dl
G) resting electrocardiographic results (values 0,1,2)
H) maximum heart rate achieved
I) exercise induced angina
J) oldpeak = ST depression induced by exercise relative to rest
K) the slope of the peak exercise ST segment
L) number of major vessels (0-3) colored by fluoroscopy
M) thal: 3 = normal; 6 = fixed defect; 7 = reversible defect

Attributes 1 to 12 are the condition attributes while attribute 13 is the decision attribute. In order to determine the value of attribute 13, attributes 1 to 12 should be known. However there are instances where not all the values of the 12 attributes are known so the decision attribute cannot be determined. In order to solve this problem Rough Set theory is used so only the essential attributes are needed to determine the value of the decision attribute. The rules produced will where only the essential attributes are needed are called nominal rules.

Figure-2 shows the block diagram of the Methodology. The values of the Statlog Heart Data Set of the UCI machine learning repository will be inputted in the Rough Set Data Explorer (ROSE). The ROSE will implement the Rough Set Theory algorithm then output the rules which nominal [15]. These rules will then be inputted in the MATLAB Expert System. Condition attributes will be inputted in the Expert System. After the attributes are inputted the correct Expert System rules will fire. The correct diagnosis will then be outputted by the System. Empirical testing will be used to validate the diagnosis of the Expert System [16].

Empirical testing is performed by:

a) Count the total number of rules of the Expert System. The result will be variable *a.*

b) The rules produced by the Expert System and the information system will be compared if they match.

c) The total number of matched rules will be defined as variable *b.*

d) The percent validity c will then be computed. It is obtained by using the formula $c = (b/a) \times 100$.

5. DATA AND RESULTS

5.1 Information system of the UCI Statlog heart data set

The UCI Statlog Heart Data Set is then converted to an Information System a shown in Figures 1 and 2. For coding purposes a "0" is added at the end of the value of the condition attributes. The condition attributes are A to L while the decision attribute is the thal M. Knowing the correct value of the decision attribute will help the physician in diagnosing cardiovascular disease.

ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

Table-1. Information system of the UCI Statlog heart data set from A to G.

#	A	B	C	D	E	F	G
1	670	00	30	1150	5640	00	20
2	570	10	20	1240	2610	00	00
3	640	10	40	1280	2630	00	00
4	740	00	20	1200	2690	00	20
5	650	10	40	1200	1770	00	00
6	560	10	30	1300	2560	10	20
7	590	10	40	1100	2390	00	20
8	600	10	40	1400	2930	00	20
9	630	00	40	1500	4070	00	20
10	590	10	40	1350	2340	00	00
11	530	10	40	1420	2260	00	20
12	440	10	30	1400	2350	00	20
13	610	10	10	1340	2340	00	00
14	570	00	40	1280	3030	00	20
15	710	00	40	1120	1490	00	00
16	460	10	40	1400	3110	00	00
17	530	10	40	1400	2030	10	20
18	640	10	10	1100	2110	00	20
19	400	10	10	1400	1990	00	00
20	670	10	40	1200	2290	00	20
21	480	10	20	1300	2450	00	20
22	430	10	40	1150	3030	00	00
23	470	10	40	1120	2040	00	00
24	540	00	20	1320	2880	10	20
25	480	00	30	1300	2750	00	00
26	460	00	40	1380	2430	00	20
27	510	00	30	1200	2950	00	20
28	580	10	30	1120	2300	00	20
29	710	00	30	1100	2650	10	20
30	570	10	30	1280	2290	00	20

Table-2. Information system of the UCI Statlog heart data set from H to M.

#	H	I	J	K	L	M
1	1600	00	16	20	00	7
2	1410	00	03	10	00	7
3	1050	10	02	20	10	7
4	1210	10	02	10	10	3
5	1400	00	04	10	00	7
6	1420	10	06	20	10	6
7	1420	10	12	20	10	7
8	1700	00	12	20	20	7
9	1540	00	40	20	30	7
10	1610	00	05	20	00	7
11	1110	10	00	10	00	7
12	1800	00	00	10	00	3
13	1450	00	26	20	20	3
14	1590	00	00	10	10	3
15	1250	00	16	20	00	3
16	1200	10	18	20	20	7
17	1550	10	31	30	00	7
18	1440	10	18	20	00	3
19	1780	10	14	10	00	7
20	1290	10	26	20	20	7
21	1800	00	02	20	00	3
22	1810	00	12	20	00	3
23	1430	00	01	10	00	3
24	1590	10	00	10	10	3
25	1390	00	02	10	00	3
26	1520	10	00	20	00	3
27	1570	00	06	10	00	3
28	1650	00	25	20	10	7
29	1300	00	00	10	10	3
30	1500	00	04	20	10	7

5.2 Rule reduction of the UCI Statlog heart data set

The rules of the Information System were reduced using the Expert System algorithm with the aid of the Rough Set Data Explorer (ROSE).

ARPN Journal of Engineering and Applied Sciences

Table-3. Rough set rules of the UCI Statlog heart data set.

Num.	Rule	Case	Value of symptoms in cases
1	(C = 40) & (D = 1120) => (M = 3)	M = 3	C = 40, D = 1120
2	(C = 30) & (K = 10) => (M = 3)	M = 3	C = 30, K = 10
3	(I = 10) & (K = 20) & (L = 0) => (M = 3)	M = 3	I = 10, K = 20, L = 0
4	(A = 610) => (M = 3)	M = 3	A = 610
5	(C = 20) & (G = 20) => (M = 3)	M = 3	C = 20, G = 20
6	(E = 3030) => (M = 3)	M = 3	E = 3030
7	(A = 560) => (M = 6)	M = 6	A = 560
8	(B = 10) & (C = 40) & (I = 10) => (M = 7)	M = 7	B = 10, C = 40, I = 10
9	(C = 30) & (I = 0) & (K = 20) => (M = 7)	M = 7	C = 30, I = 0, K = 20
10	(C = 40) & (G = 20) & (I = 0) & (K = 20) => (M = 7)	M = 7	C = 40, G = 20, I = 0, K = 20
11	(A = 650) => (M = 7)	M = 7	A = 650
12	(A = 400) => (M = 7)	M = 7	A = 400
13	(D = 1240) => (M = 7)	M = 7	D = 1240
14	(D = 1350) => (M = 7)	M = 7	D = 1350

Table-3 shows the reduced rules of the Information System and the Empirical Testing Data.

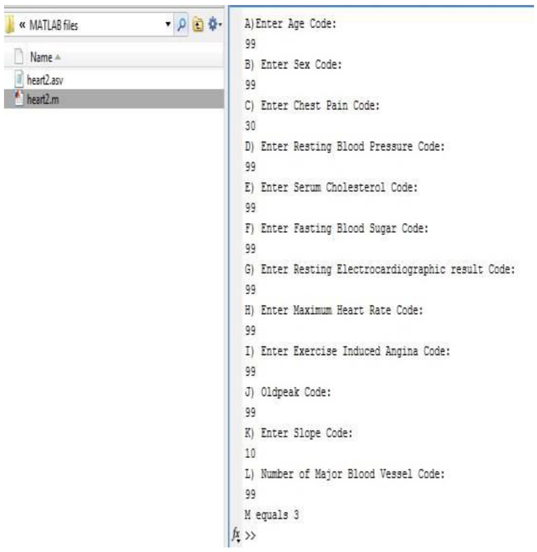

Figure-3. MATLAB implementation of the system.

Figure-3 shows the MATLAB implementation of the system. MATLAB [17] can be used as a tool to apply the rules created by the algorithm with the aid of ROSE.

5.3 Analysis of data

This research showed an Expert System algorithm for diagnosing cardiovascular disease using Rough Set Theory. A program was created in MATLAB to implement the algorithm. A total of 30 sample data from the UCI Statlog Heart Data Set was used to implement the algorithm. Empirical testing was used to verify the accuracy of the algorithm. From the data c = 14 and b = 14. Using the formula for percent validity $c = (b/a) \times 100$. The percent validity of the algorithm is 100%.

6. CONCLUSIONS AND RECOMMENDATIONS

An expert System algorithm for use in the diagnosis of cardiovascular disease was created in this research. The dataset used was the UCI Statlog Heart Data Set. The implementation of the algorithm was done using MATLAB. A total of 30 dataset was used. When the algorithm was applied the rules were reduced to 14 giving a 53.33% reduction. This algorithm cab be used in cardiovascular diagnosis especially when there are missing data or information that cannot be verified. This algorithm can also speed up the process in diagnosis therefore saving precious time.

For future studies it is recommended to try the algorithm to a larger dataset sample. Theoretically the larger the dataset sample this algorithm is used, it will give a more accurate conclusion. This algorithm can also be used to other datasets and can be used as a tool for proper Expert System diagnosis.

REFERENCES

[1] Falk E. 2006. A Pathogenesis of Atherosclerosis. ARPN Journal of Engineering and Applied Sciences. (47)8: 7-12.

[2] Young L., Bamason S. and Kupzyk K. 2016. Mechanism of engaging self-management behavior in rural heart failure patients. Applied Nursing Research. (30)1: 222-227.

[3] Nahar J., Imam T., Tickle K. and Chen Y. 2013. Association rule mining to detect factors which contribute to heart disease in males and females. Expert System with Applications. (40)4: 1086-1093.

[4] Kim K., Roh M. and Ha S. 2015. Expert System based on the arrangement evaluation model for the arrangement design of a submarine. Expert System with Applications. (42)22: 8731-8744.

[5] Jet M. and James M. 2001. An Autonomous Diagnostic and Prognostic Monitoring System for NASA's Deep Space Network. IEEE Aerospace.

[6] Anto S. and Chandramathi S. 2015. An Expert System for Diabetes Diagnosis using Extreme Learning Machine and Simulated Annealing. International Journal of Applied Engineering Research. (10)2: 32087-32101.

[7] Africa A. 2016. A Rough Set Based Data Model for Heart Disease Diagnostics. ARPN Journal of Engineering and Applied Sciences. (11)15: 9350-9357.

[8] Zhou T., Ding J., Wang X. and Xheng X. 2016. Long noncoding RNA and Atherosclerosis. Atherosclerosis. (248)1: 51-61.

[9] Lynch E., Cumming T., Jensenn H. and Bernhardt J. 2015. Early Mobilization after Stroke: Changes in Clinical Opinion despite an Unchanging Evidence Base. Journal of Stroke and Cerebrovascular Diseases. (26)1: 1-6.

[10] Statlog Heart Dataset of the UCI Machine learning repository. 2017. https://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%.

[11] Palwak Z. 1982. Rough Sets. International Journal of Information in Computer Science. (11)5: 341-356.

[12] Palwak Z. 1991. Rough Sets: Theoretical Aspects of Reasoning about data. Boston, MA: Kluwer.

[13] Palwak Z., Slowinsky K. and Slowinsky R. 1998. Rough classification of patients after highly selective vagotomy for duodenal ulcer. International Journal of Information of Machine Studies. (24)1: 413-433.

[14] Anitha K. and Venkatesan E. 2013. Feature Selection by Rough Quick Reduct Algorithm. International Journal of Innovative Research in Science, Engineering and Technology. (2)8: 3989-3993.

[15] ROSE 2.0. 1999. http://www.idss.cs.put.poznan.pl/rose.

[16] Preece A. 2001. Evaluating Verification and Validation Methods in Knowledge Engineering. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.17.3766.

[17] MATLAB 2017. http://www.mathworks.com.