

Loan Repayment Status Prediction

Natthaphong Luangnaruedom¹, Somchai Prakanchaen², ¹Lecturer of Information System for Digital Business Department, College of Innovative Business and Accountancy, Dhurakij Pundit University, 110/1-4 Prachachuen Road, Lak Si District, Bangkok 10210 Thailand, e-mail : Natthaphong.Lua@dpu.ac.th, ²Research consultant, College of Innovative Business and Accountancy, Dhurakij Pundit University, 110/1-4 Prachachuen Road, Lak Si District, Bangkok 10210 Thailand e-mail : Somchai.Pra@dpu.ac.th

The aim of this research is to construct the debtor's loan repayment status pattern. These patterns were used to classify the new arrival of a prior debtor. There were two kinds of loan repayment status patterns: certain, uncertain rule (can't classify to particular decision variable value). Exclusion of inconsistent classifying of a decision attribute observation, the research classification rules are generated by the Rough set technique, and a decision tree. The best accuracy classification rule is used to predict loan repayment status of a new interest observation. A new arrival debtor could be classified by their possibility of their loan repayment status by using a similarity measurement with trained loan repayment status pattern. The attribute selection is performed in order to choose the significance attribute which is related to the decision variable. The experimental classification rules, that are generated from only a relevant attribute on a class variable loan repayment status, gives more accuracy than the model that is generated with all the conditional variables. The inconsistent observations are composed of type I and type II error which reduce the accuracy in the classification prediction. Nevertheless, inconsistent observations were also generated in their pattern. These patterns are used to predict whether the newly arrived customer and current customer will be an uncertain LRS level or not.

Key words: *Loan repayment status data pattern, classification rule, Rough set, classification tree.*

Introduction

Moneylender departments have to consider the loan application of a borrower if they should be the good debtor. Much loan repayment status prediction research attributes are depended on "credit bureau" customers or debtor status such as a black list loan repayment status feature. This research suggests a mathematical method of loan repayment status prediction, without

information from credit-bureau departments. Loan repayment status: LRS is defined as three types: NPL, irregularly repay, and regularly repay. The experimental data and attributes were specifically chosen from private moneylender companies. The borrower's credit bureau data was excluded in this research since this research is focused on small debt/loan applications. In order to present the practical calculations, demonstrable observations were considered and picked up especially to cover three types of repayment status. The conditional attributes were all gathered from existing loan application processing of the research's loan application borrower. Each attribute was examined whether it is significantly effective to the decision variable (LRS) or not. The remaining effective attributes were used to construct the data pattern of the loan repayment status. There are three types of generated data patterns. The first is consist rule. There were many observations or objects that were presented in consistency pattern. Some observations were not presented in the consistent data pattern. Two clusters of observation were separately generated in their data patterns. Each cluster of data pattern type may consist of many Boolean expressions. In this situation, it should cause the severe problem of much computational time. Therefore, Boolean expressions were reformatted to the small and less computation Boolean expression form (Binary form). After training data for classification rule generated, these rules were used to classify loan repayment status of the new loan application of the borrower. Normalised similarity measurement was applied to measure the most similarity of testing observation about loan repayment status. Testing observations were cross validated to measure the accuracy of the three generated classifying rule. Moreover, the cluster of inconsistent classification data rule was also generated for their data rule.

Related theory and research

Rough set (Pawlak, 1991)

Rough set analysis is a mathematical technique that can discover the hidden pattern in experimental data. It can be used to extract the features, approximate the concepts, generate decision rule, etc. The Rough set analysis can discriminate the observation, under specific boundary region of set, to certain "yes", certain "not", and uncertain classification. Information system (IS) is a table that contains the conditional attributes of an interested domain. If this table has also contains a decision variable then it is called the decision system (DS).

Table 1. Information system and decision system

	Conditional Attribute - A			Decision attribute-D
Object(x)	a ₁	a ₂	a ₃	y
1	1	1	2	y
2	1	2	1	y
3	1	1	2	y
4	1	2	1	n
5	2	2	1	n
6	2	2	1	y
7	2	1	1	n

Where,

“U” is a non-empty finite set of objects or observation.

“A” is a nonempty finite set of conditional attributes. There exist “a” is element of set A, $a \in A$. “d”, $d \notin A$, is decision attribute which “V_a” is value set of a,

$$a: U \rightarrow V_a. IS: T = (U, A) \quad (1)$$

$$DS: T = (U, A \cup \{d\}). \quad (2)$$

“X” is non-empty set of objects which all objects contain information in B-set of attributes, $B \subseteq A$. Indiscernibility (or similar) relation is the set of objects which has B-indiscernibility binary relation, $IND_{IS}(B)$. For example, there are two objects, x and x'. The objects will be contained in IND_{IS} if they follow the condition,

$$IND_{IS}(B) = \{(x, x') \in U^2 \mid \forall a \in B, a(x) = a(x')\}. \quad (3)$$

In short, B-indiscernibility relation can denote by $[x]_B$.

The Set approximation of “X”, under B-boundary information, is defined in two situations called B-lower and B-upper, where

$$B-lower = \underline{B}X = \{x \mid [x]_B \subseteq X\} \quad (4)$$

$$B-upper = \overline{B}X = \{x \mid [x]_B \cap X \neq \emptyset\}. \quad (5)$$

B-boundary region of “X” is calculated from

$$BN_B(X) = \overline{B}X - \underline{B}X. \quad (6)$$

The B-lower represents the set of objects that are certain compiled to B relation and decision attribute, The B-boundary represents the set of objects that are not certain compiled on B – relation and decision attribute. These three type of approximation are used to discriminate a

many objects approach to their B-relation and specific decision attribute. For example, according to a sample decision system in table 1. The result of calculations as shown in table 2. Assume that set “B” covers all attributes (a1, a2 and a3). Therefore, $IND_{IS}(B) = \{\{x1, x3\}, \{x2, x4\}, \{x5, x6\}, \{x7\}\}$.

Case# 1, if $W = \{x | d = Y\}$ therefore $\underline{BX} = \{\{x1, x3\}\}$, $\overline{BX} = \{\{x1, x3\}, \{x2, x4\}, \{x5, x6\}\}$, $BN_A(W) = \overline{BX} - \underline{BX} = \{\{x2, x4\}, \{x5, x6\}\}$, and $B\text{-outside-region} = U - \overline{BX} = \{x7\}$.

Case#2, if $W = \{x | d = N\}$ therefore $\underline{BX} = \{\{x7\}\}$, $\overline{BX} = \{\{x2, x4\}, \{x5, x6\}, \{x7\}\}$, $BN_A(W) = \overline{BX} - \underline{BX} = \{\{x2, x4\}, \{x5, x6\}\}$, $B\text{-outside-region} = U - \overline{BX} = \{x1, x3\}$.

Table 2. Result of set approximation

Object	Decision attribute					
	Y			N		
	Certain "Y"	Uncertain "Y"	Certain not "Y"	Certain "N"	Uncertain "N"	Certain not "N"
x1	*					*
x2		*			*	
x3	*					*
x4		*			*	
x5		*			*	
x6		*			*	
x7			*	*		

Accuracy of approximation (α), in each case, can devise B-lower by B-upper,

$$\alpha_B(X) = \frac{|\underline{BX}|}{|\overline{BX}|} \quad (7)$$

Where, “X” is crisp if the accuracy of approximation is equal to “1”. “X” is rough if the accuracy of approximation is less than “1”. Reducts are the subset of minimal attributes that can preserve the generated data pattern or classification rule. There are some attributes that are redundant. These redundant attributes can be removed without worsening the classification rules. Reduct attributes can be found out from discernibility matrix processing. Let $IS, DS: T = (U, A \cup \{d\})$, has “n” objects thus discernibility matrix is symmetric $n \times n$ matrix. Their element is C_{ij} . All element are calculated from

$$C_{ij} = \{a \in A | a(x_i \neq x_j)\} \text{ for } i, j = 1, n. \quad (8)$$

After all elements are calculated then discernibility function is calculated from

$$f_s(a_1^* \dots a_m^*) = \bigvee \{ \exists c_{ij}^* \mid 1 \leq j \leq i \leq n, c_{ij}^* \neq \emptyset \} \quad (9)$$

while $c_{ij}^* = \{a^* \mid a \in c_{ij}\}$. The discernibility function results will be included with only the reducts attribute that are related to “B” relation. For example, let us consider IS in table 1. Discernibility matrix of IS_A is shown in table 3.

Table 3. Discernibility matrix (DM)

Object	x1	x2	x3	x4	x5	x6	x7
x1	*			a2,a3	a1,a2,a3		a1,a3
x2		*			a1		a1,a2
x3			*	a2,a3	a1,a2,a3		a1,a3
x4	a2,a3		a2,a3	*		a1	
x5	a1,a2,a3	a1	a1,a2,a3		*		
x6				a1		*	a2
x7	a1,a3	a1,a2	a1,a3			a2	*

The discernibility function is calculated from $f_s(a_1^* \dots a_m^*) = (a2 \cup a3) \cap (a2 \cup a3) \cap (a1 \cup a2 \cup a3) \cap (a1) \cap (a1 \cup a2 \cup a3) \cap (a1) \cap (a1 \cup a3) \cap (a1 \cup a2) \cap (a1 \cup a3) \cap (a2) = (a1 \cap a2) \cup (a1 \cap a3)$.

This discernibility function value are presented in DNF format. The result of Boolean expression is $(a1 \cap a2) \cup (a1 \cap a3)$ thus a1, a2, and a3 are reducts attributes. The DM are also used to generate the data pattern or rule. For example, object x 4’s discernibility attributes to all other objects is (a2, a3), and (a2, a3). Thus, DNF is extracted from $(a2 \cup a3) \cap (a2 \cup a3) = (a2 \cup a3)$. Thus, the rule can be present as: if a2=1 or a3=1 -- > d=n. Since there should have been other attributes that are not specified in the interest rule attribute “a1” did not occur in this rule. Indeed, the “a1” possible occurring values are “1”, and “2”. Therefore, all possible occurring rules are: if a2=1 or a3=1 or a1=1 -- > d=n, and if a2=1 or a3=1 or a1=2 -- > d=n. Hence, the opportunity of occurring rules of object “x4” that followed to DNF is (the number of rule that follow DNF) / (total number of possible rule). In this example the probability of specific observation’s DNF rule is 0.5; (1/2). The possibility of the rule can represent the strength of the generated rule. Normally, there would be observations which follow the same DNF. RSES is a practical tool that provides calculation of reducts, data pattern or rule generation.

Similarity measurement (Stone, 2004)

The measurement of one object to other objects is to measure the similarity of all attributes of both objects. There are many similarity measure techniques. Each technique is suitable for some data type attributes. All techniques measure the distance (or distance) between two objects (i and j) on every attribute. If the distance value (d_{ij}) is transformed to normalisation value, such as “0-1” range, then similarity (s_{ij}) is $s_{ij} = 1 - d_{ij}$.

Table 4. Similarity measurement techniques

Technique	Data type			
	Nominal	Ordinal	Binary	Quantity
City block	*	*		*
Jaccard	*		*	
Euclidean				*

There are many techniques of distance measurement as follows.

A. City block (or Manhattan)

A City block similarity measurement is used to measure distance of two objects on specific attribute, whose attribute type is ordinal or quantity. The City block distance formula is

$$d_{ij} = \sum_{k=1}^n |x_{ik} - x_{jk}| \quad (10)$$

Where, “i” and “j” are index of object-i and object-j, k is index of attribute-k: k=1 to n.

B. Hamming distance for binary data type attribute formula is $d_{ij} = q + r$, where p is a number of variables that their value is positive in object x_i and x_j , and q is the number of variables of a object x_i is positive but negative in object x_j .

C. Euclidean distance is the distance measurement technique that is mostly used to measure distance of the object’s attribute with the quantitative data type. The formula of Euclidean distance is

$$d_{ij} = \sqrt{\sum_{k=1}^n (x_{ik} - x_{jk})^2} \quad (11)$$

D. Jaccard index or coefficient is the similarity value between two objects. The similarity of two objects is counted from the number of intersects (similar) over union. The Jaccard index formula is

$$S_{AB} = \frac{A \cap B}{A \cup B}. \quad (12)$$

Where, $0 \leq S_{AB} \leq 1$.

Jaccard index has a dominant character that the index is normalised in similarity with coefficient “0.0 – 1.0” value.

Ordinary, the distance values of many techniques are not in the same data range. This should cause a problem if the similarity measurements are made on a multivariate attribute. Each attribute distance value must be normalised to the same data range such as “0 – 1”. Thenormalised distance value of x_i and x_j all attributes were aggregated (summation) to form up total normalised distance. Subsequently, this aggregated is divided with the total number of attributes. This value is used to calculate the similarity between x_i and x_j . Nevertheless, the problem can be overcome by transforming the data type of all attributes to a binary data.

Data normalisation

Data normalisation is the process of transformed distance value from some data range, for example $d_{\min} - d_{\max}$, to other data range, such as $0.0 \leq d \leq 1.0$. For a quantitative data type, transformation of data range d_{\min}, d_{\max} can be calculated from the formula;

$$\frac{d - d_{\min}}{d_{\max} - d_{\min}}. \quad (13)$$

For example, if the original index value is “10, 40, 50” then the normalised value is “ $\frac{10-10}{50-10}, \frac{40-10}{50-10}, \frac{50-10}{50-10}$ ” or “ $0, \frac{3}{4}, 1$ ”. For ordinal data type, the original value must be set to the new range start with 1(r) to maximum number of value (R). For example, if the original index is “-1, 0, 1” then these indexes are set to “1, 2, 3”. The value of $r=1$ (starting number), and $R=3$ are used to normalise the set index by formula $d = \frac{r-1}{R-1}$. Thus, “1, 2, 3” are normalised to “

$$\frac{(1-1)}{3-1}, \frac{(2-1)}{3-1}, \frac{(3-1)}{3-1}” or “0, \frac{1}{2}, 1”.$$

F-measure (or F-score) (Marina, 2015)

The F-score is a measurement technique that is mostly used to measure accuracy of classification. The result of classification can be measured if the objects are specified as the true positive-TP, false positive-FN (type II error), false positive-FP (type I error), and true

negative-TN. F-score value is ranged between 0-1. The F-score equal to “1” means that the classification is perfect precision and recall. The F-score formula is

$$F - score = 2 \cdot \frac{P \cdot R}{P + R} \quad (14)$$

$$\text{While, } P = \frac{TP}{TP + FP} \text{ and } R = \frac{TP}{TP + FN} .$$

Cross validation testing (Type I and II error) (Anita, 2020)

Trained classification rules are generated from training data. These rules must be tested in their accuracy in the classification. The cross validation test tests the model accuracy of classification on another dataset. Since some experiment observations are all the good data content so that there may trend to generate a good classification rule. Therefore, a cross validation test should be performed on other data observations in order to prove if the accuracy of rule is believable.

Predicting Loan Repayment (Imad, 2018)

The lending industry has to consider borrowing features if there is a risky borrower. Data mining techniques are used to create the model to predict the level of a borrower’s risk in loan repayment. Many features are defined by the lending industry such as credit policy, purpose, installment, annual income, debt to income, borrower's revolving balance, credit score, number of derogatory public records, etc. Model learning techniques are gradient boosting, Support Vector Classifier, Random Forest. The result of studying show that there are no single best model for prediction since the economics is imbalanced. Other problems occurred from a lot of missing data and type I & type II error prediction. Therefore, prediction models are necessary for periodic testing in its prediction accuracy.

Is SME loan repayment influenced by specific attributes of borrowers?(Rania, 2015)

Rania kabir has experimented with loan repayments of borrowers regarding factors which influence the loan repayment. The research sample is two private commercial banks. There are many hypotheses that are proof of their significance by hypothesis testing. The findings are that the influencing factors on loan repayments such as amount of the borrowing, the longer loan term, age of the borrower, experience of the borrower, personal assets of the borrower.

What is a loan and types of loans (Peter, 2020)

Peter suggests the loan borrower about type of loan and consideration topics of lending. The consideration topic that borrower has to evaluate. There are many question to answer from a



borrower's point of view. Examples of questions are; the loan amount, loan security, interest rate, personal or business loan, fees and charges, long or short term, prepay policy, loan payment due, late payments fee. These questions are also used by money lender companies. These answers must be detected by lender whether there are true or fake.

Loan default prediction (Zakari, 2017)

Zakaria present his study about peer to peer loans. Private investors want to know of the possibility that a borrower or customer may default on aa loan. The sample was collected from a private lending company. The attributes of the borrower are state of residence, income, credit score, current status of a current loan, amount of principal and interest paid on a current loan, payment date of current loan, etc. Data mining techniques are used to discover the loan default prediction. The results show that a priori association rule on multi features give more accuracy in prediction.

The impact of socio-economic factors and attributes on repayment ability in Micro financing

Zeinab (2020) has studied the effect of borrower's socio-economic factors on loan repayment ability. The sample data waas collected from loan borrowers from Amhara Credit and Savings Institution, Ethiopia. The results of the study show that there are some significant socio-economic effects on loan repayments such as; the poor borrowers have bias in cash management more than wealthier and post moral hazard borrowers seem to fail on loan repay, etc.

Research methodology

This research is conducted with six research activities as illustrated in figure 1.

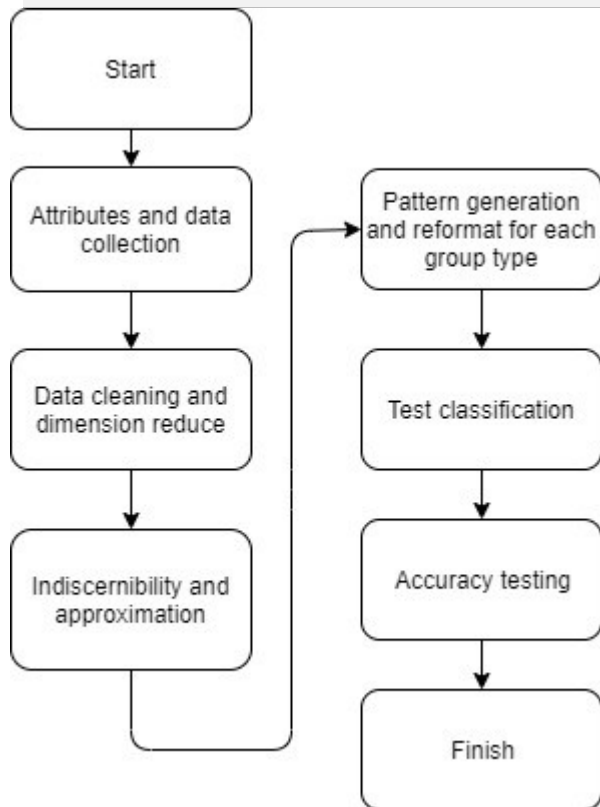


Figure 1. Research methodology

Loan Repayment Status attribute

The research is a definitive study on capital and short-term loans that is provided by a case study of private moneylender’s businesses in Thailand, therefore the conditional and decision attributes were all collected from the experimental company’s loan data processing system.

Table 5. Research’s conditional and decision attributes description

Variable label	Variable name	Data type	Data range	Type
1.Economic situation effect	ESE	Quantitative	1- effect,2-moderate or no effect	Conditional
2.Salary-stability	SS	Quantitative	1-nostability,2-moderate stability	Conditional
3.Education Level	EL	Nominal	1- secondary school or vocational school, 2-above	Conditional
4.Marriage status	MS	Nominal	1-divorce or marriage,2-single	Conditional
5.Number of children	NoC	Quantitative	1- ≥ 2 childs,2- ≤ 1 children	Conditional
6.Home owner	HO	Quantitative	1-house installment,2-rent a house or debt free house owner	Conditional
7.Informal debt	ID	Quantitative	1-exist,2-no informal debt	Conditional

8. Debt ratio	DR	Quantitative	1- effect, 2- moderate or no effect	Conditional
9. Loan repayment status	LRS	Quantitative	1-NPL, 2-irregularly repay, 3-regularly repay	Decision

Data collection

Experimental observations were considered and collected from the customer (debtor) loan's database. All debtor's privacy identification data was concealed. This sample data was gathered during 2018-2019. There were three hundred observation samples. In research planning, eighty percent of observations (240 observations) were chosen for model training; while sixty observations were used in the model cross validation test.

Table 6. Partial research sample data

	ESE	SS	EL	MS	NoC	HO	ID	DR	LRS
1	1	1	1	1	1	1	1	1	1
2	2	1	2	2	2	1	2	1	2
3	1	2	2	1	1	2	1	1	1
4	2	1	2	2	2	2	2	1	1
5	2	1	2	2	2	1	2	1	1
6	2	1	2	2	2	1	2	1	2
7	1	1	1	1	1	1	1	1	2
8	2	1	2	2	2	1	2	1	2
9	2	1	2	1	2	1	2	2	3
10	2	2	1	2	2	2	2	2	3
...

Data cleaning and attribute selection

Observation data was entered by the customer and verified by the moneylender company clerks. Therefore, the data could not be altered without debtor permission. Nevertheless, some data content may cause an effect to some attributes. The one who has SS=3 seems not to effect attribute ESE=1. The ESE value seems to be considered as an incorrect data value. This value may be changed to value "2" or "3". Otherwise, this observation may be deleted. All attributes were checked whether it is related to class variable or whether it has a significant relation to the decision variable or not. Reducts attribute, or essential significance attribute, were tested by discernibility matrix (DM) of the Rough set analysis. The result of the DM shows that the reduct variables are Economic situation effect-ESE, Salary-stability-SS, Number of children-NoC, Home owner-HO, and Debt ratio-DR. Whilst, dispensable variables are Education Level-EL, Informal debt-ID, and Marriage status-MS.

Indiscernibility and approximation

The remaining attributes, after feature selection, are used to construct the indiscernibility relation. If, $A = \{ESE, SS, NoC, HO, DR\}$ then (partial example) indiscernibility relation under “A” is $IND_{IS}(A) = \{\{1, 7, \dots\}, \{2, 5, 6, 8, \dots\}, \{3, \dots\}, \{4, \dots\}, \{9, \dots\}, \{10, \dots\}, \dots\}$. If “B” is the set that coverage on decision attribute “LRS”:

“LRS”=1 then $\underline{BX} = \{\{3\}, \{4\}, \dots\}$, $\overline{BX} = \{\{1, 7, \dots\}, \{2, 5, 6, 8, \dots\}, \{3\}, \{4\}, \dots\}$, and

$BN_1 = \overline{BX} - \underline{BX} = \{\{1, 7, \dots\}, \{2, 5, 6, 8, \dots\}, \dots\}$. If

“LRS”=2 then $\underline{BX} = \{\{\phi\}, \dots\}$, $\overline{BX} = \{\{1, 7, \dots\}, \{2, 5, 6, 8, \dots\}, \dots\}$, and

$BN_2 = \overline{BX} - \underline{BX} = \{\{1, 7, \dots\}, \{2, 5, 6, 8, \dots\}, \dots\}$. If

“LRS”=3 then $\underline{BX} = \{\{9\}, \{10\}, \dots\}$, $\overline{BX} = \{\{\phi\}, \dots\}$, and $BN_3 = \overline{BX} - \underline{BX} = \{\{9, \dots\}, \{10, \dots\}, \dots\}$

From all calculated approximation, calculation results show that some observations are not consistent from point to decision variable in a particular value, such as LRS=1,2, or 3. These observations will be defined as members of the inconsistent group. The inconsistent observations are not believable observations thus they will not be further used in data rule generating. The remaining set of indiscernibility (or similar) observations will be sent to perform their data pattern. The data pattern of consistent observations can be constructed by many techniques such as the Rough set, Decision tree, Association rule, etc. There are only two hundred and fifty observations remaining after inconsistent observation deletion.

Table 7. Example of inconsistent and consistent observation

	ESE	SS	NoC	HO	DR	LRS	Note
1	1	1	1	1	1	1	Inconsistent
2	2	1	2	1	1	2	Inconsistent
3	1	2	1	2	1	1	Consistent
4	2	1	2	2	1	1	Consistent
5	2	1	2	1	1	1	Inconsistent
6	2	1	2	1	1	2	Inconsistent
7	1	1	1	1	1	2	Inconsistent
8	2	1	2	1	1	2	Inconsistent
9	2	1	2	1	2	3	Consistent
10	2	2	2	2	2	3	Consistent
...

All inconsistent observations are deleted from the dataset. The remaining observation are the observations that will not cause any inconsistency rules generation.

Table 8. Partial observation detail (only consistent observations)

Observation	ESE1	ESE2	SS1	SS2	NoC1	NoC2	HO1	HO2	DR1	DR2	LRS	Note
...												
3	1			1	1			1	1		1	Consistent
4		1	1			1		1	1		1	Consistent
...												
9		1	1			1	1			1	3	Consistent
10		1		1		1		1		1	3	Consistent
...												

Data pattern generation (Rule)

Two hundred observations were tried out for their data pattern by RSES2, rough set software tool. Normally, RSES2 will generate a huge number of data pattern rules. This will cause a problem in computing time delay. Nevertheless, inconsistent observation deletion diminishes the inconsistency data rule and increases the rule generation strength or high probability. Normally, the generated rules are presented in DNF or CNF format. This research suggests problem-solving by DNF to Boolean reformatting. The reformatting process performing has to separate each attribute to the new variable along its data range. All attributes, except “LRS” were reformatted from their data type to Binary data type.

Table 9. Attributes in new sub-attribute data range

Variable label	Variable name	Data range
Economic situation effect	ESE1, ESE2	1- effect, 2-moderate or no effect
Salary-stability	SS1, SS2	1-nostability, 2-moderate stability
Number of children	NoC1, NoC2	1->= 2 children, 2-<=1 children
Home owner	HO1, Ho2	1-house installment, 2-renter or debt free home owner
Debt ratio	DR1, DR2	1- effect, 2-moderate or no effect
Loan repayment status	LRS*	1-NPL, 2-irregularly repay, 3-regularly repay

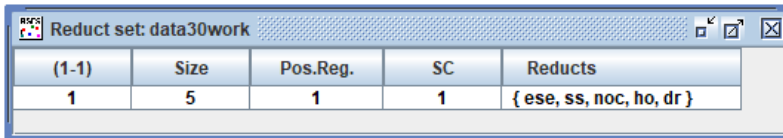
* LRS is not set to new variable

The dataset is prepared and saved as “.arff” file format in order to be processed in RSES2, and WEKA software tool. Partial observation and their description in “.arff” file formats follow and are shown in figure 2.

```
%ESE,SS,NoC,HO,DR,LRS
@relation LRS_200_Data set training
@attribute ese {1,2}
@attribute ss {1,2}
@attribute noc {1,2}
@attribute ho {1,2}
@attribute dr {1,2}
@attribute class {1,2,3}
@data
1,1,1,1,1,1
1,2,2,1,1,1
2,1,2,2,2,2
1,2,2,1,2,2
2,2,1,1,2,2
1,2,2,2,2,3
2,1,2,1,2,2
```

Figure 2. Partial “.arff” of training two hundred data observations.

RSES2, Rough set tool, has detected the training data set again in reduct (significance) attributes. The result of the reduct attribute seeking are still coverage on the former reduct attribute; ESE, SS, NoC, HO, and DR. The possibility of a rule occurring (or rule strength), rule significant, is “1.00”.



(1-1)	Size	Pos.Reg.	SC	Reducts
1	5	1	1	{ ese, ss, noc, ho, dr }

Figure 3. Reduct attributes with pos. reg 1.00 value

These significant attributes are sent to rule generate processing. There are nineteen data patterns. Class 1, 2, and 3 represent as “NPL, Irregular repay, and Regular repay”. Numbers in the square brackets represent the number of observations that follow a particular data rule.

RULES 19

```
(ese=1)and(ss=1)and(noc=1)and(ho=1)and(dr=1)=>(class=1[5])
(ese=2)and(ss=1)and(noc=2)and(ho=1)and(dr=1)=>(class=1[12])
(ese=1)and(ss=2)and(noc=1)and(ho=2)and(dr=1)=>(class=1[13])
(ese=1)and(ss=2)and(noc=2)and(ho=2)and(dr=1)=>(class=1[8])
(ese=2)and(ss=1)and(noc=1)and(ho=1)and(dr=2)=>(class=1[6])
(ese=1)and(ss=2)and(noc=2)and(ho=1)and(dr=1)=>(class=1[11])
(ese=2)and(ss=1)and(noc=2)and(ho=2)and(dr=2)=>(class=2[12])
(ese=1)and(ss=2)and(noc=2)and(ho=1)and(dr=2)=>(class=2[6])
(ese=2)and(ss=2)and(noc=1)and(ho=1)and(dr=2)=>(class=2[15])
(ese=1)and(ss=2)and(noc=1)and(ho=1)and(dr=2)=>(class=2[11])
(ese=2)and(ss=1)and(noc=2)and(ho=1)and(dr=2)=>(class=2[8])
(ese=2)and(ss=2)and(noc=1)and(ho=2)and(dr=1)=>(class=2[14])
(ese=1)and(ss=2)and(noc=2)and(ho=2)and(dr=2)=>(class=3[9])
(ese=2)and(ss=2)and(noc=2)and(ho=2)and(dr=1)=>(class=3[17])
(ese=2)and(ss=2)and(noc=1)and(ho=2)and(dr=2)=>(class=3[7])
(ese=1)and(ss=2)and(noc=1)and(ho=2)and(dr=2)=>(class=3[21])
(ese=2)and(ss=2)and(noc=2)and(ho=2)and(dr=2)=>(class=3[14])
(ese=2)and(ss=1)and(noc=1)and(ho=2)and(dr=2)=>(class=3[9])
(ese=1)and(ss=1)and(noc=1)and(ho=1)and(dr=2)=>(class=3[2])
```

Figure 4. Generated data pattern of LRS by RSES2

Data pattern reformat

The generated rules are separated into three groups. Each group contains only the same class variable. Each group Boolean expression are shown in Table 10.

Table 10. Generate rule separate on class

Class-LRS	
NPL (Non-performing loan)	(ese=1)and(ss=1)and(noc=1)and (ho=1)and(dr=1)
	(ese=2)and(ss=1)and(noc=2)and(ho=1)and(dr=1)
	(ese=1)and(ss=2)and(noc=1)and(ho=2)and(dr=1)
	(ese=1)and(ss=2)and(noc=2)and(ho=2)and(dr=1)
	(ese=2)and(ss=1)and(noc=1)and(ho=1)and(dr=2)
	(ese=1)and(ss=2)and(noc=2)and(ho=1)and(dr=1)
Irregular loan repay	(ese=2)and(ss=1)and(noc=2)and(ho=2)and(dr=2)
	(ese=1)and(ss=2)and(noc=2)and(ho=1)and(dr=2)
	(ese=2)and(ss=2)and(noc=1)and(ho=1)and(dr=2)
	(ese=1)and(ss=2)and(noc=1)and(ho=1)and(dr=2)
	(ese=2)and(ss=1)and(noc=2)and(ho=1)and(dr=2)
	(ese=2)and(ss=2)and(noc=1)and(ho=2)and(dr=1)
Regular loan repay	(ese=1)and(ss=2)and(noc=2)and(ho=2)and(dr=2)
	(ese=2)and(ss=2)and(noc=2)and(ho=2)and(dr=1)
	(ese=2)and(ss=2)and(noc=1)and(ho=2)and(dr=2)
	(ese=1)and(ss=2)and(noc=1)and(ho=2)and(dr=2)
	(ese=2)and(ss=2)and(noc=2)and(ho=2)and(dr=2)
	(ese=2)and(ss=1)and(noc=1)and(ho=2)and(dr=2)
(ese=1)and(ss=1)and(noc=1)and(ho=1)and(dr=2)	

Test observation classification

In order to decrease the similarity computational time, all Boolean expressions are transformed to a binary number. For example, observation #1 ESE attribute value is “1” then ESE1=1 and ESE2=0. All attributes, except LRS, are transformed.

Table 11. Transformed data patterns from binary number to Boolean expression.

Rule		ESE1	ESE2	SS1	SS2	NoC1	NoC2	HO1	HO2	DR1	DR2	LRS
NPL	1	1	0	1	0	1	0	1	0	1	0	1
	2	0	1	1	0	0	1	1	0	1	0	1
	3	1	0	0	1	1	0	0	1	1	0	1
	4	1	0	0	1	0	1	0	1	1	0	1
	5	0	1	1	0	1	0	1	0	0	1	1
	6	1	0	0	1	0	1	1	0	1	0	1
Irregular loan repay	1	0	1	1	0	0	1	0	1	0	1	2
	2	1	0	0	1	0	1	1	0	0	1	2
	3	0	1	0	1	1	0	1	0	0	1	2
	4	1	0	0	1	1	0	1	0	0	1	2
	5	0	1	1	0	0	1	1	0	0	1	2
	6	0	1	0	1	1	0	0	1	1	0	2
Regular loan repay	1	1	0	0	1	0	1	0	1	0	1	3
	2	0	1	0	1	0	1	0	1	1	0	3
	3	0	1	0	1	1	0	0	1	0	1	3
	4	1	0	0	1	1	0	0	1	0	1	3
	5	0	1	0	1	0	1	0	1	0	1	3
	6	0	1	1	0	1	0	0	1	0	1	3
	7	1	0	1	0	0	1	0	1	0	1	3

There are some data observations that their class variable are unknown. Therefore, its observation attributes are used to seek for the most similarity on the Boolean expression rules. The most similar Boolean expression implies the type of class variable, LRS. For example, the unknown class type observation of all attribute content and its transformed Boolean expression is shown in table 12.

Table 12. Detail of original data of the test observation and its transformed data

Observation	ESE	SS	NoC	HO	DR	LRS						
Test	2	2	1	2	2	?						
Observation	ESE1	ESE2	SS1	SS2	NoC1	NoC2	HO1	HO2	DR1	DR2	LRS	
Test	0	1	0	1	1	0	0	1	0	1	?	

The normalised similarity is calculated with a Jaccard index (coefficient). The calculation similarity of the experiment test observation must be undertaken with all Boolean expression generate rules. The similarity results is presented in table 13. For example, the similarity index of the test observation with observation #1 is $S_{Test\&rule\#1} = \frac{2}{10} = 0.20$.

Table 13. Jaccard similarity index of the test observation with all Boolean generated rules

Rule	#	Similarity	LRS
NPL	1	0.20	1
	2	0.20	1
	3	0.60	1
	4	0.40	1
	5	0.60	1
	6	0.20	1
Irregular loan repay	1	0.60	2
	2	0.40	2
	3	0.80	2
	4	0.60	2
	5	0.40	2
	6	0.80	2
Regular loan repay	1	0.60	3
	2	0.60	3
	3	1.00*	3
	4	0.80	3
	5	0.80	3
	6	0.80	3
	7	0.40	3

*note the most similarity value 1.00 indicates LRS class Regular loan repay #3

Since the most similar test observation is 1.00 value with rule#3 of regular loan repayment, therefore the test observation class type is the “Regular loan repay” class.

LRS prediction by other classification technique

With the same two hundred observation data set, this research also tried out the rule generation by another data mining technique. Two hundred training observations with all research original attributes were trained with classification technique such as ID3, J48, etc. The best technique is classification tree ID3. The classification rules are illustrated in table 14.

Table 14. Generated rules that are separate on class-LRS by classification tree –ID3

Class-LRS	
NPL (Non-performing loan)	(ese=1)and(dr=1)
	(ese=2)and(ss=1)and(dr=1)
	(ese=2)and(ss=1)and(noc=1)and(ho=1)and(dr=2)
Irregular loan repay	(ese=2)and(ss=2)and(noc=1)and(ho=1)and(dr=1)
	(ss=1)and(noc=2)and(ho=2)and(dr=2)
	(ss=2)and(noc=1)and(ho=1)and(dr=2)
	(ss=1)and(noc=2)and(ho=1)and(dr=2)
Regular loan repay	(ese=2)and(ss=2)and(noc=2)and(dr=1)
	(ese=1)and(ss=1)and(noc=1)and(ho=1)and(dr=2)
	(ss=1)and(noc=1)and(ho=2)and(dr=2)
	(ese=1)and(ss=2)and(noc=1)and(ho=2)and(dr=2)
	(ss=2)and(noc=2)and(ho=2)and(dr=2)

Cross validation test

Fifty test observations were cross validation tested by both classification techniques. The results show that the Rough set technique gives more accurate classification than the classification tree ID3.

Table 15. Comparison of accuracy test between Rough set and decision tree ID3

Technique	Detail about accuracy				
	Precision	Recall	F-measure	Correct classification	Incorrect classification
Rough set	0.89	0.86	0.88	87.42	12.58
ID3 (All attributes)	0.49	0.53	0.51	53.33	46.67
ID3 (only relevant attributes)	0.85	0.83	0.83	83.33	16.67

4. Research results

Two classification techniques were tried out in this research. Both techniques provide a desirable result of correct classification in the dataset whose attributes were significant or related to a class variable, LRS. In further research, the accuracy of classification should be increased if the other relevant data attributes are considered additional in the classification. Rough set analysis can specify the observations that are inconsistent to classify a class variable. Therefore, the remaining observations are good and validity supports rule generating.

“Table 16.” represents the generated rules that are calculated from the inconsistent observations. The rule represent that in the same value of all attributes have classified more value of LRS. For example, “1.(ese1=1)and(noc1=1)and(ho1=1)and(dr1=0)” can predict LRS value in “2” and “3”.

Table 16. Generated rules separated on class (inconsistent observation)

Class-LRS	
NPL(1)	-
Irregular loan repay(2) & Regular loan repay(3)	1.(ese1=1)and(noc1=1)and(ho1=1)and(dr1=0)
	2.(ese1=1)and(noc1=1)and(ho2=0)and(dr1=0)
	3.(ese1=1)and(noc1=1)and(ho1=1)and(dr2=1)
	4.(ese1=1)and(noc2=0)and(ho1=1)and(dr1=0)
	5.(ese2=0)and(noc1=1)and(ho2=0)and(dr1=0)
	6.(ese2=0)and(noc1=1)and(ho1=1)and(dr2=1)
	7.(ese2=0)and(noc2=0)and(ho1=1)and(dr1=0)
	8.(ese2=0)and(noc1=1)and(ho2=0)and(dr2=1)
	9.(ese2=0)and(noc2=0)and(ho2=0)and(dr1=0)
	10.(ese1=1)and(noc1=1)and(ho2=0)and(dr2=1)
	11.(ese1=1)and(noc2=0)and(ho2=0)and(dr1=0)
	12.(ese1=1)and(noc2=0)and(ho1=1)and(dr2=1)
	13.(ese2=0)and(noc1=1)and(ho1=1)and(dr1=0)
	14.(ese1=1)and(noc2=0)and(ho2=1)and(dr2=1)
	15.(ese2=0)and(noc2=0)and(ho1=1)and(dr2=1)
	16.(ese2=0)and(noc2=0)and(ho2=0)and(dr2=1)

Summary and Suggestion

Even through the inconsistent observations are undesirable observations for most correct classification rule generation, but the inconsistent class classification observation serves as valuable support in a loan request consideration. These debtor transaction may be calculated for similarity index to all generated rule.s The generated rules represent the pattern of observation that indicates more than one LRS value. The most similarity generated rule is then picked up. The moneylenders company has to seek for the attributes which cause the customer’s application failed consideration. If there are possible improvements on some customer features then the moneylender’s company may ask debtors to adapt or improve something in order to pass the customer borrowing request. Another advantage of this research is the data of proportional loan LRS class of all debtors. Some debtors may come to reduce their good LRS class failing to the lower LRS class. The generated rule pattern should forewarn the manager about the loan debtor who has a similarity score or the possibility to get close to “NPL”. Therefore, the manager can do something to prevent this business risk.



Acknowledgements

This experimental research dataset was provided by some moneylender companies in Thailand. Within the data there were many observations that occurred about twenty years ago. The economic situation then was different and showed recession, or a prosperous economy. These criteria should make the data classification an unbelievable result. Thus, sample data were gathered during 2018-2019. With help from the experimental moneylenders company, this research proceeded successfully.

REFERENCES

- Anita Wasilewska, Classification: Testing Classifier Accuracy, Stony Brook University, New York, USA, 2020.
- Imad Dabbura, Predicting Loan Repayment, Baylor Scott & White Health, Texas, USA, 2018.
- Marina Santini, Basic concept of data mining, Uppsala university, Sweden, 2015.
- Pawlak, Zdzisław (1991). Rough Sets: Theoretical Aspects of Reasoning About Data Dordrecht: Kluwer Academic Publishing. ISBN 978-0-7923-1472-1.
- Peter Miller, What is a loan and learn about types of loans, Money rates company, California, USA, 2020.
- Rania Kabir, Does SME loan repayment influenced by specific attributes of borrowers – A study on selected banks, Indian Journal of Commerce & Management Studies ISSN: 2240-0310 EISSN: 2229-5674, 2015.
- Stone and Jame v., Similarity Searching and Information Retrieval, Data mining, MIT press, 2004.
- Zakaria and Dmitriy, Loan default prediction and identification of interesting relations between attributes of peer-to-peer loan applications, New zeal and Journal of Computer-Human Interaction ZJCHI 2,2 (2017), ISSN 2463-5626.
- Zeinab Hassano & Felicia Nordgren, The impact of socio-economic factors and attributes on repayment ability in Micro financing, Södertörn University, Sweden, 2020.