Practical Assessment, Research, and Evaluation

Volume 27 Article 25

2022

Rough Set Theory (RST) Data Analysis with R and its Application on Studying Relative Significances of Self-Regulated Learning (SRL) Strategies of Gifted Students

Tze-ho Fung
Hong Kong Academy for Gifted Education

Wing-yi Li
Hong Kong Academy for Gifted Education

Recommended Citation

Fung, Tze-ho and Li, Wing-yi (2022) "Rough Set Theory (RST) Data Analysis with R and its Application on Studying Relative Significances of Self-Regulated Learning (SRL) Strategies of Gifted Students," *Practical Assessment, Research, and Evaluation*: Vol. 27, Article 25.
Available at: https://scholarworks.umass.edu/pare/vol27/iss1/25

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 27 Number 25, September 2022 ISSN 1531-7714

Rough Set Theory Data Analysis with R and its Application on Studying Relative Significances of Self-Regulated Learning Strategies of Gifted Students

Tze-ho Fung, *Hong Kong Academy for Gifted Education*
Wing-yi Li, *Hong Kong Academy for Gifted Education*

Rough set theory (RST) was proposed by Zdzistaw Pawlak (Pawlak,1982) as a methodology for data analysis using the notion of discernibility of objects based on their attribute values. The main advantage of using RST approach is that it does not need additional assumptions – like data distribution in statistical analysis. Besides, it provides efficient algorithms and tools for finding hidden patterns. Despite its advantages, the adoption of RST in educational research is still limited, which may be due to limited quality software available for RST. Recently, the *RoughSets* package in R statistical system has been developed (Riza et al., 2014), providing various utilities of the RST methods. In the paper, we will first describe the basic RST concepts and steps of data analysis under RST. We will then apply RST to a study, which aimed to determine the relative significance of various SRL strategies used by student participants, so as to illustrate the steps of data analysis using the *RoughSets* R package. From the illustration, it is expected that more researchers in the field of education will be encouraged to try RST methods in their own studies.

Keywords: rough set theory, discernibility, *RoughSets* R package, online learning, self-regulated learning strategies

Introduction

Rough set theory (RST) was proposed by Zdzisław Pawlak (Pawlak, 1982) as a methodology for data analysis. It revolves around the notion of the approximation of concepts in information systems and the fundamental concepts of discernibility, which is the ability to distinguish between objects, based on their attribute values. Given an indiscernibility relation, equivalence classes within the given data could be established. All the data tuples forming an equivalence class are indiscernible, that is, the samples are identical with respect to the attributes describing the data. In a real-world context, it is common that some concepts (e.g., persons with a certain disease) cannot be uniquely distinguished in terms of the available attributes (e.g., symptoms or testing results of a person). Based on these equivalence classes, we can construct lower and upper approximations of concepts (e.g., the presence of a disease in a person). Objects included in the lower approximation can be classified with certainty based on the attribute values as members of the concept. In contrast, the upper approximation contains objects possibly belonging to the concept, i.e., with the same set of attribute values, some of them belong to the concept, while some of them do not. Furthermore, dependencies between the concept concerned and attributes available could be defined and measured. An important advantage of RST is that it does not require additional parameters or assumptions to analyse the data.

For more than three decades RST has been attracting researchers and practitioners in many different areas. For example, RST is applied in diverse domains such as water quality analysis (Karami et al., 2014), intrusion detection (Ma et al., 2013), bioinformatics (Li et al., 2014), and character pattern recognition (Liu, 2014). However, its application in educational research is still not popular nowadays. One of the reasons is that researchers in the educational field are not familiar with the basic notions of RST and the basic steps to implement the data analysis under the RST approach. Besides, a lack of popular software that provide comprehensive facilities for an implementation of fundamental concepts of RST for academic research impedes the adoption of RST in educational research. Recently, the *RoughSets* package in R statistical system has been developed (Riza et al., 2014) that allows researchers and practitioners to explore both the basic knowledge of the theories and their applications. It was written in the R language, which is a widely used analysis environment for scientific computing and visualization, such as statistics, data mining, bioinformatics, and machine learning. Currently, over 5000 packages are included in the repositories of the Comprehensive R Archive Network (CRAN) and the Bioconductor project. The *RoughSets* R package is free for download on CRAN.

In the remaining parts of the paper, we will first describe the basic concepts related to RST. We will then apply RST to a study, which aimed to determine the relative significance of various self-regulated learning (SRL) strategies used by student participants, so as to illustrate the steps of data analysis using the utilities provided in *RoughSets* R package. The main contribution of our study is that from the introduction of the rudiments of RST and *RoughSets* R package, and the illustration of a case study using a stepwise manner, more researchers in the field of education will have the basic understanding and knowledge of RST in order to try these methods in their own studies under the popular R statistical environment.

Rudiments of Rough Set Theory

Basic Concepts

The rough set approach was proposed as a tool for dealing with imperfect knowledge, in particular with vague concepts. Rough set theory has gained interest of many researchers and practitioners from all over the world. In the following, the basic concepts of RST and the analysis tools related to the present study are briefly described.

Rough Set Theory (RST) is a fundamental mathematical tool for studying uncertainty that may arise in various areas closely related to data analysis (Lin & Cercone, 1997; Orlowska, 1998; Slowinski, 1992; Pawlak, 1991). The main advantage of standard rough set theory in data analysis is that it does not need any preliminary or additional information about data like probability distributions in statistics, basic probability assignments in Dempster–Shafer theory, a grade of membership or the value of possibility in fuzzy set theory. Thus, it is sometimes called a non-invasive data analysis approach.

When an object (say, object *a*) possesses the same set of information (e.g., same values for a certain set of attributes) with others, this class of objects (denoted as $R(a)$) are indiscernible (similar) with respect to the available information about them. The equivalence class ($R(a)$) generated from this indiscernibility is the mathematical basis of RST, which is called an elementary set and forms a basic granule of knowledge about the domain concerned. In the standard RST, $R(a)$ is defined by values of a (sub)set of attributes possessed by the object. RST approximates another set (called a concept or decision attribute) concerned (say D) using a pair of sets named the lower and upper approximation of the set, namely: the sets $R_{low}(D)$ and $R_{up}(D)$. With respect to the set of attributes defining the relation R, the set $R_{low}(D)$ is consisted of all those elementary sets, which are certainly in the set D (i.e., the elementary sets are subsets of D). The set $R_{up}(D)$ is consisted of all those elementary sets, which have the possibility of belonging to the set D (i.e., their intersections with D are non-empty). The boundary region of D is defined as $R_{Br}(D) = R_{up}(D) - R_{low}(D)$. Based on these two approximations, a set D is said to be crisp (with respect to R) if the boundary region of D is empty; otherwise, D is said to be rough.

Attribute Dependency

One of the important aspects of data analysis is the discovery of attribute dependencies, i.e., it is aimed to discover which attributes are strongly related with the target concept concerned and thus to retain only those mostly related attributes for predictive modelling. In rough set theory, the notion of dependency is defined very simple without imposing any stringent assumptions such as normality and linear relationship.

Specifically, let us consider the target concept *D* and its complement $D^c = U - D$, where *U* stands for the universal set, i.e., the set consisting of all the objects concerned. Then the dependency of the target concept *D* on attribute set R is denoted as $\gamma_R(D)$, which is defined as follows:

$\gamma_R(D)$= *(cardinality of the* $R_{low}(D)$ *+ cardinality of the* $R_{low}(D^c)$*)/ cardinality of U*

The cardinality of a set is simply the number of elements in the set. $\gamma_R(D)$ is the proportion of objects that could be correctly classified as either belonging to *D* or belonging to D^c based on the attribute set R. Such a definition could be easily generalized when the target concept could take on *n* distinct values, i.e., the universal set *U* can be mutually and exclusively classified into $\{D_1, D_2\ldots, D_n\}$. Accordingly, the significance of an attribute, say A_1 could be assessed using the change in the dependency due to dropping the attribute concerned, i.e., the significance of the attribute $A_1 = \gamma_R(D) - \gamma_{R-\{A1\}}(D)$.

Reduct – Minimal Set of Attributes

A related interesting question is whether there is a subset of all attributes which can "almost" fully characterize the information in the data set. Such a minimal set of attributes is called a reduct in RST. To derive a reduct, we first decide a quality measure of a subset of attributes, say *quality F*. In the *RoughSets* R package, a number of options are provided to users. One of the possible candidates is the entropy of the attribute subset concerned, which is used to assess the amount of information gained or lost after the addition or deletion of an attribute. The current study chose entropy as the measure of quality when searching a reduct. Additionally, we employ the epsilon parameter in order to compute an ε-approximate reduct, which is defined as an irreducible subset of attributes *R'* with respect to the full set of attributes R such that:

Quality(R') $\geq (1-\varepsilon)$ *Quality* (R)

It should be noted that *Quality*(R') and *Quality*(R) are the values of quality measures respectively for the attribute sets *R'* and R and ε is a numeric value between 0 and 1, expressing the approximation threshold. In this study, ε was set to 0.1. Under such a criterion, it is expected that the dependency of target concept *D* on the ε-approximate reduct *R'* would remain more or less the same with that of the full set of attributes R.

Rule Induction

After determining the significance of each attribute and the reduct, the next important question is to use the attributes remained in the reduct to predict the decision attribute. This kind of decision, determining the value of a decision attribute based on the values of other attributes, is frequently expressed in terms of rules and rule induction is one of the fundamental tools in data mining. They are in the form of:

if (attribute-1; value-1) and (attribute-2; value-2) and … (attribute-n; value-n)

then (decision; value)

It implies that when an object's *attribute-1* has *value-1*, and its *attribute-2* has *value-2*…and its *attribute-n* has *value-n* (the condition), the object's attribute of *decision* will possess the corresponding *value* (the conclusion). In RST, Learning from Examples Module version 2 (LEM2; Dhandayudam & Krishnamurth, 2013) is a common rule induction algorithm, which was used in the study. It finds regularities hidden in the data by treating all possible attribute-value pairs as the searching space and express the regularities found in terms of rules, representing a "local" covering for each class of the decision attribute. When applying a set of derived rules to a new object, in general, more than one rules will be applicable for the object, i.e., the condition of a rule is fulfilled. Each rule will have a weight. For each decision class, the weights of the corresponding applicable rules will be aggregated. The object will be assigned to the decision class with the "heaviest" aggregated weights. In the following, LEM2 is briefly outlined.

First, we define a block of an attribute-value pair *t = (A, v)*, denoted by *[t]*, is the set of all examples that for attribute *A* have value *v*. A concept, described by the value *w* of the decision attribute *D*, is denoted *[(D, w)]*, and it is the set of all examples that have value *w* for the decision attribute *D*. Now let *B* be a concept and let *T* be a set of attribute-value pairs. The concept *B* depends on a set *T* if and only if

[T] = Intersection of all *[t]* in $T \subseteq B$

Set *T* is a minimal complex of concept *B* if and only if *B* depends on *T* and *T* is minimal, i.e., *B* no longer depends on *T* if any one element in *T* is dropped. Let τ be a nonempty collection of nonempty sets of attribute-value pairs. Set τ is a "local" covering of *B* if and only if the following three conditions are satisfied:

1. Each member of τ is a minimal complex of *B*,
2. Union of all *[T]* where *T* in τ = *B*, and
3. τ is minimal (τ has the smallest possible number of members)

For a set *X*, | *X* | denotes the cardinality of *X*. The pseudo-code of the procedure of LEM2 is shown in Table 1 for the sake of easy understanding.

Rough Set Theory (RST) Data Analysis

The advantages of using RST approach in data analysis are summarized in the following:

1. It does not need any preliminary or additional information about data – like probability in statistics.
2. It provides efficient methods, algorithms, and tools for finding hidden patterns.
3. It allows to reduce original data, i.e., to find minimal sets of data with "more or less" the same knowledge as in the original data.
4. It allows to evaluate the significance of data (attributes).
5. It allows to generate in automatic way the set of decision rules from data, which is easy to understand.

RST can be used to discover structural relationships within imprecise or noisy data. Under the classical RST, it only applies to discrete-valued attributes. Continuous-valued attributes must therefore be discretized before its use. The basic steps of data analysis under RST approach are as follows:

Step 1. Characterization of a set of objects in terms of discrete attribute values.

Step 2. Finding dependency between the attributes and reducing superfluous attributes by retaining the significant ones.

Step 3. Decision rule generation.

Table 1. The Pseudo-code of LEM2 Algorithm for Finding a Local Covering of the Concept B

```
Input: The concept $B$ concerned.

Start: Let the Goal Set, $G$ be the concept $B$. Let the local covering, $\tau$ be empty.
Beginning of Loop I: While $G$ is not empty, continue to do the following:
    Let the set $T$ be empty.
    Beginning of Loop II: While $T$ is empty or $[T]$ is larger than $B$, continue to do the following:
        1. Let $T(G)$ be the set of attribute-value pairs, $t$ s.t. $t$ is not in $T$ and intersection
           of $[t]$ with $G$ is not empty.
        2. From $T(G)$, choose the attribute value, $t$ s.t. amongst all elements in $T(G)$ s.t.
            a. $|[t] \cap G|$ is the maximum. If a tie occurs, select the one with the
               smallest $|[t]|$.
        3. Update $T$ and $G$: Add $t$ to $T$ and $[t] \cap G \rightarrow G$.
        4. Go back to the Beginning of Loop II above.
    End of Loop II
Check the resultant set $T$ is minimal, in term of set inclusion.
Add $T$ to $\tau$.
    Update $G$: $G$ – (Union of all $[T]$ in $\tau$) $\rightarrow G$.
    Go back to the Beginning of Loop I above.
    Check the resultant set $\tau$ is minimal, in term of set inclusion.
End of Loop I
```

To apply RST on datasets, a number of software systems had been developed in the past. Amongst them, the well-known packages included Rose2 (Rough Sets Data Explorer), Rosetta (Rough Set Toolkit for Analysis of Data), RSES (Rough Set Exploration System) and Rough Set Analysis provided in WEKA (Waikato Environment for Knowledge Analysis). These packages were developed using C++/Java during mid/late 20th century (Abbas & Burney 2016). Some of them are now not under active development and maintenance. Recently, the R package, *RoughSets*, which facilitates data analysis using techniques put forth by Rough Set and Fuzzy Rough Set Theories, has been available for free. It does not only provide implementations for basic concepts of RST and FRST but also popular algorithms that derive from those theories. Besides, a large number of statistical and graphical utilities provided under R statistical platform could be directly tapped. Despite its advantages, the adoption of RST in educational research is still limited. In the following, we illustrate its application in assessing the relative significances of various SRL strategies and their associations with the online learning performance of students using the utilities provided in *RoughSets* R package.

Study on the Relative Significances of Various SRL strategies

Background and Aims

Research on SRL has been emerged more than two decades. Over the past decades, different researchers have defined SRL in various ways and offered different models of SRL (Zimmerman & Martinez-Pons, 1988; Pintrich, 2000). Despite there is no single agreed upon model for what the various components of SRL strategies are, it is commonly agreed that students who utilize SRL strategies tend to perform better in their learning (Pintrich et al., 1993; Zimmerman & Martinez-Pons, 1986). This relationship was demonstrated with students across subject areas and grade levels (Hattie et al., 1996; Dignath et al., 2008). In recent years, much more work focuses on the impact of SRL within the context of online or computer-assisted environments (Azevedo, 2004; Winne et al., 2006). Several studies of Massive Open Online Courses (MOOCs) have suggested that some specific SRL strategies may have more positive impacts to the online program outcome (Kizilcec & Halawa, 2015; Broadbent & Poon, 2015; Kizilcec et al. 2017).

In Hong Kong, a number of online programs have been jointly offered by the Gifted Education Section of the Hong Kong Education Bureau (EDB) and the Hong Kong Academy for Gifted Education (HKAGE). These online programs provide opportunities for primary and secondary school high-ability students or gifted students of HKAGE to self-learn at home. In fact, numerous studies have reported online programs may be a good option for gifted students to learn outside of school (Thomson, 2010; Ng & Nicholas, 2007). The present study was designed to examine the gifted students' uses of self-regulation strategies in these online programs and to those SRL strategies that have the prominent associations with their online learning performance. Knowing what specific strategies are preferred or used most often by those students with better performance will be valuable when considering various provisions of support in the online programs for these students.

Participants of the Study

The targeted participants of this study were students enrolled in the five online learning programs offered for gifted or high-ability students by the Gifted Education Section of the Hong Kong Education Bureau (EDB) and the Hong Kong Academy for Gifted Education (HKAGE). All the students participating in the programs aged 10 to 18 and were nominated by their schools. These online learning programs covered five subjects including Earth Science, Palaeontology, Astronomy, Mathematics and the Changing Hong Kong Economy. Each of these online programs comprised three levels of study, while the highest level being up to senior secondary. All these programs were followed a self-paced format and students could complete all the three levels at a pace being commensurate with their own abilities. However, students must complete their learning and obtain the passing mark in the End-of-Level Test of each level in order to enter the next level. Given the three-level design of these online programs, the student performance at the end of the programs would be classified into the following three categories in an increasing order of achievement: (1) Elementary Level – incomplete (i.e., either failed the End-of-Level Test or dropped out before attempting the End-of-Level

Test); (2) Elementary Level - completed; and (3) Advanced Level – completed.

The number of student participants of this study totally amounted to 157 (84 males and 73 females), who aged from 10 to 16 (*M*=13.02, *SD*=1.99). Out of these participants, more than half of the students (63%) could attain Advanced Level in the online learning programs, while 12% of them could attain Elementary Level. The rest of them (25%) were incomplete in Elementary Level.

Measurement Instruments

In this study, we measure students' SRL strategies based on the Self-Regulated Learning Interview Schedule (SLRIS) developed by Zimmerman and Martinez-Pons (1986). In the SLRIS, 14 SRL strategies were identified and grouped into three categories: motivation, metacognitive and behavioral. In accordance with SLRIS, 14 question items of these SRL strategies were developed. The statements of these question items were scrutinized to ensure that all items were in simple language and readily comprehended by primary and secondary school students (see Table 2). Respondents were asked to rate their frequencies of using a particular SRL strategy using a 5-point scale, ranging from 0 (*never*) to 4 (*almost always*). The higher rating indicates the higher use of the specific strategy.

Data Collection Procedures

Participants of this study were recruited from Hong Kong Academy for Gifted Education (HKAGE). Students who enrolled in any one of the online programs offered by the Gifted Education Section of the Hong Kong Education Bureau (EDB) and HKAGE between 2016/17 and 2018/19 were

Table 2. Fourteen Question Items of Self-regulated Learning Strategies

Strategy		Statement
1. Self-evaluation		When I study, I check if I understand what I have learnt.
2. Organizing and transforming		When I study, I outline the important points to help me organize my thoughts.
3. Goal-setting and planning		When I study, I set goals and organize my study time to accomplish my goals.
4. Seeking information		When I study and I don't understand something, I look for additional information to clarify this (internet/books).
5. Keeping records and monitoring		When I study, I take notes and try to figure out what I need to learn.
6. Environmental structuring		When I study, I choose a place and time to avoid distractions.
7. Self-consequences		When I study, I promise myself I can do something I want later if I get my studying done.
8. Rehearsing and memorizing		When I study (or prepare for a test), I read the material over and over until I remember.
Seeking social assistance	9. Peers	When I study and I don't understand something, I ask peers to help.
	10. Teachers	When I study and I don't understand something, I ask teachers to help.
	11. Adults	When I study and I don't understand something, I ask adults (e.g., parents) to help.
Reviewing records	12. Test	When I study (or prepare for a test), I review the previous tests that I took before.
	13. Texts	When I study (or prepare for a test), I review the material of the programme.
	14. Notes	When I study (or prepare for a test), I review my notes of the programme.

invited to complete the questionnaire survey, including the measures of SRL strategies, some basic demographic information, and previous experiences in online programs. The survey was conducted in an online mode, and student respondents were fully informed about the study purposes and the participation in the study was entirely voluntary. They were asked to indicate how often they used various SRL strategies when completing the online programs. The retrospective self-reported use of SRL strategies from each individual respondents were matched with their highest level achieved at the end of the programs. Finally, a total of 157 students have completed the questionnaire.

Analysis and Findings

In the following, the steps of data analysis under RST approach were applied to the data collected, and the corresponding results and findings are presented and discussed.

Attributes Concerned. In the current study, conditional attributes and decision attributes were discrete values. The decision attribute D and conditional attributes R of the study were defined in Table 3. 70% of the original data (110 data instances) were randomly selected as the training data set to establish the decision rules, while the rest of them (47 data instances) were used as the testing set to verify the classification accuracy.

Dependency between Attributes and Minimal Set of Attributes. The degree of dependency of the decision attribute *D* on the set of conditional attributes *R*, $\gamma_R(D)$, which is simply the proportion of data instances that can be certainly classified with respect to their decision attributes using their conditional attributes *R*. The collection of all these data instances is called a positive region. In *RoughSets* R package, the function *BC.positive.reg.RST* is provided. This function implements a fundamental part of RST and computes a positive region and the degree of dependency. It can be used as a basic building block for development of other RST-based methods. For the current study, $\gamma_R(D)$ was found to be 0.97, which was very high in value.

Next, we aimed to reduce the number of conditional attributes when maintaining "more or less" the same degree of dependency. One of the methods provided in *RoughSets* R package is *FS.greedy.heuristic.reduct.RST*. This function implements a greedy heuristic algorithm for computing an ε-approximate reduct. In the implementation, some attribute subset quality measure can be passed to the algorithm as the parameter called *qualityF*. The measure guides the computations in the search for an ε-approximate reduct. The current study adopted entropy, (which in general, provides a measure of the average amount of information needed to represent an event drawn from a probability distribution for a random variable) as the information gained or lost when adding or deleting an attribute with ε being set to 0.1. The following set *R'* with only five attributes were resulted:

R'= {*ask.teachers, ask.peers, rev.test, set.goal, check*}

The dependency of *D* on *R'*, $\gamma_{R'}(D)$ was found to be 0.88, which was still high after dropping the rest of the nine attributes. The significance of each attribute in *R'* was measured based on the change in dependency value after dropping the attribute concerned. The values of significance of these five attributes sorted in the decreasing order are shown in the Table 4.

Rule Generation. Finally, the rule induction algorithm, LEM2 was deployed to generate rules for each class of decision attribute. The function *RI.LEM2Rules.RST*, being an implementation of LEM2 for induction of decision rules was provided in the *RoughSets* R package. A total of 51 rules were resulted: 16 rules for decision class = 1 (i.e., Elementary Level - Incomplete), 10 rules for decision class = 2 (i.e., Elementary Level - completed), and 25 rules for decision class = 3 (i.e., Advanced Level - completed). When applying this set of rules to the testing data set, the percentage of correctness was 62%. It should be noted that the percentage of correctness was up to 91% when applying this set of rules to the training data set. For sake of reference, when randomly guessing the decision class of a student (i.e., level achieved in online learning), the mean of the percentage of correctness was 34% obtained from 1000 random trials. When randomly guessing according to the proportions of decision classes in the training data set, the mean of the percentage of correctness was 49%. Therefore, the use of decision rules did improve the classification accuracy. As an illustration, some rules for decision class = 1 and decision class = 3 are shown in Table 5 for reference.

Table 3. Decision Attribute D and Conditional Attributes R of the Study

Decision attribute, D
1. *level.achieved*: Level achieved in online learning
Conditional attributes, R
1. *check*: Self-evaluation 2. *outline*: Organizing and transforming 3. *set.goal*: Goal-setting and planning 4. *seek.Info*: Seeking information 5. *take.notes*: Keeping records and monitoring 6. *place.time*: Environmental structuring 7. *reward*: Self-consequences 8. *repeat:* Rehearsing and memorizing *9.* *ask.peers:* Seeking social assistance from peers 10. *ask.teachers:* Seeking social assistance from teachers 11. *ask.adults:* Seeking social assistance from adults 12. *rev.test:* Reviewing records – Previous tests 13. *rev.prog:* Reviewing records – Programme materials 14. *rev.notes:* Reviewing records – Programme notes

Table 4. Values of Significance of Five Attributes

Attribute	Significance
ask.teachers	0.200
ask.peers	0.182
rev.test	0.173
set.goal	0.155
check	0.127

Table 5. Rules for Decision Class = 1 and Decision Class = 3

Decision class =1 (Elementary Level – incomplete)	

Rule 1:	IF *rev.test* is 2 and *set.goal* is 3 and *ask.peers* is 3
	THEN Outcome is 1

Rule 2:	IF *check* is 3 and *ask.teachers* is 1 and *rev.test* is 2
	THEN Outcome is 1

Decision class = 3 (Advanced Level – completed)	

Rule 3:	IF *ask.peers* is 4 and *ask.teachers* is 4
	THEN Outcome is 3

Rule 4:	IF *check* is 4 and *set.goal* is 4
	THEN Outcome is 3

Two-way Tables: Supplementary Analyses. To better understand the impacts of these five attributes on the online learning performance, the row percentages of the cross-tabulations for each attribute vs *level.achieved* based on the training data set are shown in Table 6. From the table, it is observed that those students, who scored high in the attributes (i.e., score =4) concerned, had a much better chance of attaining a high level of achievement. Therefore, it would be beneficial for students to ask teachers and peers when they do not understand some issues. When studying or preparing tests, the effective way is to review the previous tests. Besides, the students should set goals in their study and organize time to achieve them. In addition, they should keep on checking whether they do understand what have learnt.

Table 6. Row Percentages of the Five Conditional Attributes vs the Decision Attribute

Level achieved in online learning	Elementary level – incomplete	Elementary level – completed	Advanced level – completed
	%	%	%
Attribute: *ask.teachers**			
1	55.0%	10.0%	35.0%
2	30.6%	13.9%	55.6%
3	22.9%	14.3%	62.9%
4	10.0%	10.0%	80.0%
Attribute: *ask.peers**			
1	43.8%	12.5%	43.8%
2	30.6%	13.9%	55.6%
3	25.0%	12.5%	62.5%
4	25.0%	15.0%	60.0%
Attribute: *rev.test**			
1	20.0%	30.0%	50.0%
2	44.1%	11.7%	44.1%
3	28.6%	11.4%	60.0%
4	15.3%	7.7%	76.9%
Attribute: *set.goal**			
1	33.3%	20.1%	45.8%
2	19.4%	13.9%	66.7%
3	41.7%	8.3%	50.0%
4	9.1%	9.1%	81.8%
Attribute: *check*			
1	25.0%	25.0%	50.0%
2	24.3%	13.5%	62.2%
3	41.2%	13.7%	45.1%
4	5.6%	5.6%	88.9%

* The number of students, who responded "0" (*never*) to the related questions, are excluded in the tabulation, as it was too few and thus unreliable.

Discussions and Conclusions

Rough set theory (RST) was proposed by Zdzisław Pawlak (Pawlak,1982) as a methodology for data analysis. It revolves around the notion of the approximation of concepts in information systems and the fundamental concepts of discernibility, which is the ability to distinguish between objects, based on their attribute values. The main advantage of using RST approach in data analysis is that it does not need any preliminary or additional information about data – like probability in statistics.

As an illustration of how to use RST for data analysis, we applied RST to a study, which aimed to determine the relative significance of various SRL strategies used by student participants using the utilities provided in *RoughSets* R package. In the present study, the number of participants amounted to less than two hundred. Furthermore, most the responses collected are ordinal data in nature. In this regard, sophisticated statistical modelling, e.g., linear regression and Structural Equation Modeling (SEM), which relies heavily on the continuous nature of the data collected and the normality assumptions of the data distribution, may not be appropriate. Thus, data analysis under RST approach is adopted, which require no additional assumptions at all, to detect the extent of attribute dependency and relative significances amongst the various attributes. The present study showed that five SRL strategies used by students had prominent associations with their achievement in online programs, namely: “seeking teacher assistance”, “seeking peer assistance”, “reviewing tests”, “goal setting and planning”, and “self-evaluation” was found to be the most significant. From the study, it is illustrated that RST is a promising approach for data analysis in the educational research, especially under some specific circumstances. It is expected that more researchers in the field of education will be encouraged to try RST methods in their own studies.

References

Abbas, Z., & Burney, A. (2016). A survey of software packages used for rough set analysis. Journal of Computer and Communications, 4(9), 10-18. https://www.scirp.org/journal/paperinformation.aspx?paperid=68752

Azevedo, R., & Cromley, J. G. (2004). Does Training on Self-Regulated Learning Facilitate Students' Learning With Hypermedia? *Journal of Educational Psychology, 96*(3), 523–535. https://doi.org/10.1037/0022-0663.96.3.523

Broadbent, J., & Poon, W. L. (2015). Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review. *The Internet and Higher Education*, *27*, 1-13. https://doi.org/10.1016/j.iheduc.2015.04.007

Dhandayudam, P., & Krishnamurthi, I. (2013). Customer behavior analysis using rough set approach. *Journal of theoretical and applied electronic commerce research*, *8*(2), 21-33. http://dx.doi.org/10.4067/S0718-18762013000200003

Dignath, C., Büttner, G., & Langfeldt, H. (2008). How can primary school pupils learn SRL strategies most effectively? A meta-analysis on self-regulation training programmes. *Educational Research Review*, *3*(2), 101-129. http://dx.doi.org/10.1016/j.edurev.2008.02.003

Hattie, J., Biggs, J., & Purdie, N. (1996). Effects of learning skills interventions on student learning: A meta-analysis. *Review of Educational Research, 66*(2), 99–136. https://doi.org/10.2307/1170605

Karami, J., Alimohammadi, A., & Seifouri, T. (2014). Water quality analysis using a variable consistency dominance-based rough set approach. *Computers, environment and urban systems*, *43*, 25-33. https://doi.org/10.1016/j.compenvurbsys.2013.09.005

Kizilcec, R. F., & Halawa, S. (2015). Attrition and achievement gaps in online learning. In *Proceedings of the second (2015) ACM conference on learning@ scale* (pp. 57-66). http://dx.doi.org/10.1145/2724660.2724680

Kizilcec, R. F., Pérez-Sanagustín, M., & Maldonado, J. J. (2017). Self-regulated learning strategies predict learner behavior and goal attainment in Massive Open Online Courses. *Computers & education*, *104*, 18-33. https://doi.org/10.1016/j.compedu.2016.10.001

Li, C., Yang, Y., Jia, M., Zhang, Y., Yu, X., & Wang, C. (2014). Phylogenetic analysis of DNA sequences based on k-word and rough set theory. *Physica A: Statistical Mechanics and its Applications*, *398*, 162-171. http://dx.doi.org/10.1016/j.physa.2013.12.025

Lin, T. Y., & Cercone, N. (Eds.). (1997). *Rough sets and data mining: Analysis of imprecise data*. Kluwer Academic Publishers, Dordrecht/Boston/London.

Liu, Y. L. (2014). Research on information technology with character pattern recognition method based on rough set theory. In *Advanced Materials Research* (Vol. 886, pp. 519-523). Trans Tech Publications Ltd. https://doi.org/10.4028/www.scientific.net/AMR.886.519

Ma, S., Liao, H., & Yuan, Y. (2013). Intrusion detection based on rough-set attribute reduction. In *Proceedings of the International Conference on Information Engineering and Applications (IEA) 2012* (pp. 363-369). Springer, London. https://doi.org/10.1007/978-1-4471-4853-1_47

Ng, W., & Nicholas, H. (2007). Conceptualizing the use of online technologies for gifted secondary students. *Roeper Review*, *29*(3), 190-196. https://doi.org/10.1080/02783190709554408

Orlowska, E. (Ed.). (1998). *Incomplete information: Rough set analysis*. Physica, Heidelberg.

Pawlak, Z. (1982). Rough sets. *International journal of computer & information sciences*, *11*(5), 341-356. https://doi.org/10.1007/BF01001956

Pawlak, Z. (1991). *Rough Sets: Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers, Dordrecht/Boston/London.

Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In *Handbook of self-regulation* (pp. 451-502). Academic Press.

Pintrich, P. R., Smith, D. A. F., Garcia, T., & Mckeachie, W. J. (1993). Reliability and Predictive Validity of the Motivated Strategies for Learning Questionnaire (Mslq). *Educational and Psychological Measurement*, *53*(3), 801–813. https://doi.org/10.1177/0013164493053003024

Riza, L. S., Janusz, A., Bergmeir, C., Cornelis, C., Herrera, F., Śle, D., & Benítez, J. M. (2014). Implementing algorithms of rough set theory and fuzzy rough set theory in the R package “RoughSets”. *Information sciences*, *287*, 68-89. https://doi.org/10.1016/j.ins.2014.07.029

Slowinski, R., ed. (1992). *Intelligent Decision Support. Handbook of Advances and Applications of the Rough Set Theory.* Kluwer Academic Publishers, Dordrecht/Boston/London.

Thomson, D. L. (2010). Beyond the Classroom Walls: Teachers’ and Students’ Perspectives on How Online Learning Can Meet the Needs of Gifted Students. *Journal of Advanced Academics*, *21*(4), 662–712. https://doi.org/10.1177/1932202X1002100405

Winne, P. H., Nesbit, J. C., Kumar, V., & Hadwin, A. F., Lajoie, S. P., Azevedo, R. A., & Perry, N. E. (2006). Supporting self-regulated learning with gStudy software: The Learning Kit Project. *Technology, Instruction, Cognition and Learning, 3*(1), 105-113.

Zimmerman, B. J., & Martinez-Pons, M. (1986). Development of a Structured Interview for Assessing Student Use of Self-Regulated Learning Strategies. *American Educational Research Journal*, *23*(4), 614–628. https://doi.org/10.3102/00028312023004614

Zimmerman, B. J., & Martinez-Pons, M. (1988). Construct validation of a strategy model of student self-regulated learning. *Journal of Educational Psychology, 80*(3), 284–290. https://doi.org/10.1037/0022-0663.80.3.284

Citation:
Fung, T., & Li, W. (2022). Rough set theory (RST) data analysis with R and its application on studying relative significances of self-regulated learning (SRL) strategies of gifted students. *Practical Assessment, Research, & Evaluation*, 27(25). Available online: https://scholarworks.umass.edu/pare/vol27/iss1/25/

Corresponding Author:
Tze-ho Fung
Hong Kong Academy for Gifted Education
Sha Kok Estate, Shatin, N.T., Hong Kong

Email: ericfung [at] hkage.org.hk