Rough Set Model Based Knowledge Acquisition of Market Movements from Economic Data

Yoshiyuki Matsumoto and Junzo Watada

Abstract The concept and method of rough sets were proposed by Z. Pawlak in 1982. This method enables us to mine knowledge granules as decision rules from a database, a web base, a set and so on. The obtained decision rules can be applicable for data analysis as well as used to reason, estimate, evaluate, or forecast an unknown object. The objective of this paper is to apply the rough set method to time series data for mining knowledge granules, and especially to mine knowledge granules from the data set of tick-wise price fluctuations.

Keywords Rough set model · Knowledge acquisition · Market movement · Economic data · Knowledge granule · Decision rule · Data analysis · Tick-wise price

1 Introduction

The analysis of time-series data is widely used in corporations and economics. Especially, the technical analysis is used to model the change of prices based on the graphical expression of market movement and the fundamental analysis is applied to understand the corporate performance and economic environment of a company. Also Y. Matsumoto and J. Watada employ the chaotic method to forecast the future price value [1]. This paper employs rough sets method to analyze time-series data [2, 3]. The rough sets analysis is proposed by Z. Pawlak to acquire rule-based

Y. Matsumoto
Shimonoseki City University, 2-1-1, Daigaku-Cho, Shimonoseki
Yamaguchi 751-8510, Japan
e-mail: matsumoto@shimonoseki-cu.ac.jp

J. Watada (✉)
Waseda University, 2-7 Hibikino, Wakamatsu-ku, Kitakyushu
Fukuoka 808-0135, Japan
e-mail: junzow@osb.att.ne.jp

W. Pedrycz and S.-M. Chen (eds.), *Information Granularity, Big Data, and Computational Intelligence*, Studies in Big Data 8,
DOI: 10.1007/978-3-319-08254-7_18,

knowledge from a set defined with plural attributes. The objective of this paper is to mine the knowledge granules of price movement from the intra-day trading data of stock prices called tick. Each dealt price is recorded whenever the dealing is accomplished. This paper aims to mine the knowledge of forecasting from the tick data, and using the knowledge gained, we predicted stock prices.

2 Rough Sets Theory

A rough set is especially useful for domains where the data collected are imprecise and/or incomplete about the domain objects. It provides a powerful tool for a data analysis and data mining of imprecise and ambiguous data. A reduction is the minimal set of attributes that preserves the indispensability relation, that is, the classification power of the original dataset [4]. Rough set theory has many advantages, such as providing efficient algorithms for finding hidden patterns in data, finding minimal sets of data (data reduction), evaluating the significance of data, and generating the minimal sets of decision rules from data. It is easy to understand and to offer a straightforward interpretation of the results [5]. These advantages can simplify analyses, which is why many applications use a rough set approach as their research method. The rough set theory is of fundamental importance in artificial intelligence and cognitive science, especially in the areas of machine learning, knowledge acquisition, decision analysis, knowledge discovery from databases, expert systems, decision support systems, inductive reasoning, and pattern recognition [6, 7, 8].

Rough set theory has been applied to the management of many various issues, including expert systems, empirical study of materials data [9], machine diagnosis [10], travel demand analysis [11], web screen design [12], IRIS data classfication [13], business failure prediction, solving linear programs, data mining [14] and α-RST [15]. Another paper discusses the preference-order of the attribute criteria needed to extend the original rough set theory, such as sorting, choice and ranking problems [16], the insurance market [17], and unifying rough set theory with fuzzy theory [18]. Rough set theory provides a simple way to analyze data and reduct information.

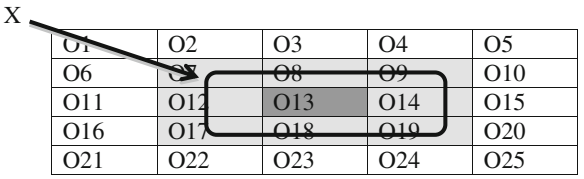
2.1 Information Systems

Generally, an information system denoted *IS* is defined as $IS = (U, A)$, where universe U consists of finite objects and A is named a universe and A is a finite set of n attributes $\{a_1, a_2, \ldots, a_n\}$. Each attribute a belongs to set A, that is, $a \in A$. An object ω ($\omega \in$ U) has a value $f_a(\omega)$ for each attribute, which is defined as f_a: $U \rightarrow V_a$. f_a means that object ω in the U has a value $f_a(\omega) \in V_a$ for attribute $a \in A$, where V_a is a set of values of attribute $a \in A$. It is called a domain of attribute a (Table 1).

Table 1 Sample Information System

Object	P1	P2	P3	P4
O1	+	−	+	−
O2	−	−	−	+
O3	+	+	+	+
O4	−	−	+	−
O5	−	+	−	−

Fig. 1 Upper and lower approximations



2.2 Lower and Upper Approximations

Rough sets analysis is based on the two basic concepts: lower and upper approximations of a focal set. The upper approximations is the set of all elements which possibly belong to the focal set and the lower approximations of all elements which doubtlessly belong to the focal set. Let X be a subset of elements in universe U, that is, X ⊂ U. Let us consider a subset in V_a, $P \subseteq V$.

The low approximation of P, denoted PX, is the union of all equivalence classes which are contained by the target set as follows: $\underline{PX} = \{x|[x]_P \in X\}$.

The upper approximation of P, denotes $\overline{PX}$, is the union of all equivalence classes which have non-empty intersection with the target set as follows: $\overline{PX} = \{x|[x]_P \cap X \neq \phi\}$.

The boundary set of X in U is defined as following: $PNX = \overline{PX} - \underline{PX}$ (Fig. 1).

2.3 Decision Ruless

An information system denoted IS is defined as $IS = (U, A)$, A can be partitioned into two disjoint classes $C, D \subseteq A$ of attributes, called condition and decision attributes, respectively. Here $IS = (U, C, D)$ is called decision system.

Decision rules can also be regarded as a set of decision (classification) rules of the form: $a_k \rightarrow d_j$, where a_k means that attribute a_k has value 1, d_j means the decision attributes and the symbol → denotes propositional implication. In the decision rule $\theta \rightarrow \varphi$, formulae θ and φ are called condition (premise) and decision (conclusion), respectively [19]. The decision rules we can minimize the set of attributes, reduce the superfluous attributes and classify elements into different

groups. In this way we can have many decision rules, each rule shows meaningful attributes. The stronger rule will cover more objects and the strength of each decision rule indicate the appropriateness of rules.

2.4 Analysis of Decision Rules

Only decision rules that are obtained rough set theory and have high C.I. are employed in reasoning. C.I. is an abbreviation of Covering Index that is a rate of objects that can sufficiently reach the same decision attribute by the rule out of the whole objects [20]. If whole objects number is 5, corresponding objects number is 3, C.I. is 0.6.

Generally speaking, decision rules with high C.I. are highly reliable and results in good reasoning. In real situations, the number of obtained decision rules is often more than several hundreds. In these cases, reasoning can not employ almost all decision rules. That is, reasoning scattered almost decision rules.

It is necessary to make decision rules effective so as to combine decision rules by means of decision rule analysis [21]. Decision rule analysis enables us to obtain new combined decision rules by means that premises of decision rules are decomposed and given some points depending on their C.I. value. This method enables us to take all decision rules into consideration even if rules have a low C.I. value. In this paper, decision rules are combined and applied to forecasting

Let us explain the detail of decision rule analysis. The decision rule analysis determines rules by calculating their column scores. The column score can be calculated in the following:

Let us consider the following three rules.

$$\begin{gathered} \text{IF a} = 1 \text{ and } \text{b} = 1 \text{ then d} = 1(\text{C.I.} = 0.4) \\ \text{IF b} = 2 \text{ then d} = 1(\text{C.I.} = 0.3) \\ \text{IF a} = 2 \text{ and b} = 2 \text{ and c} = 1 \text{ then d} = 1(\text{C.I.} = 0.6) \end{gathered} \tag{1}$$

The column score can be obtained using the combination table as shown in 0. The combination table is an n × n matrix consisting of all attributes. An element of the combination table is a score of combination of two attributes.

For example, the first rule has a = 1 and b = 1 as its premises. On this case, the vertical column has a = 1 and the horizontal row has b = 1, and the vertical column has b = 1 and horizontal row has a = 1. We describe two scores in these elements. The score value is one or C.I. value divided by the written score value.

On this case, two elements have each score value.

$$0.4/2 = 0.2 \tag{2}$$

On the case of the second rule, as the premise has one attribute, the column and row are written 0.3 for b = 2.

Table 2 Combination table

	a = 1	a = 2	b = 1	b = 2	c = 1	c = 2	Column score
a = 1			0.2				0.2
a = 2				0.1	0.1		0.2
b = 1	0.2						0.2
b = 2		0.1		0.3	0.1		0.5
c = 1		0.1		0.1			0.2
c = 2							

On the case of the third rule, as the premise has 3 attributes, 6 elements ($_3C_2 = 6$) should be written scores. The written score is

$$0.6/6 = 0.1 \tag{3}$$

The column score is the total value of scores in each column. For example, on the case of a = 2 we obtain

$$0.1 + 0.1 = 0.2 \tag{4}$$

This calculation results in Table 2. Using this combination we can derive a decision table. For example, on the case of column b = 2, since there is a score in a = 2, b = 2 and c = 1, the rule of this column results in as follows:

$$\text{IF a} = 2 \text{ and b} = 2 \text{ and c} = 1 \text{ then d} = 1 \tag{5}$$

Usually, scores under the some threshold are not accepted. For instance, when the threshold is 0.2, the rule is written in the following:

$$\text{IF b} = 2 \text{ then d} = 1 \tag{6}$$

3 Regression Line-Base Analysis

In general, rough sets analysis deals with categorical data. Therefore, in this paper we obtain regression line for the time-series data to forecast future values and use the up and down trends of the regression line as a condition attribute. For example, when we analyze past six fiscal terms, we can obtain the trends of the regression line for all six terms, the former three terms and the latter three terms. The trend a is decided as follows:

$$a = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

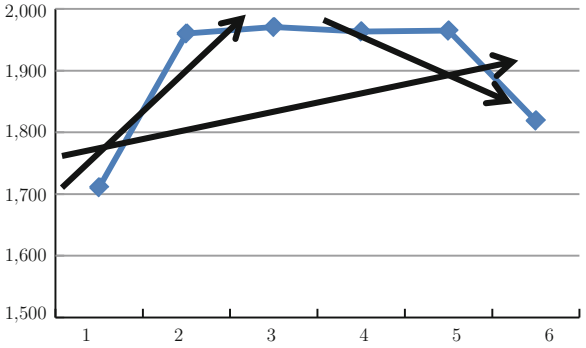


Fig. 2 Trends by linear regression

Table 3 Trends of data

No.	Condition attribute (Trend)			Decision attribute
	Total term	Former term	Latter term	Present term
1	+	+	−	−
2	+	−	+	+
3	+	−	−	−
4	+	−	−	+
5	−	−	+	−

The obtained trend a is employed as a condition attribute. That is, relating to each of the total, former and latter parts we forecasted whether each part has increasing or decreasing trend, depending on such data. On the case as shown in Fig. 2, the condition attribute shows that the total trend shows plus trend, the former part has plus trend and the latter part indicates minus trend. Such values are evaluated concerning with each of data, then we mine the knowledge by using rough sets analysis. Table 3 illustrates this process.

According to this process, we analyzed the trend of each of past data and forecasted the up and down movements of the present term depending on the data.

4 Rough Sets-Based Knowledge Acquisition of Market

In this research study, we analyzed the time series data of the market by using the regression line, and knowledge acquisition from the result by using rough sets. Using the regression line, the trend of the data can be understood, And it is possible to predict the state after a certain time. The analysis used the company's stock price data of the Tokyo stock market. We acquired knowledge granules by using the stock price of Fuji Heavy Industries on June 02, 2008. The original data are tick data, in this analysis are used to acquire the knowledge granules to create a 1 min chart data from its data. We were using data that has been trading in

between 9 and 11 a.m. Knowledge acquisition is done for the trend data of up to previous 12 min. We are classified the data into seven types (whole, first half, second half, first quarter, second quarter, third quarter, fourth quarter). We analyzed the trend of rising and falling for seven types. The decision attribute is whether the value increased or decreased after 1 min. We acquired the knowledge granule to predict the increase or decrease movement of 1 min using the previous 12 min data. We are creating a decision table. Objects in the decision table is a trend in each time. Conditions attributes of the object is the trend of the past. Decision attribute of the object is the trend of the future.

4.1 Result of Knowledge Acquisition

Tables 4, 5 and 6 show decision rules acquired by using a rough set. C.I. is an abbreviation of the Covering Index showing the proportion of the target corresponding to that rule in the target with the same decision attribute [15]. The rule is reliable means that the C.I. value is higher. We have presented only the top 10 rules of the C.I. "+" in this table means that this period was increasing. "−" shows the descent, "0" indicates no change as well.

Table 4 shows the rule was a descent in this period. If the whole was descent, it shows that the descent could be high after 1 min. Even if the second half rise and fourth quarter was descent or no change, it shows that the value of the current period was descent.

Table 5 shows the rule was no change in this period. This table includes a lot of "0". If the stock price does not change much in 12 min, which indicates no change after 1 min. Also, "−" did not exist at all in knowledge granules of the data. If the increasing trend was found before 12 min, this period had often no change.

Table 6 shows the rule was an increase in this period. This table shows that the often descent to first and third period, also increase to second and fourth period (Fig. 3).

4.2 The Prediction Based on the Knowledge Acquired

In this section, we predicted the stock price on the basis of the knowledge acquired through the rough set. The stock to very short term fluctuations, we assume affected the behavior of the past. We used the rules obtained from the transaction data in the morning of that day, and predicted the price movements based on previous 12 min after the start of the afternoon trading. The regression line is obtained using the same data that were predicted. We predicted the price movements from the start of trading before the slope of the regression line became "rising" or "falling".

Table 4 Rules of decreasing cases

No.	Decision rule							C.I.
	Whole	Former	Latter	First quarter	Second quarter	Third quarter	Fourth quarter	
1	−	−		−			0	0.143
2	−		+				0	0.095
3	−		+				−	0.095
4		+		−			−	0.095
5		−		0		0		0.095
6	−		+	+				0.095
7			+	+	−			0.095
8	+			−			−	0.095
9		−		−		−	0	0.095
10	−	−			0		0	0.095

Table 5 Rules of no changing cases

No.	Decision rule							C.I.
	Whole	Former	Latter	First quarter	Second quarter	Third quarter	Fourth quarter	
1	+			0			0	0.217
2				0		+	0	0.117
3		+	+		+			0.117
4		+		0		0		0.117
5			+		+		0	0.100
6	+		0		0			0.100
7		+			0	0	0	0.100
8	+				0	0	0	0.100
9		+	0		0	0		0.100
10		+	0		0		0	0.100

Table 6 Rules of increasing cases

No.	Decision rule							C.I.
	Whole	Former	Latter	First quarter	Second quarter	Third quarter	Fourth quarter	
1		+		0		−		0.211
2		+		0			−	0.158
3			+	−	0			0.158
4			−	0	+			0.158
5				0	+	−		0.158
6				−		−	+	0.105
7				−	0		+	0.105
8		−			0		+	0.105
9	+			0		−		0.105
10	−			0	+			0.105

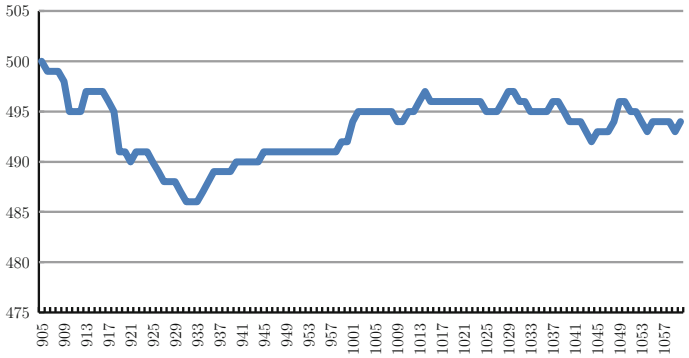


Fig. 3 One-minute chart of Fuji heavy industry

Table 7 The number of hitted rules in the afternoon

Time	Number of rules			Prediction
	Down	No change	Up	
12:30	0	3	3	Up
12:31	2	3	4	Up
12:32	1	5	3	No change
12:33	0	2	0	No change
12:34	0	8	2	No change
12:35	1	5	4	No change
12:36	0	2	6	Up
12:37	0	5	0	No change
12:38	1	4	2	No change
12:39	0	0	0	No change
12:40	4	0	0	Down
12:41	0	0	0	No change

Table 7 shows the number of rules that were used to predict the price movements of 12 min from the start of the afternoon trading. For example, this row of 12:30 shows that the number of hits to descending rule is zero, no change rule is three, increasing rule is three. We used as the predicted value the state often hits the most.

4.3 Prediction Result

We have predicted the stock price of Fuji Heavy Industries Ltd. by using the proposed method. The data used in the prediction are stock prices that were traded on June 2008. We have predicted the stock price by using a decision table in rough set. Objects in the decision table is a trend at each time. We have calculated the minimum decision rules from the decision table. Minimal decision rule is a knowledge of the prediction in specific conditions.

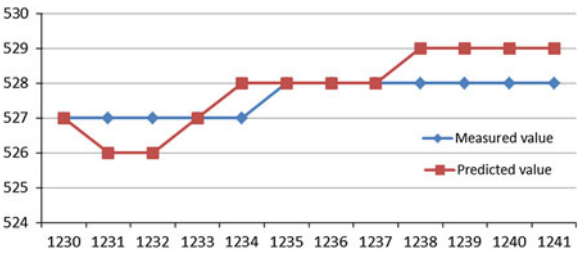


Fig. 4 The change of real and predicted values (1)

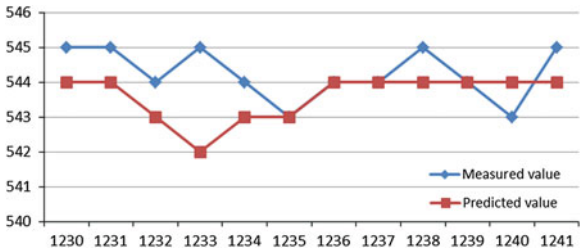


Fig. 5 The change of real and predicted values (2)

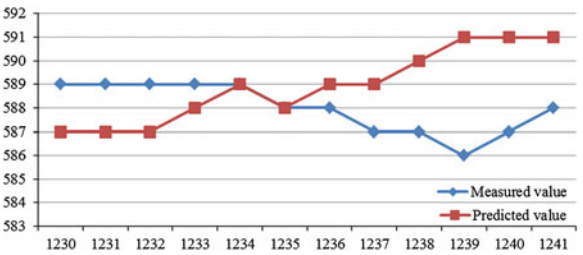


Fig. 6 The change of real and predicted values (3)

The knowledge of predictions obtained, it is to predict the stock price.

Figures 4, 5 and 6 show a part of the predicted results.

The measured values in Fig. 4 show a moderate upward trend. In addition, predicted values are all in the same general price movements. This figure shows that the prediction was correct.

Figure 5 shows that both values were vibrating and did not change much.

The measured values in Fig. 6 show a moderate downward trend. However, the predicted value was an upward trend. This figure shows that the prediction was not correct.

Table 8 is a summary of the prediction results of one month. The slope of the regression line is described as " + " for positive, described as " − " for negative. It is predicted using the knowledge granules obtained from trading in the morning. We predicted the 12 min immediately after the start of trading in the afternoon. Its value was predicted whether to fall or rise. The correct rate of predicted values in one month was 61.9 %.

Table 8 The rate of correct predictions in the month

Date	Measured value	Predicted value	Correct predictions
20080602	+	−	×
20080603	+	+	○
20080604	−	+	×
20080605	+	+	○
20080606	+	−	×
20080609	+	+	○
20080610	−	−	○
20080611	+	−	×
20080612	+	−	×
20080613	+	+	○
20080616	+	+	○
20080617	+	+	○
20080618	+	+	○
20080619	−	+	×
20080620	+	+	○
20080623	+	+	○
20080624	+	+	○
20080625	+	−	×
20080626	+	+	○
20080627	+	+	○
20080630	−	+	×
			61.9 %

Table 9 Toyota and Fuji heavy industry

Date	Comparison of the measured value	Date	Comparison of the measured value
20080602	○	20080617	○
20080603	○	20080618	○
20080604	○	20080619	×
20080605	×	20080620	×
20080606	×	20080623	○
20080609	○	20080624	○
20080610	○	20080625	○
20080611	○	20080626	○
20080612	○	20080627	○
20080613	○	20080630	○
20080616	○		81.0 %

Stock prices are affected by the stock price of other related stocks. Therefore, we propose to acquire knowledge granules from stock price data of other related stocks. Prediction was performed by using the movement of the relevant stocks. We used Daihatsu and Toyota as related stock prices. We acquired the knowledge of the prediction for stock prices of these two companies.

Table 9 shows the comparison of the measured value of Fuji Heavy Industries and Toyota Motor Corporation. Fuji Heavy Industries and Toyota Motor

Table 10 Daihatsu and Fuji heavy industry

Date	Comparison of the measured value	Date	Comparison of the measured value
20080602	○	20080617	○
20080603	○	20080618	○
20080604	○	20080619	×
20080605	○	20080620	×
20080606	○	20080623	○
20080609	×	20080624	×
20080610	○	20080625	○
20080611	○	20080626	×
20080612	○	20080627	○
20080613	○	20080630	○
20080616	○		76.2 %

Table 11 Prediction result

Date	Toyota	Daihatsu	Both companies
20080602	–	–	–
20080603	–	○	○
20080604	–	×	×
20080605	○	○	○
20080606	–	×	×
20080609	–	○	○
20080610	○	○	○
20080611	×	×	×
20080612	–	×	×
20080613	○	–	○
20080616	–	○	○
20080617	–	–	–
20080618	○	○	○
20080619	×	×	×
20080620	○	–	○
20080623	–	○	○
20080624	–	○	○
20080625	×	–	×
20080626	–	○	○
20080627	○	–	○
20080630	×	×	×
Correct predictions (%)	60.0	60.0	63.2

Corporation showed the same movement was 81.0 %. Table 10 shows the comparison of the measured value of Fuji Heavy Industries and Daihatsu Motor Corporation. Fuji Heavy Industries and Daihatsu Motor Corporation indicated the same movement was 76.2 %.

Table 11 shows the predicted results. Prediction accuracy when using Toyota Motor stock values as related data was 60.0 %, using Daihatsu Motor stock values as related data was 60.0 %, using the stock prices of both as relevant data was the best, 63.2 %.

5 Conclusion

In this paper, we investigated knowledge acquisition about the market fluctuations in stock markets by using the rough set theory.

The intra-day trading data are the record of all the trading transactions related to all the stocks. We converted 1 min chart data to the intra-day trading data and acquired the knowledge in order to predict future changes by using the regression line and rough set. We considered whether we can predict the fluctuation movements in the market using the knowledge acquired, and compared the actual movements of stock prices with the model forecasts. In addition, we predicted the price movement using the proposed model. It was predicted by using the data in June 2008. The correct rate of predicted values in one month was 63.2 %. Error rate of the predicted value will be 36.8 %. Profit margin obtained by subtracting the error rate from the correct answer rate is 26.4 %. This is a very high value. We were able to visualize as a rule knowledge. This is an advantage over other approaches. By using the rough set, we were able to improve these results.

References

1. Matsumoto, Y., Watada, J.: Improvement of chaotic short-term forecasting on fuzzy reasoning and tuning on genetic algorithm. Jpn. Soc. J. Fuzzy Theory Intell. Inf. **16**(1), 44–52 (2004)
2. Pawlak, Z.: Rough sets. Int. J. Comput. Inf. Sci. **11**(5), 341–356 (1982)
3. Pawlak, Z.: Rough Sets—Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers (1991)
4. Tan, S., Cheng, X., Xu, H.: An efficient global optimization approach for rough set based dimensionality reduction. Int. J. Innov. Comput., Info. Control **3**(3), 725–736 (2007)
5. Goh, C., Law, R.: Incorporation the rough sets theory. Chemometr. Intell. Lab. Syst. **47**(1), 1–16 (2003)
6. Azibi, R., Vanderpooten, D.: Construction of rule-based assignment models. Eur. J. Oper. Res. **138**(2), 274–293 (2002)
7. Beynon, M.J., Peel, M.J.: Variable precision rough set theory and data discrimination: an application to corporate failure prediction. Omega **29**(6), 561–576 (2001)
8. Li, R., Wang, Z.O.: Mining classification rules using rough set and neural networks. Eur. J. Oper. Res. **157**(2), 439–448 (2004)
9. Quafafou, M.: α-RST: a generalization of rough set theory. Inf. Sci. **124**(4), 301–316 (2000)
10. Greco, S., Matarazzo, B., Slowinski, R.: Rough sets theory for multi-criteria decision analysis. Eur. J. Oper. Res. **129**(1), 1–47 (2001)

11. Jhieh, Y., Tzeng, G., Wang, F.: Rough set theory in analyzing the attributes of combination values for insurance market. Expert Syst. Appl. **32**(1), 56–64 (2007)
12. Harada, T., Tanaka, R.: Analysis of Specifications for web screen-design using rough sets. J. Adv. Comput. Intell. Intell. Info. **10**(5), 688–694 (2006)
13. Kim, D., Bang, S.Y.: IRIS data classification using tolerant rough sets. J. Adv. Comput. Intell. Intell. Info. 4(5) (2000)
14. Walczak, B., Massart, D.L.: Rough set theory. Chemom. Intell. Lab. **47**(1), 1–16 (1999)
15. Predki, B., Slowinski, R., Stefanowski, R., Wilk, S.z.: ROSE-software implementation of the rough set theory. In: Polkowski, L., Skowron, A. (eds.) Rough Set and Current Trends in Computing. Lecture Notes in Artificial Intelligence, Springer, Berlin, pp. 605–608, (1998)
16. Predki, B., Wilk, S.z.: Rough set based data exploration using ROSE system. In: Ras, Z.W, Skowron, A. (eds.) Foundations of Intelligent Systems. Lecture Notes in Artificial Intelligence, Poland, Warsaw: Springer, pp. 172–180, (1999)
17. Pawlak, Z.: Rough classification. Int. J. Hum.-Comput. Stud. **51**(15), 369–383 (1999)
18. Gronhaug, K., Gilly, M.C.: A transaction cost approach to consumer dissatisfaction and complaint action. J. Econ. Psychol. **12**(1), 165–183 (1991)
19. Lin, C.,Watada, J., Tzeng, G.: Rough sets theory and its application to management engineering. In: Proceedings, international symposium of management engineering, Kitakyushu, Japan, pp. 170–176, (2008)
20. Tanaka, H Tsumoto, S.: Rough sets and expert system, Math. Sci., pp. 76–83 (1994)
21. Mori, N., Tanaka, H., Inoue, K.: Rough sets and Kansei: knowledge acquisition and reasoning from Kansei data, (2004)