Knowledge Discovery from the Data of Long Distance Travel Mode Choices Based on Rough Set Theory

[1]*Weijie Wang and* [2]Moon Namgung

*Department of Civil and Environmental Engineering, Wonkwang University,
Iksan City, Jeollabuk Do, 570-749, South Korea
[1]weijie@wonkwang.ac.kr, [2]ngmoon@wonkwang.ac.kr*

Abstract

The purpose of this study is to find the relationships between personal demographic attributes and long distance travel mode choices based on the Artificial Intelligence technique-rough set theory. Rough set theory can learn and refine decision rules or hidden facts from the incomplete observed data without the constraints of statistical assumptions. Also the induced decision rules are expressed in natural language, which can help policymakers in the decision making process. In the study, we conducted a survey to collect the peoples' most preferred travel mode choices for the given destination and people's demographic information. We analyzed the observed data based the rough set theory, calculated and discussed the approximation, core, reduct and rules of the data. The results of validation test were very promising, which showed that the induced decision rules could represent the relationships between data with the accuracy of 74.59%.

1. Introduction

Almost every person's activity is affected in some way by the quality of the transportation system. An accessible, affordable, reliable and efficient transportation system increases people's mobility providing opportunities for education, work shopping and traveling and other transportation needs. Travel mode choices are always an issue of travel research and many research results have been published [1-3].

In this study we focus on the long distance travel mode choices. As usual, the travel time and cost is relatively stationary for a long distance travel. Therefore, demographic data plays an integral part in the long distance transportation planning process, and policymakers and transportation planners are sure to face the problem of finding the relationships between demographic characteristics and travel mode choices. Previous researches [4-6] have primarily focused it on the impact of socioeconomic factors at the individual and household levels. Researches show that personal and household-level demographic characteristics have a significant impact on an individual's activity and travel choices.

All these analysis were based on the statistical analysis which suffers from some limitations, often due to the unrealistic assumption of statistical hypotheses or due to a confusing language of communication with the decision makers. In the last decades, many tools and techniques have been developed to extract patterns or relationships from empirical data in the domain of knowledge discovery. To overcome the statistical limitations, in the view of the need for formal relationships between personal demographic attributes and travel mode choices, this study investigates the feasibility of applying rough set theory from Artificial Intelligence into modeling travel mode choices and finding representative knowledge. The rough sets can extract useful information (identification and recognition of common patterns) from a knowledge system including quantitative and qualitative data, especially in the case of uncertain and incomplete data, and express this information through natural languages (decision rules) which are easy to understand [7]. To solve the NP-hard problem in logit model, Wong et al. [8] used decision rules of rough set analysis to reduce the sample space of interaction effects. And Wong and Chung [9] proved the feasibility to test the existence of accident patterns and their strength by the novel-parametric methodology-rough set analysis.

Having discussed the research background, Section 2 of this paper gives an overview of the rough set theory including indiscernibility of objects, information system, approximations of set, attributes reduction and decision rules induction. In Section 3, a real case of long distance travel mode choices is presented to show the process of finding relationships by rough set theory. Finally, we present our conclusions in Section 4.

2. Basic concepts of the rough set theory

Since rough set theory as a new approach to decision making in the presence of uncertainty and vagueness was introduced by Pawlak in 1982 [7, 10], it has attracted attention of researchers all over the world [11, 12]. During the last decades it has been applied in many different fields [13, 14] such as fault diagnosis, financial prediction, image processing, decision theory, etc. Until now many advantages of rough set theory application have been founded [10, 15], some of them are listed as follows:

1) It accepts both quantitative and qualitative attributes and specifies their relevance for approximation of the classification;

2) It discovers important facts hidden in data and expresses them in the natural language of decision rules;

3) It contributes to the minimization of the time and cost of the decision making process;

4) It is easy to understand and offers straightforward interpretation of obtained results;

5) It takes into account background knowledge of the decision maker;

6) It can be incorporated into an integrated DSS for the evaluation of corporate performance and viability.

2.1. Concept 1. Indiscernibility of objects

The rough set philosophy is founded on the assumption that with every object of the universe of discourse we associate some information (data, knowledge). Objects characterized by the same information are indiscernible in view of the available information about them. The indiscernibility relation generated in this way is the mathematical basis of rough set theory [15].

In rough set theory, given an information system S,

$\mathrm{S} = (\mathrm{U}, \mathrm{A}, \mathrm{V}, \mathrm{f})$

Where $\mathrm{U} = \{X_1, X_2, X_3, \cdots\}$, it represents the closed universe with a finite set of N objects in the system;

$\mathrm{A} = \{a_1, a_2, a_3, \cdots, a_n\}$, it is a finite set of n attributes including condition attributes and decision attribute;

V is the domain of attributes a;

f is the total decision function called the information function.

Every set of attributes A determines an equivalence relation on the universe U, denoted by $IND(A)$. $IND(A)$ divides the given universe U into a family of equivalence classes, called the elementary sets.

2.2. Concept 2. Information system

An information system in rough set theory is also called a knowledge representation system and can be intuitively expressed in terms of information table. In the table columns are labeled by attributes, rows are labeled by objects. Therefore, each row represents a piece of information about the corresponding object, and objects arranged in an information table are based their condition attributes and decision attribute.

2.3. Concept 3. Approximations of set

Given knowledge base $\mathrm{S} = (\mathrm{U}, \mathrm{A}, \mathrm{V}, \mathrm{f})$, with each subset $\mathrm{X} \subseteq \mathrm{U}$ and an equivalence relation $\mathrm{R} \in \mathrm{IND(S)}$, we associate two subsets:

$$\underline{A}R = \{x \in U : [x]_R \subseteq X\}$$

$$\overline{A}R = \{x \in U : [x]_R \cap X \neq \varnothing\}$$

called the R-lower and R-upper approximation of X. R-boundary of X is defined as follows:

$$BN_A(X) = \overline{A}R - \underline{A}R$$

Two measures are provided to examine the inexactness of approximate classifications. The first is defined as follows:

$$\alpha_A(X) = \frac{\sum card\,\underline{A}R_i}{\sum card\,\overline{A}R_i}$$

It is called the accuracy of classification, which expresses the percentage of possible correct decisions when classifying objects. If no boundary region exists, the ratio is equal to 1, which means a perfect classification. The second is defined as follows:

$$\lambda_A(X) = \frac{\sum card\,\underline{A}R_i}{card(U)}$$

It is called the quality of approximation, which expresses the percentage of objects correctly classified to classes.

2.4. Concept 4. Attributes reduction

Some attributes in an information system may be superfluous and dispensable, thus can be eliminated without losing essential classification information. If $IND(A) = IND(A - a_i)$, then the attribute a_i is called superfluous. Otherwise, the attribute a_i is indispensable in A. The attributes reduction is the process of removing of superfluous partitions (equivalence relations) and finding only that part of the really useful knowledge. Intuitively, a reduct of knowledge is its essential part, which suffices to define knowledge, and the elimination of any attribute in it does lead to a less accurate classification. The core is the interaction of all reducts and is the set of the most characteristic and important part of knowledge (information table).

2.5. Concept 5. Decision rules induction

Decision rules induction requires the partitioning of the attributes into condition and decision attributes. Given an information table (decision table), we can find all minimal decision algorithms associated with the table, which is the most important application of rough set theory. After computing the core and getting the reducts of attributes, we get a reducted information table. The decision rules could be found through determining the decision attributes value based on condition attributes values. A decision rule may either be exact or approximate, and it is be described as conditional statement that is expressed in the terms of “IF condition(s) THEN decision(s)”. The quality of the decision rule is indicated by its strength. The strength of a rule represents the number of observations or cases that accordance with that rule.

3. Application of the rough set theory

3.1. The data

The study area of this study is Iksan city, located in Jeollabuk-do of South Korea with the population of about 300,000. One of main destination for the Iksan citizens is Seoul city, the capital city, which is about 208 km away from Iksan city.

The data in this study was collected by personal interview surveys in October, 2004. The purpose of the survey was to investigate the citizens’ travel mode choices for a long distance travel from Iksan to Seoul. In the survey, the participants were asked about their most preferred mode choices, as well as their demographic characteristics such as their sex, age, occupation etc. KTX as the express train service is the abbreviation for Korea Train eXpress and was started in April 2004. Though KTX is one kind of train service, it is separately considered because it is a new mode and different from the common train services. Meanwhile, identifying the characteristics of KTX passengers is also one of our purposes.

The general demographic information provided in the survey included qualitative and quantitative attributes. All attributes were divided into condition and decision attributes, and to apply rough set theory, they were transformed into categorical information, which was shown in Table 1.

A total of 400 individual surveys were collected within the area. Of those, approximately 8.5 percent was erroneous, leaving 366 valid forms for analysis. The results of the surveys (Figure 1) showed that a half of the participants preferred train, 23% of them preferred bus, 15% of them preferred KTX and 12% of them preferred car, indicating large distribution difference among provided travel mode choices.

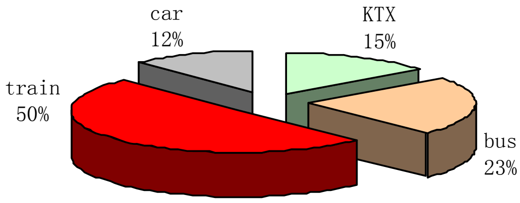

Figure 1. Distribution of preferred travel mode choices

Table 1. Categorisation of the attributes

No.	Condition attributes							Decision attributes
	a1	a2	a3	a4	a5*	a6	a7	D
	Sex	Age	Education level	Occupation	Household income	Car ownership	Travel purpose	Travel mode choices
1	male	< 20	under high school	student	< 100	yes	work	KTX
2	female	20~30	high school	technician	100~200	no	business	bus
3		30~40	college student	management	200~300		returning home	train
4		40~50	College graduation	businessman	300~400		school	car
5		50~60	postgraduate	Service worker	400~500		tourism	
6		> 60		farmer	> 500		shopping	
7				producer			visiting relatives	
8				unemployed			others	

Note: * The unit is million won (Korea).

3.2. Presentation of the rough set results

In this study, Rough Set Data Explorer (ROSE) system [16] was used to analyze the coded information in Section 3.2. ROSE is a modular software calculating basic elements of the rough set theory and discovering useful rules[17]. The software has been developed at the Laboratory of Intelligent Decision Support Systems of the Institute of Computing Science in Poznan.

The results of the operation are as follows:

(1) Approximation Table 2 describes the approximation classification results of our survey data. The accuracy of approximation classification is used to express the completeness of knowledge about decision attribute (travel mode choices) that could be obtained based on the condition attributes (personal demographic attributes). Among the accuracies of classifications for 4 kinds of travel mode choices, Class train shows us the highest accuracy value of 0.67, and Class car, bus, KTX show the accuracy values of 0.62, 0.51, 0.51, respectively. The quality of classification is 0.76, indicating that 76% of cases in the collected survey data are correctly classified.

Table 2. Approximation of the survey data

Class	Total No. of objects	No. of objects in lower approximations	No. of objects in upper approximations	Accuracy of classification	Quality of classification
KTX	54	41	80	0.51	0.76
Bus	83	57	111	0.51	
Train	185	145	216	0.67	
Car	44	34	55	0.62	

(2) Core and reducts In rough set analysis, the core contains the attributes that are the most important in the information table. In this study all seven condition attributes are included in the core. This indicates that any attribute is necessary for perfect approximation of the decision classes and removal of any of them leads to the decrease of the quality of approximation.

Reducts, also called minimal sets, contain no redundant information. In this study, 11 reducts, shown in Table 3, were obtained for the coded information table. The length of reducts is 2~5 and is smaller than that of the original set. They represent a reduced set of attributes that provide the same classification quality of the decision attribute as the original set of condition attributes.

Table 3. Reducts for the coded information table

No.	Reduct
1	{education, purpose}
2	{sex, age, income}
3	{age, income, purpose}
4	{age, education, income}
5	{age, occupation, purpose}
6	{occupation, income, car ownership, purpose}
7	{sex, income, car ownership, purpose}
8	{sex, age, occupation, car ownership}
9	{sex, income, car ownership, purpose}
10	{sex, age, education, income, purpose}
11	{sex, age, income, car ownership, purpose}

(3) Decision rules induction We used LEM2 methodology in ROSE system to reduce the information table and got 89 rules including 28 approximate rules and 61. Rough set analysis provides computation intelligence to the problem of classification. As well it clearly shows the strength or weakness of the classification through its relative strength measure. As we can see from the relative strength of the rules some are stronger than others. If the relative strength of the rule is too small, it could not be used to predict to get good results. Therefore, we threw away the rules with relatively low strength values and just took the top few greatest rules into consideration.

Table 4. Induced decision rules

No.	Rules	Relative strength (%)
1	IF (age=20~30, 40~50) AND (income=400~500 million won) AND (travel purpose=business) THEN (travel mode choice=KTX)	9.26
2	IF (occupation=management) AND (income=100~300 million won) AND (car ownership=2) AND (purpose=commuting, visiting relatives) THEN (travel mode choice=KTX)	7.41
3	IF (sex=female) AND (age=50~60) AND (income=over 200 million won) THEN (travel mode choice=bus)	10.84
4	IF (sex=female) AND (education=college student) AND (income=300~500 million won) AND (car ownership=2) THEN (travel mode choice=train)	6.49
5	IF (education=postgraduate) AND (purpose=tourism, shopping) THEN (travel mode choice=car)	9.09
6	IF (age=20~30) AND (education=college student) AND (income=200~400 million won) THEN (travel mode choice=bus or train)	22.86
7	IF (sex=male) AND (income=2) AND (car ownership=1) AND (purpose=school) THEN (travel mode choice=bus or car)	42.86
8	IF (sex=female) AND (age=30~40) AND (occupation=business) AND (car ownership=1) AND (purpose=commuting) THEN (travel mode choice=bus or car)	28.57
9	IF (sex=male) AND (income=under 100 million won) AND (car ownership=1) AND (purpose=returning home) THEN (travel mode choice=bus or car)	28.57

In this study, the selected rules, listed in Table 4, are used to model the relationships between personal demographic attributes and long distance travel mode choices. We can see that the approximate rules were induced, which represented classifications of vehicle choices (bus or car) or public transportation choices (bus or train). The approximate rules 7, 8, and 9 of vehicle choices show the great relative strengths of 42.87%, 28.57%, and 28.57%, respectively. The approximate rule 6 of public transportation choices shows the great relative strength of 22.86%. The 4 approximate rules give us good classifications of vehicle choices and public transportation choices. While approximate rules are not all helpful, those with 3 or more classifications gain us nothing though they show great relative strengths.

Compared with the approximate rules, the single travel mode choices with low relative strength don't have good classifications. Though they provide relatively weak rules, these relationships can be further analyzed by policymakers to be used in the decision making process. However, this might suggest that the seven condition attributes may be insufficient in explaining the long distance travel mode choices and the not high accuracy of classification and quality of classification in the analysis of approximation can also be one proof. In the further study additional personal or household-level demographic characteristics should be considered.

(4) Validation A confusion matrix [18] is a visualization tool typically used in supervised learning. And it contains information about actual and predicted classifications done by a classification system. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. We can be easy to see if the system is confusing two classes through a confusion matrix. ROSE can generate confusion matrix to validate the decision rules. The validation result in Table 5 shows the total accuracy of 74.59%, which indicates that the rules performed well in representing the relationships between personal demographic attributes and travel mode choices. We also can see that the hit rate is quite good for the train choice but quite bad for the car choice. One reason for the bad hit rate may be the large distribution difference among the travel mode choices, which leads to the weak decisions generated. This suggests that to improve the prediction accuracy, we can use enough or larger data without large distribution difference among decision class at the start to generate better rules.

Table 5. Validation results

	KTX	Bus	Train	Car	Non e	Hit rate (%)
KTX	35	2	8	7	2	64.81
Bus	7	57	8	3	8	68.67
Train	6	8	159	9	3	85.95
Car	5	4	9	22	4	50.00
Total accuracy (%)	74.59					

4. Conclusions

In this study, rough set theory has been applied to find the relationships between personal demographic attributes and long distance travel mode choices in South Korea. The rough set analysis need not many conventional statistical assumptions like statistical analysis, but it can generate decision rules in the form of "if-then" statements based on classifications. The induced rules, representing the relationships between data, are intuitive and straightforward to comprehend and help policymakers during the decision making process. Applying ROSE system, we generated 11 reducts and 9 decision rules with greatest relative strength. The results of analysis are promising and are partially described the classification pattern successfully, which showed us the useful decision rules and promising validation accuracy of 74.59%.

However, we also can see that the rule strength and validation accuracy in this study need to be improved so that we can obtain more useful and reliable decision rules. Therefore, we should take some measures in the future study, such as 1) collecting enough data without large distribution difference among decision classes; 2) taking into account other strong attributes to explain decision attributes.

References

[1] Jörgen Garvill, Agneta Marell & Annika Nordlund, "Effects of increased awareness on choice of travel mode", Transportation, Volume 30, Number 1, pp. 63-79(17); February 2003.

[2] Tommy Gärling and Kay W. Axhausen, "Introduction: Habitual travel choice", Transportation, Volume 30, Number 1, pp. 1-11, February 2003.

[3] Sebastian Bamberg, Icek Ajzen and Peter Schmidt, "Choice of Travel Mode in the Theory of Planned Behavior: The Roles of Past Behavior, Habit, and Reasoned Action", Basic and Applied Social Psychology, Vol. 25, No. 3, Pages 175-187, 2003,

[4] S Algers, "Integrated structure of long-distance travel behavior models in Sweden", Transportation Research Record 1413, Transportation Research Board, Washington DC, pp. 141-149, 1993.

[5] N. Georggi, R. Pendyala, "Analysis of long-distance travel behavior of the elderly and low income", Transportation Research Circular, Transportation Research Board, Washington DC, pp.121-150, 2001.

[6] W.J. Mallett, "Long-distance travel by low-income households", Transportation Research Circular, Transportation Research Board, Washington DC, pp.169-177, 2001.

[7] Z. Pawlak, ROUGH SETS: Theoretical Aspects of Reasoning about Data, Kluwer Academic, Dordrecht/Boston/London, 1991.

[8] Chorng-Shyong Ong, Jih-Jeng Huang and Gwo-Hshiung Tzeng, "Using Rough Set Theory for Detecting the Interaction Terms in a Generalized Logit Model" Lecture Notes in Computer Science, Springer, 624-629, 2004.

[9] J.T. Wong, Y.S. Chung, "Using rough sets to explore the nature of occurrence of accidents", Transportation Research Board Annual Meeting CD-ROM, Transportation Research Board, Washington DC, pp. 1597-1616, 2006.

[10] B. Walczak and D.L. Masart, "Tutorial: rough sets theory", Chemometrics and Intelligent Laboratory Systems, Elsevier, Nethersland, pp. 1-16, 1999.

[11] J.W. Guan and D.A. Bell, "Rough computational methods for information system", Artificial Intelligences 105, pp.77-103, 1998.

[12] E.Burke and G.Kendall, Introductory Tutorials on Optimization: Search and Decision Support Methodologies, Springer-Verlang, New York, 2005.

[13] T.Y. Lin and N. Cercone, ROUGH SETS AND DATA MINING: Analysis for Imprecise Data, Kluwer Academic, Boston/London/Dordrecht, 1997.

[14] R. Slowinski, INTELLIGENT DECISION SUPPORT: Handbook of Applications and Advances of the Rough Sets Theory, Boston/London/Dordrecht, 1992.

[15] Z. Pawlak, "Rough set approach to knowledge-based decision support", European Journal of Operational Research, Elsevier, Nethersland, pp. 48-57, 1997.

[16] Predki, Slowinski, Stefanowski, Susmaga, Wilk, http://www-idss.cs.put.poznan.pl

[17] Bartlomiej Predki and Szymon Wilk, "Rough set based data exploration using ROSE system", Lecture Notes In Computer Science; Vol. 1609, Proceedings of the 11th International Symposium on Foundations of Intelligent Systems, Pages: 172 – 180, 1999.

[18] F. Provost and R. Kohavi, "Guest Editors' Introduction: On Applied Research in Machine Learning", Machine Learning, Springer, Netherlands, pp. 127-132, 1998.

Authors

Weijie Wang

A Ph.D. candidate in civil and environmental engineering at Wonkwang University, Korea. He received a B.S. degree in mechanical engineering and marketing from Jingdezhen Ceramic Institute, China P.R., 2002, and a M.S. degree in civil and environmental engineering from Wonkwang University, Korea, 2006.

Moon Namgung

A professor in the Department of Civil and Environmental Engineering since 1992. He received a B.S. degree in civil engineering from Wonkwang University, Korea in 1984, a M.S. degree in civil engineering from Chonbuk National University, Korea in 1986 and a Ph.D. degree in transportation engineering from Hiroshima University, Japan in 1992.