



## PERFORMANCE COMPARISON ON FINDING PREDICTOR IMPORTANCE USING ROUGH SET THEORY AND REGRESSION

M. SUDHA and A. KUMARAVEL

Research Scholar, Department of Mathematics  
AMET Deemed to be University  
Kanathur, Chennai-603112, India  
E-mail: seedinmenew@yahoo.com

Dean, School of Computing  
Bharath University, Selaiyur  
Chennai-600073, India  
E-mail: drkumaravel@gmail.com

### Abstract

It has been two decades, since the Professor Zdzislaw Pawlak introduces the Rough set theory. The Rough set theory has attracts the researchers still for its efficiency and simplicity and many methods have been successfully created with RST so far. But, it has some issues with some areas which have become research for most of the people now. Many comparative studies has done on checking its efficiency with other algorithms and concepts. Most of the time, the significance level with RST is negligible. Moreover, RST has successfully applied in many fields like medicine, pattern recognition, as expert systems, knowledge discovery, information system, inductive reasoning, intelligent systems, data mining, pattern recognition, decision-making, and machine learning. RST has developed with many extensions and generalisations. In this paper, the dimensionality reduction of data is experimented with the concept of the rough set theory and regression. Heart failure data is used to test he efficiency and their results interpreted with the rate of selection of attributes based on the prediction.

### Introduction

Machine learning is a mathematical technique that gives the machine the ability to construct a model for learning and enhancing the performance of a

---

2020 Mathematics Subject Classification: 34Dxx, 93Dxx.

Keywords: Rough set theory, linear regression, Heart failure data, Data mining, Attribute selection, Prediction.

Received October 5, 2020; Accepted November 10, 2020

specific data function. This role hits the issues around supervised learning, clustering, reduction in dimensions and future prediction. Machine learning algorithms resolve the issues, and make decisions and predict results. Classification and regression are the function of supervised learning where clustering is not supervised [2]. Classification is the problem of defining a new pattern based on training data analysis and the pattern applied by the classification known as classifier. This role is widely needed because in the reduced space the data analysis such as classification and regression is more reliable and also the size of the data is large and becoming complicated in this advanced world. It also improves classifier performance, reduces time and space, and allows easy visualization in 2D or 3D. The techniques like PCA, LDA, GDA, CCA etc., are been used to resolve this so far [3]. The Rough set theory has been noticed as an efficient tool in machine learning for the past decades and it has been successfully overcome the above underlined tasks in many fields [4, 5]. The RST-based algorithm called the LERS method being proof for it was announced by NASA's Johnson Space Center as the successful application of the RST in data mining by adopting LERS [9] as an expert system creation tool [6]. Some institutions have made use of this helpful software about the rough sets. ROSE, ROSETTA, RSES, RStudio [10] are the tools having the concepts of the Rough set theory and the algorithms based on that [11]. Among them ROSE tool has all the fundamentals of the Rough set theory. However, issues like too much running time, lack of current RST based algorithms, and low acceptance of big data are being found make the researchers felt lacking sometimes. In this paper, we try to figure out some issues faced by the beginners with the mentioned tool and with the concept of the Rough set theory

**Methods and Materials:** We apply several machine learning classifiers to predict the patient's survival. RST can be described by means of lower and upper approximations. The set of granules is the universal set  $U$  and now the set is defined with respect to  $R$  where  $R$  is the equivalence relation assumed based on the knowledge prescribed in granules of  $U$ . To describe the vague part of the set  $X$  with respect to  $R$ , we need the approximations of rough set theory [12, 13].

**Rule induction:** LERS, the Rough Set Learning Examples is the successful minimal algorithm used to generate decision rules in the structure

of if and then rules with a perfect pattern. Pattern is the knowledge calculated by all instances regarding the set of attributes and it can identify or test any instance that belongs to that knowledge. In three approaches, it induces rules; minimum set, exhaustive set, and satisfactory set. Maximum set produces a maximum number of rules that are adequate to convey all instances. Exhaustive set uses examples to induce all rules. Satisfactory setting leads to rules that satisfy user-defined requirements. In rule induction, it is important to process numerical attribute transformation into symbolic attribute. This process of discretization may vary by algorithm. The principles of decision and their algorithms have been discussed in [14]. Decision class approximations are described as three types, i.e. minimum set, exhaustive set and satisfactory set of decision rules [15]. For this study, ROSE2 [16] is a modular software system that uses the basic concepts of the rough set theory and the methods of rule finding.

**DATASET:** We analyzed a dataset containing the medical records of 299 heart failure patients collected from UCI repository [7, 8]. The patients consisted of 105 women and 194 men, and their ages range between 40 and 95 years old. The details of the attributes are described below.

Heart failure, also known as congestive heart failure, happens when blood is not pumped as efficiently as it by the heart muscle. Many problems, like narrowed arteries in your heart (coronary artery disease) or high blood pressure, slowly leave the heart too weak or rigid to effectively fill and pump. It is not possible to cure all problems that lead to heart failure but therapies will reduce the signs and symptoms of heart failure and help you live longer. One way to reduce heart failure is to stop and monitor heart failure problems such as coronary artery disease, high blood pressure, diabetes or obesity [1].

**Table1.** Attributes description.

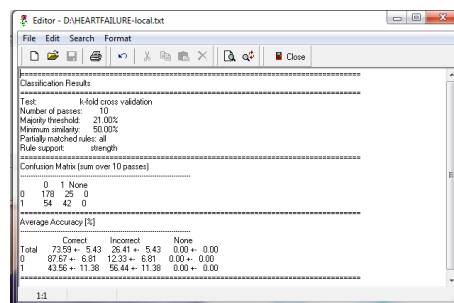
Feature	Explanation	Measurement
Age Anaemia	Age of the patient Decrease of red blood cells or hemoglobin	Years Boolean
High blood pressure Creatinine phosphokinase (CPK)	If a patient has hypertension Level of the CPK enzyme in the blood	Boolean mcg/L
Diabetes Ejection fraction	If the patient has diabetes Percentage of blood leaving the heart at each contraction	Boolean Percentage
Sex Platelets	Woman or man Platelets in the blood	Binary kiloplatelets/m L
Serum creatinine Serum sodium Smoking Time Class: death event	Level of creatinine in the blood Level of sodium in the blood If the patient smokes Follow-up period If the patient died during the follow-up period	mg/dL mEq/L Boolean Days Boolean

The above data was carried out for classification based on RST. For that we used the ROSE 2.2 tool which is completely constructed using the fundamentals of RST. Initially the RST calculated the approximations of the dataset based on the number of equivalent sets. The lower and upper approximation gives the quality of classification of the dataset which should not be changed or lessened while mining. Here the dataset is having quality of classification is equals to 1, tells this dataset is a crisp set and the number of equivalence sets (atoms) is 299. Afterward the boundary region is empty here.

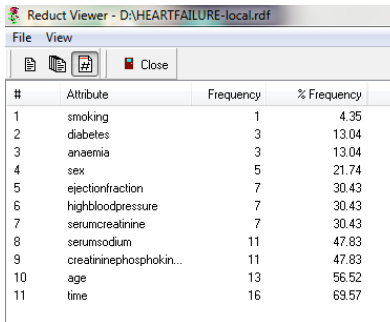
Based on the RST reduction concept, here the CORE attributes are none because of vagueness not found in the boundary region. Also reducts are found using discernibility matrix which are detailed in the table. To find the best reduct we need to use frequency of attributes to get the equal number of atoms that are in the original set. Otherwise the reduct set will loss the original information.

The percentage of accuracy of the classification by stratified cross validation using basic minimal covering is 73.59%, which is found by the number of appropriately classified elements over the total amount of elements [2]. The number of correctly classified instances is 178 and incorrectly classified is 42. The other parameters like 50% similarity of partially matched rules with 21% of majority threshold in voting and rule strength in class support have given similar percentages of accuracy only vary in decimals. Among we found, the mentioned percentage is the highest.

Though the rough set is best for reduction, here the set is a crisp set. The original scores well and need not to be reduced. But the frequency based reduct has found among the reducts created in the tool is shown in figure [2]. From that, we got four irrelevant attributes (sex, anemia, diabetes, platelets and smoking) as they are very low in frequency of threshold 0.30. The accuracy of the same was calculated by the cross validation is not up to the original accuracy (Figure 1). It is decreased to 69.52%. So we pushed to take the original set accuracy instead the reduct.



**Figure 1.** Accuracy of validation.



The screenshot shows a window titled 'Reduct Viewer - D:\HEARTFAILURE-local.rdf'. The window contains a table with the following data:

#	Attribute	Frequency	% Frequency
1	smoking	1	4.35
2	diabetes	3	13.04
3	anaemia	3	13.04
4	sex	5	21.74
5	ejectionfraction	7	30.43
6	highbloodpressure	7	30.43
7	serumcreatinine	7	30.43
8	serumsodium	11	47.83
9	creatininephosphokin...	11	47.83
10	age	13	56.52
11	time	16	69.57

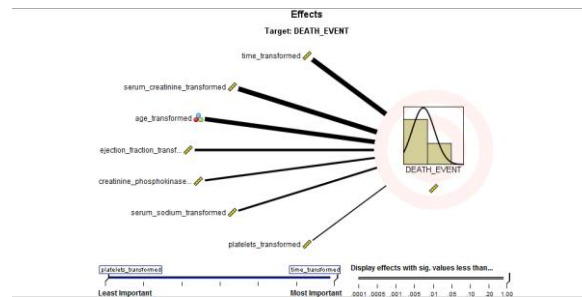
**Figure 2.** Frequency of Reducts.

**Linear Regression with SPSS modeler tool:** Regression models are used to define the unknown relationship between multiple condition variables and a decision variable. It can predict the value of the decision variable using the dependencies between the variables. The cross validation validate the model and prevent over fitting, the build model is normally applied to test data to verify its predicting pattern. Particularly to evaluate classification models, such as KNN or SVM. For further information on the concept and variants of cross-validation refer [19]. This process is to split data into a training dataset 70% and a test dataset 40%. The training dataset is used to assess the parameters of model and to build it. The testing dataset is to execute the model with the considered parameters which helps to measure the ability of model and to find the specified pattern to predict future. The IBM SPSS Modeler is a powerful tool [18], which is widely used to analyzing data and developing predictive models. Here created the models can be organized and executed in data analytics production processes. It helps to handle data analytics methods in an appropriate way with all the recent algorithms.

**Predictor Importance:** The above heart failure data was processed with linear aggression in SPSS modeler tool. The results of them are given below.

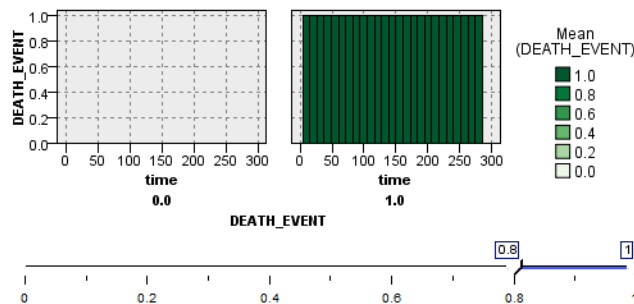
The model built by regression shows the importance of the attributes on the basis of prediction accuracy. We know that once the regression models build from the training set, the best model fit can be find by evaluation of validation set and it will be executed in the test data to measure the prediction with the expected error. Actually the root mean square (RME) is used to measure fitness of the model and helps to find the best among the all

models. RMSE is the standard deviation of difference between the Actual and predicted and it is a measure for model accuracy used in cross-validation to match model and their estimating errors with each other which helps to decide the best model [20, 21]. The built model with least root mean square using reference model with the accuracy of 41.8%. The predictor importance calculated by the regression between the target and the other predictors (inputs) by the build model is shown below.



**Figure 3.** Predictor importance.

From the above we can see the relevant attributes by the correlation between the input variables and the target. The deviation between the target and the predictor ‘time’ is shown below. Here time is highly correlated and its dependency level lies between 0.8 and 1 (figure 3).



**Figure 4.** Dependency between ‘Death event’ and ‘Time’.

The screenshot displays two sections of statistical data. The first section, labeled 'smoking', shows a table of statistics and a Pearson correlation table. The second section, labeled 'time', also shows a table of statistics and a Pearson correlation table. Both sections indicate the strength of the correlation with 'DEATH\_EVENT'.

smoking	
Statistics	
Count	299
Mean	0.321
Min	0.000
Max	1.000
Range	1.000
Variance	0.219
Standard Deviation	0.468
Standard Error of Mean	0.027
Pearson Correlations	
DEATH_EVENT	-0.013 Weak

time	
Statistics	
Count	299
Mean	130.261
Min	4.000
Max	285.000
Range	281.000
Variance	6023.965
Standard Deviation	77.614
Standard Error of Mean	4.489
Pearson Correlations	
DEATH_EVENT	-0.527 Strong

**Figure 5.** Importance predictors by Pearson correlation.

With the help of least error value we can identify the weak and strong attribute when it is correlated with the target (figure 5).

**Discussion:** The rough set theory is an efficient tool for attribute reduction where you can see the relevant and irrelevant attributes. RST based on the approximation of lower and upper bound that helps to measure the accuracy and quality of the classification. The quality of the classification is nothing but the number of equivalence classes attained by the relation defined as the target. In the reduction concept, the RST produces the core attributes by finding the intersection of all reductions. The core attributes are the attributes cannot be removed from the system. If the core attributes have the same quality as the original set has, then the other attributes other than core are irrelevant. They can be removed from the system. If it is not in the case, we should find the important attributes among the remaining attributes which may score quality to the reduct. Here in Heart Failure data analysis, the quality of classification is 1 and there is no core attributes since the reducts are in more numbers and they have no attributes in common. In this case, we go for finding important attributes using frequency of attributes. We should add an attribute of high frequency to the core and we should check the quality of the new reduct. We should repeat this until we get the quality of the original dataset with threshold. Here the threshold is 0.3 and we found four irrelevant attributes (table 2). The selected attributes with ranking is listed in table.



**Table 2.** Attribute selection of RST and Regression.

Attributes	Rough set theory	Linear regression
	Frequency	Predictor importance
time	69.57	0.57
age	56.52	0.01
Creatinine phosphokinase (CPK)	47.83	0.06
Serum sodium	47.83	0.04
Serum creatinine	30.43	0.27
High blood pressure	30.43	0.01
Ejection fraction	30.43	0.04
Sex	21.74	0.00
Anemia	13.04	0.00
diabetes	13.04	0.00
Smoking	4.35	0.00

**Rules induction:** In Rough Set theory, the rules are induced by LERS system with the rule support and strength by minimal algorithm. The no. of rules induced are 16 whereas in the association rule method induced rules generate 30 interesting rules with rule support.

### Conclusion

In our work, the analysis with the concept of Rough set theory has selected Time, age and creatinine phosphokinase as relevant attributes for this heart failure medical records which is confirmed by the relevance of RST whereas in regression, it is in different way. The selection of attributes has been done by correlation between the variables and the standard deviation of difference between the Actual and predicted. The error rate was mainly considered to get the weak and strong attribute which gives the final ranking list based on the predictor importance. Even though the selection of attribute

is different, both concept selected the attributes 'time' and 'age' as strong attributes. Also they are have high frequency and predictor importance. The remaining selected attributes other than the above two are different. But the performance of prediction between them are negligible. Since many hybrid methods have been developed with regression, a bio inspired algorithm with the regression and with the selection concept of RST will be studied in future.

### References

- [1] Davide Chicco and Giuseppe Jurman, Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone, *BMC Medical Informatics and Decision Making* (2020), 2-16.
- [2] Ian H. Witten and Eibe Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Second Edition, Elsevier 62-82.
- [3] C. Ding, X. He, H. Zha and H.D. Simon, Adaptive Dimension Reduction for Clustering High Dimensional Data, *Proceedings of International Conference on Data Mining*, 2002.
- [4] Z. Pawlak, *Rough Sets and Intelligent Data Analysis*, *Information Sciences* 147(1) (2002), 1-12.
- [5] R. Mienko, J. Stefanowski, K. Taumi and D. Vanderpooten, *Discovery-Oriented Induction of Decision Rules Cahier du Lamsade*, No. 141, Université Paris Dauphine 1996.
- [6] J. W. Grzymala-Busse, *Rough Set Theory with Applications to Data Mining Real World Applications of Computational Intelligence* Springer, Heidelberg, 2004.
- [7] T. Ahmad, A. Munir, SH. Bhatti, M. Aftab MA. Raza, Survival analysis of heart failure patients: a case study *PLoS ONE*12 (7) 0181001, 2017.
- [8] T. Ahmad, A. Munir, SH. Bhatti, M. Aftab, M. Ali Raza Survival analysis of heart failure patients: a case study. Dataset Accessed 25, [https://plos.figshare.com/articles/Survival\\_analysis\\_of\\_heart\\_failure\\_patients\\_A\\_case\\_study/5227684/1](https://plos.figshare.com/articles/Survival_analysis_of_heart_failure_patients_A_case_study/5227684/1) 2019.
- [9] J. W. Grzymala-Busse, LERS-a system for learning from examples based on rough sets. In: Slowinski R, editor. *Intelligent decision support handbook of applications and advances of the rough sets theory*, Kluwer Academic Publishers (1992), 3-18.
- [10] Z. Abbas and A. Burney, A Survey of Software Packages Used for Rough Set Analysis, *Journal of Computer and Communications* 4 (2016), 10-18.
- [11] T. Slimani, Application of Rough Set Theory in Data Mining, *International Journal of Computer Science and Network Solutions* 1(3) (2013), 1-10.
- [12] J. W. Grzymala-Busse, *Rough Set Theory with Applications to Data Mining Real World Applications of Computational Intelligence* Springer, Heidelberg, 2004.
- [13] T. Ahmad, A. Munir, SH. Bhatti, M. Aftab and MA. Raza, Survival analysis of heart

failure patients: a case study PLoS ONE12 (7) 0181001, 2017.

- [14] Zdzislaw Pawlak, Rough Sets and Decision Algorithms, W. Ziarko and Y. Yao (Eds.): RSCTC 2000, LNAI 2005, (2001), 30-45.
- [15] L. Polkowski and A. Skowron, (Eds.), Rough Sets in Data Mining and Knowledge Discovery, PhysicaVerlag 1 (1998), 500-529.
- [16] ROSE Software <http://idss.cs.put.poznan.pl/site/rose.html>.
- [17] H. Almarabeh, Analysis of Students Performance by Using Different Data Mining Classifiers, I. J. Modern Education and Computer Science 9(8) (2017), 9-15.
- [18] SPSS MODELER: <https://www.ibm.com/support/pages/downloading-ibm-spssmodeler-1821>.
- [19] G. James, D. Witten, T. Hastie and R. Tibshirani, An introduction to statistical learning New York: Springer, 103 (2013).
- [20] R. J. Hyndman, and A. B. Koehler, Another look at measures of forecast accuracy International Journal of Forecasting 22(4) (2006), 679-688.
- [21] M. Kuhn and K. Johnson, Applied predictive modeling New York: Springer, (2013).