Advances and Applications in Mathematical Sciences
Volume 21, Issue 6, April 2022, Pages 3445-3456

ROUGH SET APPROACH TO ANALYSE CLINICAL DATA OF HEART FAILURE PATIENTS

LEENA SHARMA[1], HEMLATA SAXENA[2] and MINAKSHI PANCHAL[*]

Associate Professor
Applied Science and Hamanities
Pimpri Chinchwad College of Engineering
Nigdi, Pune, India
E-mail: leena.jv@gmail.com

[2]Professor, [3]Research Scholar
Department of Mathematics
Career Point University, Kota (Rajasthan)
E-mail: saxenadrhemlata@gmail.com
mins.29feb@gmail.com

Abstract

Cardiovascular diseases are becoming one of the major reasons for increasing deaths, heart failure is one of them. Always there is a misconception that heart failure means the heart has stopped working. Heart failure can be defined as the heart is not working normally. In the heart, failure death may occur consequently, but it is not only the cause of death. Death in Heart Failure maybe because of the history of the patients, like anaemia, high blood pressure, diabetes, smoking habits, etc. This paper will help to identify leading causes of death in heart failure by using the data reduction which is a key feature of Rough Set Theory (RST). Also, a set of decision rules are defined, which will help to the set of affecting factors which causing death in heart-failure. Aim of the paper is to give a solution to reduce the death rate during heart failure.

1. Introduction

Cardiovascular diseases (CVDs) are one of the most affecting factors on global death rates by the reports of the World Health Organization (WHO).

2020 Mathematics Subject Classification: Primary 94A16; Secondary 68T09.
Keywords: RST, Data Analysis, Reducts, Core, Feature Reduction.

[*]Corresponding author; E-mail: mins@gmail.com

Received September 30, 2021; Accepted December 2, 2021

80% of CVD deaths are because of heart attacks and heart strokes, and 1/3rd deaths due to heart failure occur prematurely in people. WHO takes all measures to prevent, manage, and monitor to reduce the effects of this disease. WHO is also working with different countries to develop cost-effective and equitable health care innovations to handle the disease effectively.

Heart failure is one of the CVDs, it is a chronic progressive condition that disturbs the pumping functioning of your heart muscles. Heart failure specifies the step in which fluid accumulated around the heart and disturbs it from pumping ineffectually. Once heart failure was known as a disease of old age but now it is increasingly occurring in younger people also. Worldwide millions of people are at life threat due to this disease. In Heart Failure, multiple complications are there, such as hospitalization, lethal arrhythmia, and death during the disease progression. Heart failure has also become a growing economic problem worldwide as almost 2% of the healthcare budget was used for heart failure and this percentage may be increase by observing an increasing mortality rate. Not only physically and economically but heart failure impacts badly on the patient's quality of life. So, we can say that heart failure is a social, economic, and medical issue of the current situation. This is high time that we should take it very seriously and consider it as a global health priority.

Nowadays handling and storing big data is a huge problem as data is increase in volume and uncertainty. There are many mathematical methods are available in which Rough Sets Theory is a proven method to deal with vague and uncertain data. So RST has the main advantage over the other methods is that it is very effective for uncertain data also. In medical data where many times we deal with uncertain data without any prior knowledge, RST is a better option. Therefore to avoid the effects of uncertain data on the final results, RST is used for data reduction.

This research tries to apply the RST (RST) to heart failure clinical data to find out the factors causing an increasing death rate. In the clinical data age, anaemia, creatinine, diabetes, ejection fraction, high blood pressure, serum creatinine, serum sodium, and platelet count are the different factors affecting death rate due to heart failure. In this study, we will apply a method of RST to find out the least number of characteristics affecting the

death rate. It will be easier to give more attention to these factors. This study will help to prevent premature death due to heart failure. RST uses indiscernibility relation method for the feature selection from clinical dataset. For the feature selection many machine learning techniques can be used. RST is superior to other technique because of the following points,

- No need of prior knowledge and prior process
- Effective for uncertain data also, so subsequent decisions will be not affected
- Gives optimum number of features which are sufficient to describe the original data.

They found that reducts with a fewer number of attributes give the maximum classification correctness in all datasets. The proposed method of RST uses the indiscernibility relation method to obtain reducts from a set of attributes.

2. Literature Survey

Aaron Don M. Africa [1] applied the RST in the study of determining Angiographic Disease Status (ADS) to diagnose heart disease. The researcher concluded that insufficient data also can be determined by using RST. The results are verified using empirical testing and showed 100% accuracy.

Grzegorz Ilczuk and Alicja Wakulicz-Deja [3] proposed the AQDT-2, which gives decision trees for decision rules by using RST to avoid the difficulty in analysis and validation of decision tables due to the large number and complexity of decision rules. During this research, a tree model was generated from decision rules for comparison of prediction accuracy. It was observed that the AQDT-2 method has severed benefits in the medical field like Graphical representation of decision rules.

Kindie Biredagn, Khanna Nehemiah, Kannan Arputharaj [5] used a back-propagation neural network with RST to classify three clinical datasets. The researchers are given the scope of RST as a complementary method with some optimization techniques.

Qinghua Zhang, Qin Xie, Guoyin Wang [9] briefly introduced the basic

concepts and other rules of RST, as well as different areas of applications, which are discussed. RST is playing a key role in granular computing and optimizing many existing soft computing methods. Need for more work in the process of data mining like effective reduction procedure, parallel computing etc.

B. K. Tripathy, D. P. Acharjya, and V. Cynthya [2] used domain intelligence to generate rules and minimized them by the validation process and threshold value. Using formal concept analysis, more affecting attributes for heart disease can be identified which helps in prior detection of heart disease.

M. Sudha and A. Kumaravel [8] used LERS system-based algorithms, Lem and Modlem (entropy, Laplace method) to study the similarities between rule induction algorithms based on RST. At the same time, the dependency between the quality of classification and the percentage of certain decision rules was examined. Authors emphasize the good quality of classification for the certainty of rules.

3. Preliminaries

Concept of RST was proposed by Zdzislaw Pawlak in 1982, since then continuous progress and knowledge-addition is going on. RST is a methodology to treat uncertain and vague data [12]. RST was constructed on a mathematical foundation and complementary to other methods. In many research works, it is frequently used conjointly with other methods of research e.g., statistical methods, neural networks, genetic algorithms, fuzzy sets, etc.

It is a leading method of data mining or knowledge discovery in relational databases, advantage of this method is that it can be applied to imperfect data [7]. RST possesses a variety of applications in various branches like Artificial Intelligence, Cognitive Sciences, data mining, machine learning, and pattern recognition [4]. Some other uses of RST are given below

Reduce the original data to find essential sets of data without losing the authenticity of the original data. Used for classification according to Rules and hidden patterns in data to find out the relation between the data variables [13]. A good option to statistical methods for finding a relationship

offers straight forward interpretation of obtained results. It does not need previous data like other data analysis methods of probability in statistics. RST is applied in many real-life fields [10], it gives a suitable solution for artificial intelligence methods to industrial procedures like, new materials design and investigating material proper-ties, intelligent control in industrial processes, decision support systems in business decisions, and machine diagnosis of Mechanical objects [15].

Rough sets are used prominently in medical diagnosis since rough sets give methods that are very effective in treating uncertainty in data. From data sets of patients records containing large data, the least set of attributes are obtained through reducts [5]. RST is also applied in the field of stock marketing to analyse the effect on the stock market due to the pandemic situation of Covid [6].

4. Basic Philosophy of RST

RST is based on the concept that every item is associated with some data in terms of data or knowledge e.g. if observations are data of patients suffering from some diseases then symptoms give corresponding information set [11]. Objects with similar characteristics give indiscernible relations, which is the basis of RST. Collectively such a set of all indiscernible observations is called an elementary set which forms the basic atom of knowledge. Unitedly some basic atoms sets form a crisp set (precise) otherwise it is a rough set (imprecise, vague). The rough set has boundary lines, either object certainly belongs to the classified set as a member or belongs to its compliment. In a crisp set, such boundary elements are absent [14].

- **Information system and decision table**

An information system is a data collection in form of a table that provides information about objects and their attributes. Here the set of objects is said to be Universe. Attributes are of two types, condition attributes, and decision attributes. The table consisting of data of objects along with condition and decision attributes is said to be a decision table.

- **Positive Region of R**

Positive region of R contains all that elements of U that can be uniquely classified into partitions U/IND (P) by R, it is denoted by POSR (P), where P denotes attribute. Positive and the boundary regions together represents upper approximation

- **Negative Region of a Subset**

The negative region contains all the elementary sets that does not belong to set Y that is the elements belongs to compliment of Y with respect to relation R.

$$NEGR(Y) = U - R_*(Y)$$

- **Reduct**

In supervised learning, the reduct relative to the decision attribute is useful. Several times may be the information system contains similar or indistinguishable data, some of the attributes may be excessive and unconnected. Without loss of classification performance such data can be extracted. This extracted data or attributes are called reduct.

- **Core**

The attributes that appeared commonly in the set of all reducts are said to be the core. It can be also defined as essential attributes that are common to all genuine reduct and hence these attributes cannot be uninvolved from the information system. In simple words, Core is the intersection of all reducts of an information system.

$$CORE(C) = \cap RED(C)$$

5. Proposed Work

Rough set is a better option for an explanation from uncertain data as it explores the real properties of data even though it's uncertain. When no prior knowledge is available about the process under consideration and complete dependency is on the available data Rough sets are always a better option for data analysis as compared to other existing methods of data analysis. In this study, we consider clinical data of heart failure collected from the website https://www.kaggle.com/andrewmvd/heart-failure-clinical-data. As data is big

there may be uncertainty in measuring the medical facts due to errors in measuring equipment. Uncertainty of data will effect on the quality of subsequent decisions and therefore use of RST for data mining is the most suitable option.

This clinical data contains 300 patients' data of different medical factors. Our aim is to find out the leading causes for death during heart failure. Data 300 rows and 12 columns, here 11 attributes are considered as conditional attributes and Death-Event as a decisional attribute. From the study we get following outputs.

1. All possible Reducts

2. Core

3. Decision rules

The set $A = \{A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_8, A_9, A_{10}, A_{11}\}$ represents conditional attributes an $D = \{D\}$ represents decisional attribute. The description of conditional and decisional attributes and their value set are shown in Table 1

Table1. Conditional and Decisional Attributes and their values.

Sr. No.	Abbreviation Used	Description	Value set of Conditional Attribute
1.	A_1	Age	{20-40, 40-60, 60 and Above}
2.	A_2	Anaemia	{Yes, No}
3.	A_3	Creatinine phosphokinase	{Low, Normal, High, rhabdomyolysis, severe rhabdomyolysis }
4.	A_4	Diabetes	{Yes, No}
5.	A_5	Ejection fraction	{Severely below normal,

			Moderately below normal, Slightly below normal, Normal, High }
6.	A_6	High blood pressure	{Yes, No}
7.	A_7	Platelets	{Critical, Low, Normal, High}
8	A_8	Serum creatinine	{Low, Normal, High, Severe}
9.	A_9	Serum sodium	{Hypernatremia, Normal, Hypernatremia}
10.	A_{10}	Sex	{F, M}
11.	A_{11}	Smoking	{Yes, No}

Information System is $T = \{U, A, D\}$

Where U is the set of all observations:

$$A = \{A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_8, A_9, A_{10}, A_{11}\}$$

$D = \{\text{Decision}\}$

Equivalence class for the data as per decisional attribute's value is:

$$X_1 = \{X \mid \text{decisional value dose Not die}\}$$

$$X_2 = \{X \mid \text{decisional value is Died}\}$$

$$X = \{X_1, X_2\}$$

$$IND_T(B) = \{(x, x') \in U^2 \mid \forall\ a \in B, a(x) = a(x')\}$$

After applying concepts of RST six reducts are identified

- Reduct -1. A feature subset consisting of 7 attributes

  R_1 ={Age, Anaemia, Creatinine phosphokinase, Diabetes, High blood pressure, Serum sodium, Smoking}

- Reduct-2. A feature subset consisting of 8 attributes:
  R_2 ={Age, Anemia, Creatinine phosphokinase, Diabetes, Ejection fraction, Platelets, Serum sodium, Smoking}

- Reduct-3. A feature subset consisting of 8 attributes:
  R_3 ={Age, Anaemia, Creatinine phosphokinase, Diabetes, Platelets, Serum sodium, Sex, Smoking}

- Reduct-4. A feature subset consisting of 8 attributes:
  R_4 ={Age, Anaemia, Diabetes, Ejection fraction, High blood pressure, Serum creatinine, Serum sodium, Smoking}

- Reduct-5. A feature subset consisting of 8 attributes:
  R_5 ={Age, Anaemia, Diabetes, Ejection fraction, Platelets, Serum creatinine, Serum sodium, Smoking}

- Reduct-6. A feature subset consisting of 8 attributes:
  R_6 ={Age, Anaemia, Diabetes, High Blood Pressure, Platelets, Serum creatinine, Serum sodium, Smoking}

Core

$Core = \bigcap \mathrm{Re}\, ducts$

$Core$ ={Age, Anaemia, Diabetes, Serum sodium, Smoking}

Hence the core i.e. factors which are mainly responsible for death in heart failure are Age, Anaemia, Diabetes, Serum sodium and Smoking.

Table 2. Decision rules.

Rule No.	Conditions	Outcome	Laplace Value
1	Ejection Fraction is Severely below normal Creatinine phosphokinase is High and	Death event Yes	0.875

	Anaemia is yes		
2	Age is 60 and Above and Serum sodium is Hyponatremia and A8 is High and A5 is Moderately below normal	Death event Yes	0.75
3	Ejection fraction is Severely below normal and Serum sodium is Hyponatremia and Diabetes is yes	Death event Yes	0.8
4	Age is 60 and Above and Serum creatinine is Normal and Ejection fraction is Normal	Death event Yes	0.6667
5	High blood pressure is Yes and Smoking is Yes	Death event Yes	0.8571
6	Creatinine phosphokinase is Normal and Anaemia is Yes and Diabetes is Yes	Death event Yes	0.75
7	Ejection fraction is Severely below normal and Platelets is Low	Death event Yes	0.75
8	Serum sodium is Normal and Smoking is NO and Age is 40-60	Death event No	0.9375
9	High blood pressure is NO and Age is 40-60 and Serum sodium is Normal	Death event No	0.9412
10	Serum sodium is Normal and the Ejection fraction is Slightly below normal	Death event No	0.9167
11	Serum creatinine is High and Serum sodium is Normal and	Death event No	0.8333

	Diabetes is NO		
12	Age is 40-60 and Ejection Fraction is Moderately below normal and Smoking is NO	Death event No	0.8571

A set consisting of 12 decision rules are identified after applying RST as given Table-2 that specify the decision class of attribute death rate based on its values on some condition attributes.

5. Conclusion

Extraction of usable or essential data from big data is one of the important features of RST. In this paper, this feature is used to identify the most affecting factors causing death in case of heart failure. In this study, a data containing 11 different factors (age, anaemia, creatinine phosphokinase, diabetes, ejection fraction, high blood pressure, serum creatinine, platelets, serum sodium, sex, and smoking) causing death in heart failure are studied.

By the method of reduction, the data is reduced and we identified 5 leading attributes (age, anaemia, diabetes, serum sodium, and smoking) which are more responsible for death event during heart failure. This results will be help full to reduce death rate during heart failure. Major care should be taken for the identified 5 factors which are more affecting and it will result in good health of heart patients.

Also for any other studies related to this data, this reduced form is very useful as it is easy to deal. Also using RST 12 rules are identified with a large number of support datasets, which clearly shows the relation of conditional and decisional attributes i.e. the set of causes which results in death during heart failure.

References

[1] M. Aaron Don, Rough set based data model for heart disease diagnostics, ARPN Journal of Engineering and Applied Sciences 11 (2016), 9350-9357.

[2] B. K. Tripathy, D. P. Acharjya and V. Cynthya, A framework for intelligent medical Diagnosis using rough set with formal Concept analysis, International Journal of Artificial Intelligence and Applications (IJAIA) 2(2) (2011), 45-66.

[3] Grzegorz Ilczuk and Alicja Wakulicz-Deja, Visualization of rough set decision rules for

medical diagnosis systems, Proceedings of the 11th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing (2009), 371-378.

[4] R. Kalaivani, M. V. Suresh and N. Srinivasan, A study of rough sets theory and it's application over various fields, Journal of Applied Science and Engineering Methodologies 3(2) (2017), 447-455.

[5] Kindie Biredagn Nahato, Khanna Nehemiah Harichandran and Kannan Arputharaj, Knowledge mining from clinical datasets using rough sets and backpropagation neural network, Computational and Mathematical Methods in Medicine, (2015).

[6] Leena Sharma, Aditya V. Yadav, Shravani P. Ahirrao, Varun S. Manik and Aman S. Ramani, Stock market analysis of 10 different countries in the period of disease COVID-19, International Journal of Engineering and Management Research 10(4) (2020).

[7] T. Y. Lin and N. Cercone, Rough sets and data mining-analysis of imperfect data, Kluwer Academic Publishers, Boston, London, Dordrecht (1997), 430-435.

[8] M. Sudha and A. Kumaravel, Quality of classification with, LERS system in the data size context, Applied Computing and Informatics 16 (2018), 29-38.

[9] Qinghua Zhang, Qin Xie and Guoyin Wang, A survey on RST and its applications, CAAI Transactions on Intelligence Technology 1(4) (2016), 323-333.

[10] Silvia Rissino and Germano Lambert-Torres, RST-Fundamental Concepts, Principals, Data Extraction, and Applications, Data Mining and Knowledge Discovery in Real Life Applications, Book edited by: Julio Ponce and Adem Karahoca (2009), 438

[11] Zbigniew Suraj, An introduction to RST and its applications a tutorial, 6th International Conference, RSKT 2011, Banff, Canada, October (2011), 9-12

[12] Z. Pawlak, Rough Sets, International Journal of Computer and Information Sciences, 11 (1982), 341-356.

[13] Z. Pawlak, Rough set approach to knowledge-based decision support, European Journal of Operational Research 99 (1997), 48-57.

[14] Z. Pawlak, Rough sets and intelligent data analysis, Information Sciences 147 (2002), 1-12.

[15] Z. Pawlak, RST for intelligent industrial applications, Second International Conference on Intelligent Processing and Manufacturing of Materials, IPMM'99 (Cat. No.99EX2).