# Knowledge Acquisition in Modal Choice Modelling with Decision Tree Algorithms

**Takamasa Akiyama and Masashi Okushima**
**Gifu University**

## 1. Introduction

The decision tree is regarded as an inductive technique of machine learning. The study aims at proposing the knowledge base estimation for travel behaviour analysis with learning process. The fundamental decision three algorithms such as ID3 and C4.5 are introduced to produce the rule bases to describe the knowledge in the decision of trip makers. Furthermore, fuzzy ID3 and fuzzy C4.5 are similarly applied as extended algorithms. The advantages of the decision tree algorithms can be summarized through the analysis of modal choice for commuters in urban area with empirical data.

## 2. The theoretical background of decision tree

The decision tree consists of nodes and links to specify the individuals to the classes. The node indicates attribute of classification. The links are determined to divide the clusters according to values of attribute. The effective algorithms to produce the decision tree such as ID3 and C4.5 have been proposed by Quinlan[1]. It is proposed that the decision tree would be designed to minimize the expected number of tests in classification of the data. The algorithm of ID3 is summarized as follows:

The probability of occurrence for class $C_k$ is determined as $P_k$. The entropy of information is assumed as:

$$M(D) = -\sum_{j=1}^{m}(P_k \log_2 P_k), \quad P_k = |D_k| / |D| \tag{1}$$

Since the probability $P_{ij}$ is defined to the attribute $a_{ij}$, the conditional entropy can be determined similarly as:

$$B(A_i, D) = \sum_{j=1}^{m}(P_{ij} M(D_{ij})), \quad P_{ij} = |D_{ij}| / |D| \tag{2}$$

The gain of information for the conditional attribute is

$$G(A_i, D) = M(D) - B(A_i, D) \tag{3}$$

The set of attributes are selected through ID3 algorithm as test nodes of the decision tree corresponding to the maximum gain. Similarly, the following gain ratio would be applied as another criterion in C4.5 algorithm:

$$GR(A_i, D) = G(A_i, D) / S(A_i, D) \tag{4}$$

where $S(A_i, D) = -\sum_{j=1}^{m}(P_{ij} \log_2 P_{ij})$

Therefore, ID3 and C4.5 are introduced as primitive methods. Furthermore, fuzzy ID3 and fuzzy C4.5 have been proposed as extended models with fuzzy number[2].

## 3. The Modal Choice Model

The decision process of trip maker might be described properly with the proposed methods.

The modal choice model can be formulated with the decision tree methods. The dataset is obtained from the PT survey in 1991. The objective area and zones are illustrated in Figure 1. The zone 1 and 2 correspond to the central area of Gifu city. It is observed that 1,183 commuters make trips from the surrounding zone to the central two zones. In the estimation, 215 samples are

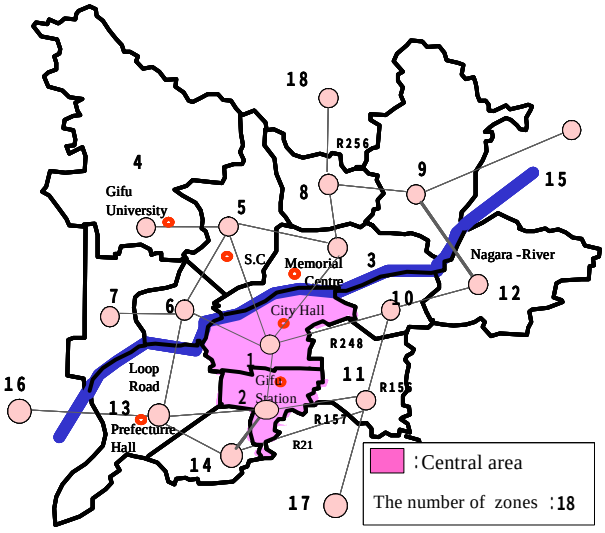


**Fig. 1 The objective area and zones**

randomly preserved as training dataset to determine the decision trees. The eight attributes are selected to estimate the mode of individual commuting as follows:

(1)NMH: The number of members in the household, (2)NCH: The number of owner cars in the household, (3)SEX: Male or female, (4)DLO: Driving license

ownership, (5)AGE: Age of the trip maker, (6)DTM: Departure time of the trip, (7) TTM: Travel time for the mode, (8) TCM: Travel cost for the mode. Three modes are assumed in estimation such as car, mass transit, the others (walk, bicycle and motorbike).

The decision trees are determined according to the algorithm. The upper level of the decision tree in ID3 and fuzzy ID3 are shown in Figure 2 and 3 respectively.. Even though the overall structures are rather different, the important elements of mode classification are commonly appeared such as travel time for car, licence ownership and size of household and etc. Similar tendency can be observed in the results of C4.5 and fuzzy C4.5 as well. The inference rules can be derived from the shape of decision trees. The inference rules for this example are counted as 105 for ID3, 121 for C4.5, 111 for fuzzy ID3, and 119 for fuzzy C4.5 respectively as results of knowledge acquisition by inductive learning. It might be true that many different decision processes exist in the modal choice of trip makers.

## 4. Discussion for the applications

The result of estimation is summarized in Table 1. The multinominal logit model is estimated for comparison with the same database. It is known that equivalent estimations with small errors might be performed in four different methods. The sample cases with error in ID3 and C4.5 are slightly different among 9 individuals. On the other hand, 8 individual error samples are commonly observed in fuzzy ID3 and fuzzy C4.5.

Furthermore, the actual share of modes (car, mass transit, others) in the area is observed as (50.7%, 18.1%, 31.2%). Each method gives quite similar estimations. Therefore, it is known that the decision tree provides high accuracy in terms of aggregate estimation as well.

The determined decision tree is applied to all samples with hold out dataset as well. The estimation for 1,183 should be carried out. As a results, the ratio of fitness for each algorithm remains over 60%. Particularly, fuzzy C4.5 provides fitness of 67.7% only with originally acquired knowledge as a structure of decision tree. The result demonstrates the applicability of machine learning techniques in modelling of travel behaviour analysis.

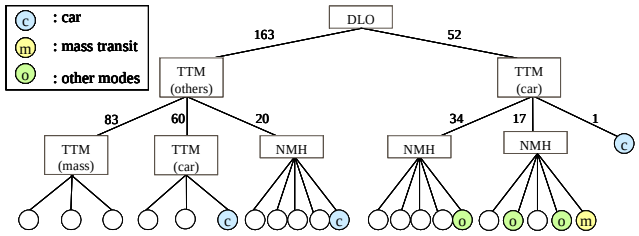


**Fig. 2 Upper levels in decision tree (ID3)**

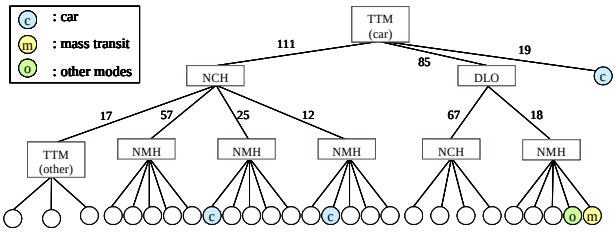


**Fig. 3 Upper levels in decision tree (fuzzy ID3)**

**Table 1 Summary of estimations for the methods**

| Method | Fitness ratio (errors) | Estimation of share |
|---|---|---|
| ID3 | 95.81 (9/215) | (54.0, 16.3, 29.8) |
| C4.5 | 95.81 (9/215) | (48.4, 19.1, 32.6) |
| Fuzzy ID3 | 96.28 (8/215) | (52.6, 17.2, 30.2) |
| Fuzzy C4.5 | 96.28 (8/215) | (52.6, 17.2, 30.2) |
| Logit | 67.40 (70/215) | (61.9, 6.0, 32.1) |

## 5. Concluding remarks

The decision tree methods are proposed to describe the knowledge of modal choice. The results of the study are summarized: (1)The decision tree methods might provide the high performance in estimation. (2) The knowledge in modal choice can be formulated in the shape of tree automatically through the algorithm. Some further studies are recommended: (1) Advantages of fuzzy versions of decision tree may not be quite obvious in the study. Another specific example with human perception would be discussed. (2) The problem as over fitting and pruning should be investigated to reduce the complexity in the structure of decision tree.

**Acknowledgement:**

The authors would like to express the acknowledgement to Mr. M, Toyoda, Kinki Regional Development Bureau, Ministry of Land, Infrastructure and Transport for helping the research.

**References:**

1) Quinlan, J. R.: C4.5: Programs for Machine Learning, Morgan Kaufmann Series in Machine Learning, 1993.

2) Hori, K., Umano, M. et. Al: Fuzzy C4.5 for Generating Fuzzy Decision Trees and Its Improvement, Proc. 15th fuzzy system Symposium, pp.515-518, 1999.