The imperfection of *rough sets* in the education field

Jie Chen†, Shi-Jing Wu† & Kun-Li Wen‡

Wuhan University, Wuhan, People's Republic of China†
Chienkuo Technology University, Changhua, Taiwan‡

ABSTRACT: Based on the characteristics of rough sets, they are seen to overlap with many other theories, especially with fuzzy set theory, evidence theory and Boolean reasoning methods. The rough set methodology has found many real-life applications, such as medical data analysis, finance, banking, engineering, voice recognition, image processing and others. But until now, there has been little research associated with imperfections of rough sets. Hence, the main purpose of this article is to study the imperfections of rough sets in the field of education. First of all, the mathematical model of rough sets is previewed, and an example is given to verify the approach in this article. It concerns the weighting of evaluation factors in English learning in the education field. After the examples are presented, a complete human-machine interface toolbox is described to support the complex calculation and verification of the complex data The toolbax uses Matlab/GUI. Finally, some suggestions are made for future research.

INTRODUCTION

Dr Pawlak presented *rough set theory* in 1982. The basic topics of rough sets include: set theory, conditional probability, membership function, attributes analysis and uncertainty description of knowledge. The main purpose of rough sets is to examine the difference between lower and upper approximations, subjective results for clustered sets, and to find the weighting factors for systems [1][2].

Therefore, the main function of rough set theory is classification by analysis of characteristics in rough set theory. To use rough set theory to find the weighting of an influence factor in a system, two conditions must be met. First, the data must be discrete and the second is that the attribute factor and decision factor satisfy the indiscernibility or discernibility condition (see Table 1) [3].

Table 1: The characteristics of rough set.

Attribute	Decision factor	Results
1. Discernibility	1. Discernibility	Cannot use
2. Discernibility	2. Indiscernibility (Likert's five points scale)	Cannot use
3. Indiscernibility	3. Discernibility	Can use
3. Indiscernibility	4. Indiscernibility (Likert's five points scale)	Cannot use

A review of the past research about this field shows that, although there have been many studies in this field, there is only one article that has discussed the limitations of rough sets [3]. Therefore, the focus of this study is the limitations of rough sets, to provide a new approach to weighting analysis. In this article, the next section has an introduction to the basic mathematical foundation of rough sets. An example is provided in the following section to verify the approach, which concerns the evaluation of English learning in education [4]. Also, the Matlab GUI toolbox is developed to verify the approach [5-7]. Finally, some advantages and suggestions are provided for the further research.

THE MATHEMATICAL MODEL AND A PREVIEW OF ROUGH SETS

In this section, the basic concepts of rough sets are introduced [1]:

- Information system: $IS = (U, A)$ is called an information system, where $U = \{x_1, x_2, x_3, \ldots\ldots x_n\}$ is the finite set universe of objects, and $A = \{a_1, a_2, a_3, \ldots, a_m\}$ is the set of attributes;

- Information function: there exists a mapping $f_a : U \to V_a$, then V_a is the set of values of a called the domain of attribute a;
- Discrete: the equal interval width is shown in Equation (1):

$$t = \frac{V_{\max.} - V_{\min.}}{k} \tag{1}$$

where: $V_{\max.}$: Maximum value of the data. $V_{\min.}$: Minimum value of the data and the range of the attribute is $[V_{\max.}, V_{\min}]$.

The intervals corresponding to the attribute values are:

$$\{[d_0, d_1], [d_1, d_2], \cdots, [d_{k-1}, d_k]\} \tag{2}$$

where: $d_0 = V_{\min}$, $d_k = V_{\max}$, $d_{i-1} < d_i$, $i = 1,2,3,\cdots,k$, k is the level of discreteness.

- Lower approximations and upper approximations:

  If $A \subseteq U$, then the lower approximation is defined as:

$$\underline{R}(A) = \{x \in U \mid [x]_R \subseteq A\} = \bigcup \{[x]_R \in \frac{U}{R} \mid [x]_R \subseteq A\},\ [x]_R = \{y \mid xRy\} \tag{3}$$

and the upper approximation is defined as:

$$\overline{R}(A) = \{x \in U \mid [x]_R \cap A \neq \phi\} = \bigcup \{[x]_R \in \frac{U}{R} \mid [x]_R \cap A \neq \phi\},\ [x]_R = \{y \mid xRy\} \tag{4}$$

In other words, the lower approximation of a set is the set of all elements that definitely belong to *U*, whereas the upper approximation of *U* is the set of all elements that possibly belong to *U*.

- Indiscernibility: An indiscernibility relation is defined for any x_i and x_j, if x_i is identical to x_j, then x_i and x_j have the same properties;
- Positive, negative and boundary: the positive, negative and boundary are defined as:

$$pos_R(X) = \underline{R}(X),\ neg_R(X) = U - \overline{R}(X),\ bn_R(A) = \underline{R}(A) - \overline{R}(A) \tag{5}$$

- The dependents of attributes: The dependents of attributes are defined as:

$$\gamma_c(D) = \frac{|posc(D)|}{U} \tag{6}$$

if $a \in C$, the ratio in the whole set.

- The significant value of attributes: In a decision system, $S = (U, C \cup D, V, f)$, if $a \in C$, the significant values of attributes are defined as:

$$\sigma_{(C,D)}(a) = \frac{\gamma_c(D) - \gamma_{c-\{a\}}(D)}{\gamma_c(D)} \tag{7}$$

This means significant values of attributes can be considered as the weighting of each factor.

A STUDY OF ENGLISH VOCABULARY LEARNING STRATEGIES FOR TAIWAN COLLEGE STUDENTS

Recently, English has played a key role in the dissemination of ideas and thoughts throughout the world. A main national objective has been to foster high quality international participation in developing English capabilities. In the *Education Policy 2005-2008* the Ministry of Education required all levels of school to develop a General English Proficiency Test (GEPT) to achieve a level of proficiency in English. In the highly competitive modern society of Taiwan, English language proficiency is a powerful asset in seeking employment and securing promotion. Accordingly, a plethora of organisations offer proficiency tests, the most common being TOEIC, IELTS, TOEFL, GET (Global English Test), Cambridge Main Suite and FLPT (Foreign Language Proficiency Test). More and more colleges require

students to pass GEPT or a similar English proficiency test before graduation. These requirements place heavy psychological pressure on non-English major students. In addition, many colleges even require their professors to select textbooks in English and instruct in English. Therefore, listening and reading abilities become more and more important. A strategy for learning English vocabulary can improve reading ability [4].

In modern society, people are often judged not only by their appearance, but also by their ability to speak, whether students or teachers, politicians or salesmen. According to the findings of domestic researchers, college students in Taiwan have serious problems learning vocabulary. Educators and language test organisations expect senior high-school graduates to have a vocabulary of 5,000 to 7,000 words to comprehend college English textbooks. But according to past research, 171 college students did a *Vocabulary Level Test*, which was based on Laufer [9] and the Nation's vocabulary level test. The result of the study was that 48.5% of the students could understand up to 1,000 words and 17% of students up to 2,000 words. Only 2.3% of the students had an understanding of 3,000 words or more. This means that up to 32.5% of the students had a vocabulary of less than 1,000 words.

Most of the time, students cannot comprehend the meaning of new words. Researchers believed that an insufficient vocabulary will lead to poor reading comprehension and poor subsequent academic achievement. Some foreign scholars suggest that if language learners have a larger vocabulary, they can understand more content and express themselves more clearly. They are also able to read broadly and deeply on subjects. Vocabulary ability is a very important indicator of reading comprehension.

In addition, good use of vocabulary learning strategies is an effective way to achieving reading comprehension. Hence, in this example, the score is based on a Likert's five points scale (Table 2). A total of 15 students were selected from each English proficiency group, and assessed on 24 factors (Table 3).

Table 2: Likert's five points scale.

Score	1	2	3	4	5
Items	All the time	Most of the time	About half the time	Some of the time	Not at all

Table 3: Contents of the questionnaire.

No	Content
Q_1	I will analyse parts (verbs, nouns) of speech to judge the meaning.
Q_2	I will guess the meaning from textual context.
Q_3	I will consult a bilingual dictionary.
Q_4	I will consult a monolingual dictionary.
Q_5	I will consult a Chinese-English and English-Chinese dictionary.
Q_6	I will ask teachers for a sentence including the new word.
Q_7	I will discover a new word's meaning through group work activities.
Q_8	I will study and practice a new word's meaning with classmates.
Q_9	I will interact with native speakers with new vocabularies.
Q_{10}	I will connect the word to its synonyms and antonyms.
Q_{11}	I will use new words in sentences.
Q_{12}	I will group words together within a storyline.
Q_{13}	I will say a new word aloud when studying.
Q_{14}	I will underline the new word to enhance the impression.
Q_{15}	I will remember root, prefix and suffix of the word.
Q_{16}	I will learn the whole phrase including the new word.
Q_{17}	I will take notes in class.
Q_{18}	I will use the vocabulary section in the textbook to learn new words.
Q_{19}	I will listen to tapes of word lists.
Q_{20}	I will keep a vocabulary notebook to write down new words.
Q_{21}	I will learn new words from watching English films.
Q_{22}	I will learn new words from reading English newspapers.
Q_{23}	I will learn new words from reading English articles.
Q_{24}	I will learn new words from listening to English radio programmes.

Test results for the 15 students are listed in Table 4, and the calculation steps are shown below.

Discrete, Indiscernibility and Discernibility

- The result of the questionnaires in Table 4 is discrete and has satisfied the assumptions for a rough set.
- Then, the indiscernibility method was used to verify the results of the questionnaires, and the attribute factors all satisfy the discernibility condition, i.e. the same as the first and second items in Table 1.

- Also, the decision factors in the questionnaire result satisfy the indiscernibility condition, i.e. the same as the second and fourth items in Table 1. Hence, we cannot obtain the weighting of those factors.

Dependence and Significance

- For the condition attributes

$$\frac{U}{C}=\frac{U}{\{Q_1,Q_2,Q_3,\cdots,Q_{24}\}}=\{\{x_1\},\{x_2\},\{x_3\},\{x_4\},\{x_5\},\{x_6\},\{x_7\},\{x_8\},\{x_9\},\{x_{10}\},\{x_{11}\},\{x_{12}\},\{x_{13}\},\{x_{14}\},\{x_{15}\}\}.$$

- For the decision attributes: $\frac{U}{D}=\{x_1\ x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}\}=\{X_1\}$.

Hence, $pos_C(D)=\{x_1\ x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}\}$, substitute this into Equation (6), to obtain:

$$\gamma_c(D)=\frac{|pos_C(D)|}{|U|}=\frac{15}{15}=1.$$

- Omit the attributes of Q_1

  The condition attributes:

  $$\frac{U}{C}=\frac{U}{\{Q_2,Q_3,Q_4,\cdots,Q_{24}\}}=\{x_1\ x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}\}.$$

  For decision attributes: $\frac{U}{D}=\{x_1\ x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}\}=\{X_1\}$.

  Hence, $pos_C(D)=\{x_1\ x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}\}$, and substituting into Equation (6),

  gives: $\gamma_{c-\{Q_1\}}(D)=\frac{|pos_C(D)|}{|U|}=\frac{15}{15}=1$, and from Equation (7), the significant of Q_1 is $\sigma_{(C,D)}(Q_1)=\frac{1-1}{1}=0$.

- Omit the attributes of Q_2

  $$\frac{U}{C}=\frac{U}{\{Q_1,Q_3,Q_4,\cdots,Q_{24}\}}=\{x_1\ x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}\}.$$

  For the decision attributes: $\frac{U}{D}=\{x_1\ x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}\}=\{X_1\}$.

  Hence, $pos_C(D)=\{x_1\ x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}\}$, substituting into Equation (6),

  gives: $\gamma_{c-\{Q_2\}}(D)=\frac{|pos_C(D)|}{|U|}=\frac{15}{15}=1$, and from Equation (7), the significant of Q_2 is $\sigma_{(C,D)}(Q_2)=\frac{1-1}{1}=0$.

- Omit the attributes of

  $$\frac{U}{C}=\frac{U}{\{Q_1,Q_2,Q_4,\cdots,Q_{24}\}}=\{x_1\ x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}\}.$$

  For the decision attributes: $\frac{U}{D}=\{x_1\ x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}\}=\{X_1\}$.

  Hence, $pos_C(D)=\{x_1\ x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}\}$, substituting into Equation (6),

gives: $\gamma_{c-\{Q_3\}}(D)=\frac{|pos_C(D)|}{|U|}=\frac{15}{15}=1$, and from quation (7), the significant of Q_3 is $\sigma_{(C,D)}(Q_3)=\frac{1-1}{1}=0$.

- Also omitting, Q_4, Q_5, Q_6 $\cdots$, Q_{24} respectively, then the significant values are all 0.

Table 5: The Final result of the weighting for each factor.

Factor	Q_1	Q_2	Q_3	…….	…….	Q_{22}	Q_{23}	Q_{24}
Significant	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

DESIGN OF THE TOOLBOX

A toolbox based on Matlab was developed to reduce the huge and complex calculation. This allowed data to be put in easily for the calculation. The toolbox makes the results on the analysis more convincing and practical [3].

Characteristics of the Toolbox: The Toolbox processes rough set formulae and methods and uses a GUI interface. The input interface is Matlab's GUI. Set numbers can be put in randomly and the system is easy to use. The system requirements for the toolbox are: Window XP; Screen resolution at least 1024×768; Matlab 2007; Excel.

Calculations using the Matlab toolbox are shown in Figure 1 and Figure 2.

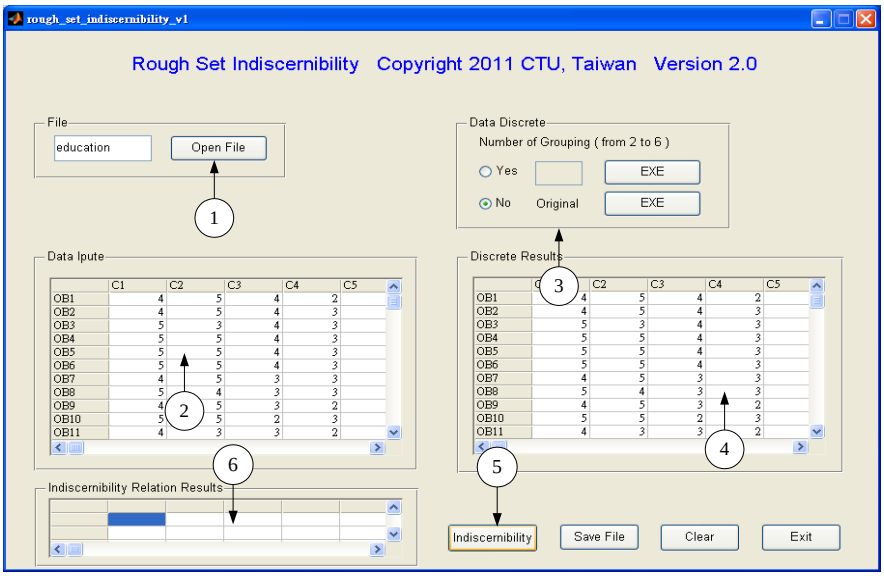

Figure 1: The indiscernibility of English learning (not exist).

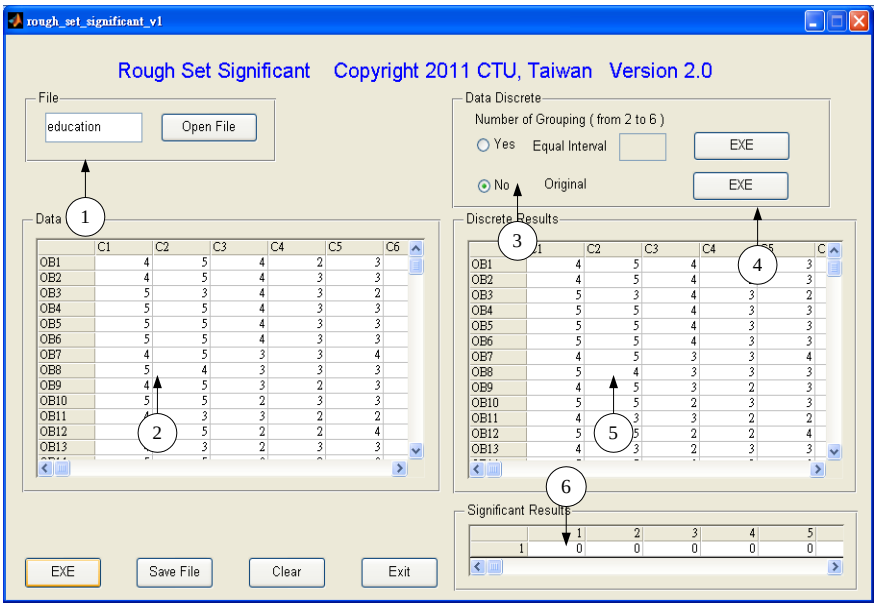

Figure 2: The significance of English learning.

Table 4: Test results from Question 1 to Question 24 and output.

No.	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_9	Q_{10}	Q_{11}	Q_{12}	Q_{13}
x_1	4	5	4	2	3	3	3	1	1	5	5	1	5
x_2	4	5	4	3	3	3	2	3	1	5	4	2	5
x_3	5	3	4	3	2	3	3	2	2	5	4	2	4
x_4	5	5	4	3	3	3	2	1	2	4	3	1	5
x_5	5	5	4	3	3	4	1	1	1	5	4	2	5
x_6	5	5	4	3	3	2	1	2	1	5	4	2	5
x_7	4	5	3	3	4	4	1	1	2	5	4	2	5
x_8	5	4	3	3	3	3	1	1	1	5	3	1	5
x_9	4	5	3	2	3	3	1	1	1	5	4	1	5
x_{10}	5	5	2	3	3	3	1	1	1	4	4	1	5
x_{11}	4	3	3	2	2	3	1	2	1	5	3	2	5
x_{12}	5	5	2	2	4	3	1	1	2	5	4	1	5
x_{13}	4	3	2	3	3	3	1	1	2	5	5	2	5
x_{14}	5	5	3	2	3	2	1	2	1	5	4	1	5
x_{15}	5	5	4	1	3	3	1	1	2	5	3	2	5

No.	Q_{14}	Q_{15}	Q_{16}	Q_{17}	Q_{18}	Q_{19}	Q_{20}	Q_{21}	Q_{22}	Q_{23}	Q_{24}	Decision
x_1	4	5	4	2	3	3	3	1	1	5	5	5
x_2	4	2	4	3	3	3	2	3	1	3	4	5
x_3	5	5	4	3	2	3	3	2	2	5	4	5
x_4	5	3	4	3	3	3	2	1	2	4	3	5
x_5	5	5	4	3	3	4	1	1	1	5	4	5
x_6	5	5	4	3	3	2	1	2	1	5	4	5
x_7	4	3	3	3	4	4	1	1	2	5	4	5
x_8	5	5	3	3	3	3	1	1	1	5	3	5
x_9	4	5	3	2	3	3	1	1	1	4	4	5
x_{10}	5	4	2	3	3	3	1	1	1	5	4	5
x_{11}	4	5	3	2	2	3	1	2	1	5	3	5
x_{12}	5	5	2	2	4	3	1	1	2	5	4	5
x_{13}	4	2	2	3	3	3	1	1	2	4	5	5
x_{14}	5	2	3	2	3	2	1	2	1	5	4	5
x_{15}	5	5	4	1	3	3	1	1	2	5	3	5

CONCLUSIONS

The main function of rough sets in the previous researches is for classification. Therefore, the analysed data have to be discrete and then the information table can be generated. Next, the attribute factors and decision factors in the information table will be judged if they are discernibility or not. Due to the process of discrete method, different discrete numbers will generate different information tables; however, if the attribute factors are discernibility while the decision factors are indiscernibility in each analysed item of the information table, it is impossible to classify the analysed object and calculate the weighting. Hence, in this article, a questionnaire analysis method in the education field is presented, to verify the approach described in this article. Under this circumstance, the values of lower approximations and upper approximations are equal, based on Equation (5). This means the boundary does not exist. Therefore, the value of dependence must be equal to 0, which implies that the significance is also equal to 0.

To sum up, rough set theory is a new classification method in soft computing. Some practical problems in relation to the application of rough sets have been investigated in many domains. In this article, the authors have provided an example with which to verify the imperfection of indiscernibility or discernibility in rough set theory. This was the main contribution of this article. As well, Matlab was used to develop a toolbox to help do the complex calculation with large amounts of data.

REFERENCES

1. Pawlak, Z., Rough sets approach to multi-attribute decision analysis, *European J. of Operational Research*, 72, 443-459 (1994).
2. Wen, K.L., Nagai, M.T., Chang, T.C. and Wen, H.C., *An Introduction to Rough Set and Application*. Taipei: Wunan Published (2008).
3. Sheu, T.W., Liang, J.C., You, M.L. and Wen, K.L., The study of imperfection in rough set on the field of engineering and education. *Proc. 3rd 2010 International Conference on Advanced Software Engineering & Its Applications*, Jeju Island, Korea, 93-102 (2010).
4. Liang, H.Y., A study of English vocabulary learning strategies in Taiwan college students via GM (0,N). Final Project Report of Chienkuo Technology University (2010).

5. Roger, J.S., *MATLAB Program Design*. Taiwan: Terasoft Company (2004).
6. Wen, K.L. and You, M.L., The development of rough set toolbox via Matlab. *Current Development in Theory and Applications of Computer Science*, *Engng and Technol.*, 1, 1-15 (2010).
7. Wen, K.L., You, M.L. and Wang. J.R., The development of Matlab toolbox for kansei factor analysis. *Inter. J. of Kensei Information*, 1, **1**, 43-52 (2010).
8. Taiwan Ministry of Education, *Education Main Policy 2005-2008* (2008), http://torfl.pccu. edu.tw/torfl8_2.htm
9. Laufer, B., *How Much Lexis is Necessary for Reading Comprehension?* In: Be Joint, H. and Arnaud, P. (Eds), Vocabulary and Applied Linguistics. London: MacMillan (1992).