Journal of Mathematics and Informatics
Vol. 20, 2021, 19-28
ISSN: 2349-0632 (P), 2349-0640 (online)
Published 11 March 2021
www.researchmathsci.org
DOI: http://dx.doi.org/10.22457/jmi.v20a03191

An Analysis of the Middle School Students' Mathematical Academic Achievements with the Attribute Model of Rough Sets

Yateng Yue* and *Zengtai Gong*

College of Mathematics and Statistics, Northwest Normal University, China
*Corresponding author. Email: zt-gong@163.com

Received 15 February 2021; accepted 9 March 2021

Abstract. In this paper, a rough set attribute model is established based on rough set theory. Through data discretization, attribute reduction, and calculate the degree of dependency, the importance of each question type in the test to the score is obtained. Finally, an experimental analysis is given.

Keywords: Rough set, Attribute reduction, Dependence degree

AMS Mathematics Subject Classification (2010): 28E10, 04A72

1. Introduction

It well known that the mathematics course is the most fundamental subject for the students whether they are in a college, a high school or an elementary school. An eternal theme of the mathematical education, especially for middle school students is that how to improve the mathematical academic achievement.

Xiufen [1] put forward in "An empirical research on improving the academic achievements of junior middle school students with mathematics learning difficulties in rural areas", to help students establish correct mathematics achievement goals. In the process of students learning, the author asks students to take the mastery goal as the process goal of learning, and the achievement goal as the end goal of learning. Yueyuan [2] et al. proposed in "The Influence Path of High School Students' Mathematics Achievement under High-efficiency Mathematics Learning", because mathematics learning performance is mainly affected by two paths, one path is the path composed of intellectual factors and mathematical literacy. Another path is a mathematical meta-cognitive, non-intelligence factors affecting mathematics and mathematics learning strategies constitute the path. Therefore, different paths can be selected according to the different learning characteristics of students to improve mathematics learning performance from many aspects. Haibo [3] and others put forward the application level of eight learning strategies in the article "The relationship between learning interest, self-efficacy, learning strategies and performance-based on Kolb learning style of junior high school mathematics learning". They believed that mathematical learning strategies contributed 17% to math achievement, and there was still a lot of space for further

exploration to play the role of learning strategies in math learning. All of the above positive studies are to observe changes in performance by changing learning methods and learning strategies, so as to find ways to improve academic performance. Commonly used methods for data processing and mining are based on sample questionnaires and simple data statistics [4,5]. This method is relatively simple and cannot dig deeper into the relationship between data. There is also probability statistical analysis [6-8], This requires some additional data or prior knowledge. However, this knowledge is not easy to obtain, the indicator system is too cumbersome, and many data are difficult to collect, so it is difficult to promote and apply.

This article changes the way of thinking. It analyzes the students' academic performance from the reverse direction, and finds the importance of each question type in the math test paper, that is, the impact of each question type on the total score. This helps students find their own learning method, scientifically arrange study energy and exam time, which is conducive to the key teaching of teachers. It has a great effect on students, especially high school students entering the review. In addition, this paper uses the rough set theory to carry out data mining, and it does not need to provide any prior information except the data set to be processed. Therefore, compared with other methods, this theory can fully mine and use the connections between known data to determine the importance, avoiding the subjective defects caused by traditional methods.

This article first establishes an information system by collecting data and discretizing the index data to obtain data that can be processed by rough set methods; then, according to the rough set attribute reduction principle, the attribute reduction is performed, and redundant attributes are removed from the index system; Finally, data mining is carried out through rough set theory to dig out the importance of question types.

2. Preliminaries

Definition 1. *[9] Information system* $S=(U,A,V,f)$. *Among them: the universe* $U=\{x_1,x_2,\cdots x_n\}$ *represents a non-empty and finite set of entities, called universe;* $A=\{a_1,a_2,\cdots a_m\}$ *is a non-empty and finite set of attributes;* $A=C\bigcup D$ *and* $C\bigcap D=\phi$, *where an element of* C *is called a condition attribute,* C *is the condition attribute set, an element of* D *is called a decision attribute, and* D *is the decision attribute set.* $V=\bigcup_{a\in A}V_a$, V_a *is called domain of attribute a;* $f:U\times A\to V$ *is an information function,* $\forall a\in A, x\in U, f(x,a)\in V_a$.

Definition 2. *For a decision table,* $S=(U,C\bigcup D,V,f)$, $B\subseteq C$, *the indistinguishable definition based on* B *is:*

$ind(B)=\{(x,y)\in U\times U\}$ *or* $ind(B)=B\{\forall b\in B, f(x,b)=f(y,b)\}$.

If $(x,y)\in ind(B)$, *then* x *and* y *are said to be indistinguishable based on* B. *The symbol* $U/ind(B)$ *represents all equivalence classes derived from the indistinguishable relationship* $ind(B)$ *on* U, *which can be abbreviated as* U/B. *Usually,* $ind(C)$ *is called the conditional equivalence class, and* $ind(D)$ *is called the decision equivalence*

class. $[X]_B$ *represents the equivalence class of* x *based on* B.

Definition 3. *Let* $S=(U,A,V,f)$ *be a knowledge expression system,* $|U|=n$. *The discernibility matrix of* S *is a* $n\times n$ *matrix, any element of which is*

$$\alpha(x,y)=\{a\in A|f(x,a)\neq f(y,a)\}. \quad (1)$$

Therefore, $\alpha(x,y)$ *is the set of all attributes that distinguish objects* x *and* y.

Definition 4. *For each attribute* $a\in A$, *specify a Boolean variable " a ", if* $\alpha(x,y)=\{a_1,a_2\cdots a_k\}\neq\phi$, *specify a Boolean function* $a_1\vee a_2\vee\cdots\vee a_k$, *and use* $\sum\alpha(x,y)$ *to represent it, if* $\alpha(x,y)=\phi$, *specify a Boolean constant as 1. The (Boolean) distinguishing function* Δ *can be defined as follows:*

$$\Delta=\prod_{(x,y)\in U\times U}\sum\alpha(x,y). \quad (2)$$

All conjunctions in the minimal disjunctive normal form of function Δ *are all reductions of attribute set* A.

Definition 5. *For a given decision table* $S=(U,C\cup D,V,f)$, $P\subseteq C$, *the positive region* $pos_P(Q)$ *of the attribute set* Q *with respect to* P *is defined as follows:*

$$pos_P(Q)=\bigcup_{X\in U/Q}\underline{P}X. \quad (3)$$

Definition 6. [10] *Let* $K=(U,R)$ *be a knowledge base,* $P,Q\subseteq R$ *:*

(1) If and only if $ind(P)\subseteq ind(Q)$, Q *relies on* P, *namely* $P\Rightarrow Q$ *;*

(2) If and only if $P\Rightarrow Q$ *and* $Q\Rightarrow P$, *namely* $ind(P)=ind(Q)$, P *is equivalent to* Q, *namely* $P\equiv Q$ *;*

(3) If and only if $P\Rightarrow Q$ *or* $Q\Rightarrow P$ *are not true,* P *is independent to* Q, *namely* $P\neq Q$ *;*

(4) If and only if

$$k=\gamma_P(Q)=|pos_P(Q)|/|U|, \quad (4)$$

Q *dependent to* P *on the degree of dependence* $k(0\leq k\leq 1)$, *denoted as* $P\Rightarrow_k Q$ *:*

①*If* $k=1$, *it is said* Q *is completely dependent on* P, *namely* $P\Rightarrow_1 Q$,*and is also denoted as* $P\Rightarrow Q$ *;*

② *If* $0<k<1$, *then it is said* Q *is depend on* P *partly;*

③ *If* $k=0$, *then it is said* Q *is completely independent of* P.

Definition 7. [11] *Let* C *and* D *be the conditional attribute set and decision attribute set respectively, and the importance of attribute subset* $C'\subseteq C$ *with respect to* D *is defined as:*

$$\sigma_{CD}(C') = \gamma_C(D) - \gamma_{C-C'}(D). \quad (5)$$

Especially when $C' = \{a\}$, *the importance of attribute* $a \in C$ *with respect to* D *is*

$$\sigma_{CD}(a) = \gamma_C(D) - \gamma_{C-\{a\}}(D). \quad (6)$$

3. Rough set property model application flow

3.1. Establishment of information system

The lowest level of index is found from the index system, which is used to constitute the attribute set of the information system. As the object of evaluation is the object set of the information system, an information system consisting of all objects and each index value can be established.

3.2. Discretize index data

Since the rough set method can only deal with discrete data, each attribute is measured through five levels of "1", "2", "3", "4" and "5", and the measured mathematical academic achievements are replaced by corresponding values from low to high according to the actual situation.

3.3. Information system attribute reduction

Based on the attribute reduction principle of rough set, a new index system is formed by removing redundant attributes from the index system and retaining necessary attributes. Rough set attribute reduction algorithms include: attribute reduction algorithm based on mutual information, inductive attribute reduction algorithm, attribute reduction algorithm based on mutual information, differentiated matrix reduction algorithm, attribute reduction algorithm based on search strategy and data analysis reduction algorithm. The discriminant matrix reduction algorithm is the most effective. In this paper, the discriminant matrix is used to express knowledge, the discriminant function is calculated through the definition of the discriminant function, and the absorptivity is used to simplify the discriminant function, and finally the reduced attribute set is obtained.

3.4. Determine the importance of the indicators

To figure out the importance of some knowledge or attribute, we remove some attributes from the table and see how the classification changes without that attribute. If the strength of the attribute is large, that is, the importance is high, the corresponding classification changes will be great if the attribute is deleted. Inversely, if the strength of the attribute is smaller, i.e. the importance is lower, the corresponding classification change of deleting the attribute will be small.

4. Experimental analysis

4.1. The data collection

The data collected in this paper comes from the first simulation test paper of 20 students in a senior high school science class in Lanzhou city. The grades of the students in this class are relatively representative in the whole school. The grades of the students in this class can be divided into different grades, and the grades of each question type can also be divided into different grades. Due to the possible changes in the type of questions to

be solved in the math test paper, this paper only takes the type of questions in the simulation test as an example, and the score table is as follows:

Table 1: Lanzhou city a senior science class math scores

Number	Choice	Completion	Question one	Question two	Question three	Question four	Question five	Optional question	Total score
1	60	15	12	12	12	10	6	10	137
2	55	20	12	11	12	10	8	9	137
3	60	15	11	11	11	8	8	8	132
4	55	15	12	12	12	9	6	10	131
5	50	15	12	12	12	9	6	10	126
6	50	20	11	11	12	4	6	7	121
7	50	15	10	10	10	9	4	10	118
8	45	15	11	10	8	7	6	8	110
9	45	10	12	10	10	6	6	8	107
10	40	15	9	10	9	4	8	9	104
11	45	10	10	8	9	7	6	7	102
12	40	10	10	8	9	6	4	6	93
13	40	10	11	8	8	6	4	6	93
14	35	10	6	9	6	6	6	10	88
15	35	10	8	8	8	4	4	8	85
16	40	10	6	6	6	4	0	4	76
17	30	10	6	8	4	0	0	4	62
18	30	5	0	6	6	4	2	6	59
19	25	5	6	0	4	8	6	0	54
20	20	5	6	6	4	6	4	2	53

The original data are discretized into five levels, namely, multiple choice questions, fill in the blanks, solution to the first question, solution to the second question, solution to the third question, solution to the fourth question, solution to the fifth question, and optional questions, which are marked as $\{1,2,3,4,5\}$, and are defined as follows:

5: The score is $90\%-100\%$ of the score of this question;
4: The score is $80\%-89\%$ of the score of this question;
3: The score is $70\%-79\%$ of the score of this question;
2: The score is $60\%-89\%$ of the score of this question;
1: The score is $0\%-59\%$ of the score of this question;

When we regard the question type and the total score as attributes, we will get the knowledge base. Discretize the data through the above method, which forms the knowledge base representation of the original data. $C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8$ is the conditional attribute of the knowledge base, representing multiple choice questions and

fill-in-the-blank , Answer the first question, answer the second question, answer the third question, answer the fourth question, answer the fifth question, and choose the question; the total score D is the decision attribute of the knowledge base [12].

After discretizing the original data, the following table is obtained:

Table 2: Knowledge base

Number	C_1 Choice	C_2 Completion	C_3 Question one	C_4 Question two	C_5 Question three	C_6 Question four	C_7 Question five	C_8 Optional question	D Total score
1	5	3	5	5	5	4	1	5	5
2	5	5	5	5	5	4	2	5	5
3	5	3	5	5	5	2	2	4	4
4	5	3	5	5	5	3	1	5	4
5	4	3	5	5	5	3	1	5	4
6	4	5	5	5	5	1	1	3	4
7	4	3	4	4	4	3	1	5	3
8	3	3	5	4	2	1	1	4	3
9	3	1	5	4	4	1	1	4	3
10	2	3	3	4	3	1	1	5	2
11	3	1	4	2	3	1	1	3	2
12	2	1	4	2	3	1	1	2	2
13	2	1	5	2	2	1	1	2	2
14	1	1	1	3	1	1	1	5	1
15	1	1	2	2	2	1	1	4	1
16	2	1	1	1	1	1	1	1	1
17	1	1	1	2	1	1	1	1	1
18	1	1	1	1	1	1	1	2	1
19	1	1	1	1	1	2	1	1	1
20	1	1	1	1	1	1	1	1	1

According to Table 1, establish discrimination matrix table 2, in which the elements

are derived from formula 1, and the discrimination matrix is a symmetric matrix [9], therefore, only half of the elements of the matrix are calculated during the calculation, and the object itself is not distinguished from itself.

Table 3: Discrimination matrix

Evaluation object	1	2	…	18	19
2	C_2, C_7	$---$	…	$---$	$---$
3	C_6, C_7, C_8	C_2, C_6, C_8	…	$---$	$---$
4	C_6	C_2, C_6, C_7	…	$---$	$---$
5	C_1, C_6	C_1, C_2, C_6, C_7	…	$---$	$---$
…	…	…	…	$---$	$---$
17	$C_1, C_2, C_3, C_4, C_5, C_6,$	$C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8$	…	$---$	$---$
18	$C_1, C_2, C_3, C_4, C_5, C_6,$	$C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8$	…	$---$	$---$
19	$C_1, C_2, C_3, C_4, C_5, C_6,$	$C_1, C_2, C_3, C_4, C_5, C_6, C_8$	…	C_6, C_8	$---$
20	$C_1, C_2, C_3, C_4, C_5, C_6,$	$C_1, C_2, C_3, C_4, C_5, C_6, C_8$	…	C_8	C_6

The discrimination matrix is a table. Because the table is too large and the space is limited, it is not listed one by one. The data in the table is calculated by formula (2) to obtain the discrimination function:

$$\begin{aligned}\Delta &= (C_6 \vee C_7)(C_6 \vee C_7 \vee C_8)C_6(C_1 \vee C_6)\cdots \\ &\cdot (C_1 \vee C_2 \vee C_3 \vee C_4 \vee C_5 \vee C_6 \vee C_8) \\ &\cdot (C_2 \vee C_6 \vee C_8)(C_2 \vee C_6 \vee C_7)(C_1 \vee C_2 \vee C_6 \vee C_7)\cdots \\ &\cdots \\ &\cdot (C_6 \vee C_8)C_8 \\ &\cdot C_6 \\ &= C_1C_2C_4C_6C_8\end{aligned}$$

It can be seen from the discrimination function that $\{C_1, C_2, C_4, C_6, C_8\}$ is the only reduction in the indicator system.

4.2. Computational dependencies and metrics

Let $C=\{C_1,C_2,C_4,C_6,C_8\}$ be the new conditional attribute, with

$$U/ind(C-C_1)=\{\{1\},\{2\},\{3\},\{4,5\},\{6\},\{7\},\{8\},\{9\},\{10\},\{11\},\{12,13\},\{14\},\{15\},\{16,20\},\{17\},\{18\},\{19\}\}$$

$$U/ind(C-C_2)=\{\{1,2\},\{3\},\{4\},\{5,7\},\{6\},\{8,9\},\{10\},\{11\},\{12,13\},\{14\},\{15\},\{16\},\{17\},\{18\},\{19\},\{20\}\}$$

$$U/ind(C-C_4)=\{\{1\},\{2\},\{3\},\{4\},\{5,7\},\{6\},\{8\},\{9\},\{10\},\{11\},\{12,13\},\{14\},\{15\},\{16\},\{17,20\},\{18\},\{19\}\}$$

$$U/ind(C-C_6)=\{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\},\{7\},\{8\},\{9\},\{10\},\{11\},\{12,13\},\{14\},\{15\},\{16\},\{17\},\{18\},\{19,20\}\}$$

$$U/ind(C-C_8)=\{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\},\{7\},\{8\},\{9\},\{10\},\{11\},\{12,13\},\{14\},\{15,17\},\{16\},\{18,20\},\{19\}\}$$

$$U/D=\{\{1,2\},\{3,4,5,6\},\{7,8,9\},\{10,11,12,13\},\{14,15,16,17,18,19,20\}\}$$

$$U/C=\{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\},\{7\},\{8\},\{9\},\{10\},\{11\},\{12\},\{13\},\{14\},\{15\},\{16\},\{17\},\{18\},\{19\},\{20\}\}$$

Then, by formula (3):

$$pos_{C-C_1}(D)=\{\{1\},\{2\},\{3\},\{4,5\},\{6\},\{7\},\{8\},\{9\},\{10\},\{11\},\{12,13\},\{14\},\{15\},\{16,20\},\{17\},\{18\},\{19\}\}$$

$$pos_{C-C_2}(D)=\{\{1,2\},\{3\},\{4\},\{6\},\{8,9\},\{10\},\{11\},\{12,13\},\{14\},\{15\},\{16\},\{17\},\{18\},\{19\},\{20\}\}$$

$$pos_{C-C_4}(D)=\{\{1\},\{2\},\{3\},\{4\},\{6\},\{8\},\{9\},\{10\},\{11\},\{12,13\},\{14\},\{15\},\{16\},\{17,20\},\{18\},\{19\}\}$$

$$pos_{C-C_6}(D)=\{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\},\{7\},\{8\},\{9\},\{10\},\{11\},\{12,13\},\{14\},\{15\},\{16\},\{17\},\{18\},\{19,20\}\}$$

$$pos_{C-C_8}(D)=\{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\},\{7\},\{8\},\{9\},\{10\},\{11\},\{12,13\},\{14\},\{15,17\},\{16\},\{18,20\},\{19\}\}$$

$$pos_{C}(D)=\{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\},\{7\},\{8\},\{9\},\{10\},\{11\},\{12\},\{13\},\{14\},\{15\},\{16\},\{17\},\{18\},\{19\},\{20\}\}$$

The dependence degree can be obtained by formulas (4):

$$\gamma_{C-C_1}(D)=\left|pos_{C-C_1}(D)\right|/|U|=\frac{14}{20}$$

$$\gamma_{C-C_2}(D)=\left|pos_{C-C_2}(D)\right|/|U|=\frac{15}{20}$$

$$\gamma_{C-C_4}(D)=\left|pos_{C-C_4}(D)\right|/|U|=\frac{18}{20}$$

$$\gamma_{C-C_6}(D)=\left|pos_{C-C_6}(D)\right|/|U|=\frac{18}{20}$$

$$\gamma_{C-C_8}(D)=\left|pos_{C-C_8}(D)\right|/|U|=\frac{17}{20}$$

$$\gamma_{C}(D)=\left|pos_{C}(D)\right|/|U|=\frac{20}{20}$$

4.3. Computational importance

The importance of each condition attribute is obtained by formula (5):

$$\sigma_{CD}(C_1)=\gamma_C(D)-\gamma_{C-C_1}(D)=\frac{20-14}{20}=0.3$$

$$\sigma_{CD}(C_2)=\gamma_C(D)-\gamma_{C-C_2}(D)=\frac{20-15}{20}=0.25$$

$$\sigma_{CD}(C_4)=\gamma_C(D)-\gamma_{C-C_4}(D)=\frac{20-18}{20}=0.1$$

$$\sigma_{CD}(C_6)=\gamma_C(D)-\gamma_{C-C_6}(D)=\frac{20-18}{20}=0.1$$

$$\sigma_{CD}(C_8)=\gamma_C(D)-\gamma_{C-C_8}(D)=\frac{20-17}{20}=0.15$$

If the value is greater, it means that the condition attribute has a greater impact on the performance.

From the results of the rough set attribute model, in this simulated test, answering the first, third, and fifth questions has almost no effect on the total scores of the students in this class. Analysis of the data, we find that this is because the students in this class are The overall level of mastery of each question type is similar. Most students can master the first and third questions, and the fifth question is the common weakness of all students. Multiple-choice questions have the greatest impact on grades, followed by fill in the blanks, Choose a question, answer the second question, and answer the fourth question.

REFERENCES

1. X.F.Yang, An empirical study on improving the academic performance of students with mathematics learning difficulties in rural areas, *Guizhou Normal University*, 2019. (in Chinese)
2. Y.Y.Kang, N.Zhang, G.M. Wang, W.J. She and Y.Y. Liu, High-efficiency mathematics learning high school students' mathematics achievement influence path, *Studys of Psychology and Behavior*, 3 (2016) 352-359. (in Chinese)
3. H.B.Yang, D.Z.Liu and R.K.Yang, The relationship between learning interest, self-efficacy, learning strategy and performance – A study on junior high school mathematics learning based on Kolb learning style, *Educational Science Research*, 10 (2015) 52-57. (in Chinese)
4. Q.L.Guo and W.J.Xu, Investigation and analysis of the factors affecting students'

learning performance, *Health Vocational Education*, 22(3) (2004) 84-85. (in Chinese)
5. J.H.Wang, S.H.Fan and Y.Q.Deng, Exploration and analysis of factors affecting students' academic performance, *Journal of Tianjin Polytechnic University*, 26(6) (2007) 86-88. (in Chinese).
6. X.Y.Jin, Applying statistics to analyze and evaluate the factors affecting students' academic performance, *Journal of Wuhan University of Science and Technology*, 5 (2006) 24-26. (in Chinese)
7. W.F.Teng, The influence of family factors on students' academic performance, *Characters of the Times·Theoretical Discussion*, (5) (2008), 157-159.(in Chinese)
8. Z.Y.Shen, A statistical analysis of various factors affecting students' learning performance, *Journal of Yanbian Education College*, 19(3) (2005) 7-10. (in Chinese)
9. Z.Pawlak, Rough sets, *International Journal of Computer and Information Sciences*, 11 (1982) 314-356.
10. S.Q.Wang, C.Y.Gao, C.Luo, G.M.Zheng and Y.N.Zhou, Research on feature selection/ attribute reduction method based on rough set theory, *Procedia Computer Science*, 154 (2019) 194-198.
11. J.Zhao, M.L.Jia, Z.N.Dong, D.YTang and Z.Liu, Accelerating information entropy-based feature selection using rough set theory with classifified nested equivalence classes, *Pattern Recognition*, 107 (2020) 107517.
12. J.Zhao, M.L.Jia, Z.N.Dong, D.YTang and Z.Liu, NEC: A nested equivalence class-based dependency calculation approach for fast feature selection using rough set theory, *Information Sciences*, 536 (2020) 431-453.