Aiming for Parsimony in the Sequential Analysis of Activity-Diary Data

Elke Moons and Geert Wets
University of Hasselt
Transportation Research Institute
Wetenschapspark 5 bus 6
3590 Diepenbeek
Belgium
Tel. +32-11-26.91.26 – +32-11-26.91.58
Email: elke.moons@uhasselt.be – geert.wets@uhasselt.be

Abstract

This paper aims at a better understanding in the impact of simplification in a sequential analysis of activity-diary data using a feature selection method. To this effect, the predictive performance of the Albatross model, which incorporates nine different facets of activity-travel behaviour, based on the original full decision trees is compared with the performance of the model based on trimmed decision trees. The more parsimonious models are derived by first using a feature selection method to determine the irrelevant variables which are then left out of the further model building process. The results indicate that significantly smaller decision trees can be used for modelling the different choice facets of the sequential system without loosing much too much in predictive power. The performance of the models is compared at two levels: the choice facet level, at which we compare the performance of each facet separately and the trip level, comparing the correlation coefficients that determine the strength of the associations between the observed and the predicted origin-destination matrices. The results indicate that the model based on the trimmed decision trees predicts activity diary schedules with a minimum loss of accuracy at the choice facet level. Moreover, the results show a slightly better performance at the trip matrix level.

Keywords: activity-travel behaviour, parsimony, sequential analysis, feature selection, decision trees

1. Introduction

In the past few years, activity-based forecasting of travel demand has become a major field of interest in transportation research. The aim of activity-based models is to predict which activities will be conducted where, when, for how long, with whom and with which transport mode. Rule-based models have proven to be very flexible when compared to utility-maximising models (Arentze *et al.*, 2001) and they also perform well in predicting transport choice behaviour if an induction technique is used (Wets *et al.*, 2000). Although these rule-based models perform very well, they also show some limitations. Most of them are based on quite complex rule sets. However, already in the Middle Ages, there was a call for simplicity: William of Occam's razor states that `Nunquam ponenda est pluralitas sin necesitate', meaning `Entities should not be multiplied beyond necessity' (Tornay, 1938). It was born in the Middle Ages as a criticism of scholastic philosophy, whose theories grew ever more elaborate without any corresponding improvement in predictive power. In the intervening centuries it has come to be seen as one of the fundamental tenets of modern science and today it is often invoked by learning theorists as a justification for preferring simpler models over more complex ones. However, Domingos (1998) learned us that it is tricky to interpret Occam's razor in the right way. The interpretation "Simplicity is a goal in itself" is essentially correct, while "Simplicity leads to greater accuracy" is not.

While a larger number of rules may be valuable when one wishes to better understand the data, from a predictive perspective a large number of rules may imply that the decision tree induction algorithm has over-fitted the data. The obtained decision tree structure may then be very unstable and sensitive to highly correlated covariates.

Feature selection offers a solution to reduce the number of irrelevant attributes and as a consequence often the size of the decision tree will also be reduced. The key notion underlying feature selection is that the number of decision rules is reduced by selecting and deleting irrelevant features, based on some statistical measure. The impact of feature selection on the predictive performance of rule-based models is however not a priori clear. On the one hand, because the irrelevant variables are deleted, feature selection may not have a substantial negative effect on predictive performance. However, a smaller decision tree may also result in a higher probability of misclassification, leading to worse predictive performance. It is against this background that this paper reports the findings of a methodological study that was conducted to gain a better understanding of the influence of a smaller set of decision rules on the predictive performance of a sequential models of activity scheduling behaviour, the Albatross model. Moons *et al.* (2002) investigated the influence of irrelevant attributes on the performance of the decision tree for the transport mode, the travel party, the activity duration and the location agent of the Albatross model system. We found that the use of considerable less decision rules did not result in a significant drop in predictive performance compared to the original larger set of rules that was derived from the activity-travel diaries. In this paper, the question `To what extent can this result be generalised to the complete Albatross model system (represented by nine different choice facets)?' is inspected.

In order to be able to look at the results in the right context, we will first shortly describe the Albatross system in the next section, followed by a brief introduction to the different methods used to perform the analysis. Then, feature selection is applied to decision rule induction and the results will be discussed in Section 3. The predictive performance will be evaluated on each facet separately (by means of the accuracy) and at trip matrix level where the correlation coefficients that determine the strength of the associations between the observed and predicted origin-destination matrices are judged against each other. Conclusions are drawn in the final section.

2. Methods

2.1 The Albatross system

The Albatross model was developed for the Dutch Ministry of Transportation (Arentze and Timmermans, 2000). In this study, we used the data that were used to find the set of rules for the original model.

This rule-based model relies on a set of Boolean decision rules that are used to predict activity-travel patterns. These rules were extracted from activity-diary data. The activity scheduling process is sequential in nature. Figure 1 provides a schematic representation of the Albatross scheduling model.

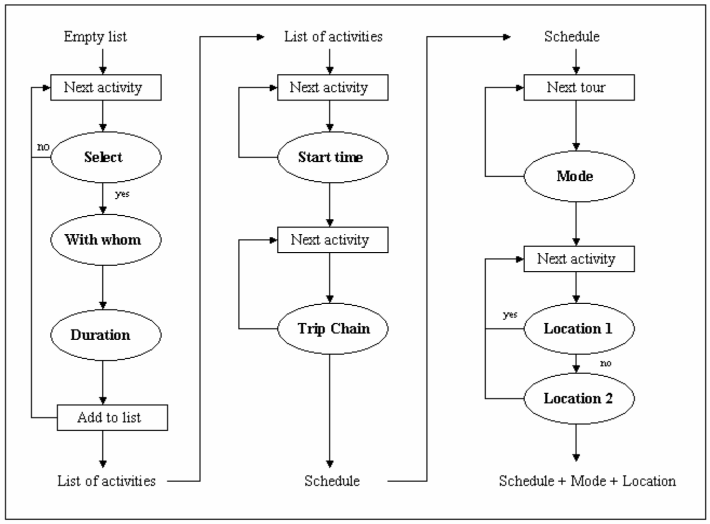

Figure 1: Albatross scheduling engine

The activity scheduling agent of Albatross is based on an assumed sequential execution of decision trees to predict activity-travel patterns. Before the sequential execution starts, the main transport mode (i.e. mode for work, referred to as mode 1) will be predicted. The model next executes a set of decision rules to predict which activity will be inserted in the schedule. It then determines, based on another sets of rules, with whom the activity is conducted and the duration of the activity. The order in which activities are evaluated is pre-defined as: daily shopping, services, non-daily shopping, social and leisure activities. The assignment of a scheduling position to each selected activity is the result of the next two steps. After a start time interval is selected for an activity, trip-chaining decisions determine for each activity whether the activity is to be connected with a previous and/or next activity. Those trip chaining decisions are not only important for timing activities but also for organizing trips into tours. The next step involves the choice of transport mode for other purposes (referred to as mode2) and the choice of location. Possible interactions between mode and location choices are taken into account by using location information as conditions of mode selection rules.

2.2 Decision Tree Induction: C4.5

Decision tree induction can be best understood as being similar to parameter estimation methods in econometric models. The goal of tree induction is to find the set of Boolean rules that best represents the empirical data. The original Albatross system was derived using a Chi-square based approach (Moons, 2005). In this paper, however, the trees were re-induced using the C4.5 method (Quinlan, 1993) because this method can be easily combined with the Relief-F feature selection. Arentze *et al.* (2000) found approximately equal performance in terms of goodness-of-fit of the two methods in a representative case study. The C4.5 algorithm works as follows. Given a set of I observations taken from activity-travel diary data, consider their values on n different explanatory variables or attributes $x_{i1}, x_{i2}, \ldots, x_{in}$ and on the response variable $y_i \in \{1,2,\ldots,p\}$ for $i = 1,\ldots, I$. Starting from the root node, each node will be split subsequently into internal or terminal nodes. A leaf node is terminal when it has no offspring nodes. An internal node is split by considering all allowable splits for all variables and the best split is the one with the most homogeneous daughter nodes. The C4.5 algorithm recursively splits the sample space on X into increasingly homogeneous partitions in terms of the response variable Y, until the leaf nodes contain only cases from a single response class. Increase in homogeneity achieved by a candidate split is measured in terms of an information gain ratio. After building the tree, pruning strategies are adopted. This means that the decision tree is simplified by discarding one or more sub-branches and replacing them with leaves. For a detailed description, we refer to Wets *et al.*, 2000.

2.3 Feature Selection: Relief-F

Feature or variable selection strategies are often implied to explore the effect of irrelevant attributes on the performance of classifier systems. One can distinguish between two types of feature selection approaches: the filter and the wrapper approach. Both methods have been compared extensively (Hall, 1999a, 1999b; Koller and Sahami, 1996). In this analysis, the filter approach, more specifically the Relief-F feature selection method, is opted for since it can handle multiple classes of the dependent variable (the nine different choice facets that we are predicting range from two to seven classes) and above that it is easily combined with the C4.5 induction algorithm.

Feature selection strategies can be regarded as one way of coping with the correlation between the attributes. This is relevant because the structure of trees is sensitive to the problem of multi-collinearity, which implies that some variables would be redundant (given the presence of other variables). Redundant variables do not affect the impacts of the remaining variables in the tree model, but it would simply be better if they were not used for splitting. Therefore, a good feature selection method for this analysis would search for a subset of relevant features that are highly correlated with the class variable that the tree-induction algorithm is trying to predict, while mutually having the lowest possible correlations.

Relief (Kira and Rendall, 1992), the predecessor of Relief-F, is a distance-based feature weighting algorithm. It imposes a ranking on features by assigning each a weight. Features with the highest weights are considered to be the most relevant, while those with values close to zero or with negative values are judged irrelevant. The weight for a particular feature reflects its relevance in distinguishing the classes. In determining the weights, the concepts of *near-hit* and *near-miss* are central. A *near-hit* of instance *i* is defined as the instance that is closest to *i* (based on Euclidean distance between two instances in the *n*-dimensional variable space) and which is of the same class (concerning the output variable), while a *near-miss* of *i* is defined as the instance that is closest to *i* and which is of a different class. The algorithm initially assigns the value zero to each attribute, and this will be adapted with each run through the instances of the data set. It attempts to approximate the following difference of probabilities for the weight of a feature *X*:

$$\begin{aligned} W_X = \; & P(\text{different value of X} \mid \text{nearest instance of different class}) \\ - \; & P(\text{different value of X} \mid \text{nearest instance of same class}). \end{aligned}$$

So, Relief works by random sampling an instance and locating its nearest neighbour from the same and opposite response class. By removing the context sensitivity provided by the "nearest instance" condition, attributes are treated as mutually independent, and the previous equation becomes:

$$\begin{aligned} \text{Relief}_X = \; & P(\text{different value of X} \mid \text{different class}) \\ - \; & P(\text{different value of X} \mid \text{same class}). \end{aligned}$$

Relief-F (Kononenko, 1994) is an extension of Relief that can handle multiple classes and noise caused by missing values, outliers, etc. To increase the reliability of Relief's weight estimation, Relief-F finds the *k* nearest hits and misses for a given instance, where *k* is a parameter that can be specified by the user. For multiple class problems, Relief-F searches for nearest misses from each different class (with respect to the given instance) and averages their contribution. The average is weighted by the prior probability of each class.

3. Analysis and Results

The overall aim of this study is to investigate whether a simplification of the rule sets underlying the Albatross model leads to a significant loss in predictive power. This simplification will be obtained by reducing the set of decision rules through the application of a feature selection method. The original Albatross model consists of nine choice facets. For each of these choice facets, a set of decision rules was extracted from activity-travel diaries. To predict activity-travel patterns, these decision trees are executed sequentially in the Albatross system according to some scheduling process model (Arentze and Timmermans, 2000). We will investigate the effect of simpler rules for each choice facet.

3.1 The Data

The analyses are based on the activity diary data used to derive the original Albatross system. The data were collected in February 1997 for a random sample of 1649 respondents in the municipalities of Hendrik-Ido-Ambacht and Zwijndrecht (South Rotterdam region) in the Netherlands.

The activity diary asked respondents, for each successive activity, to provide information about the nature of the activity, the day, start and end time, the location where the activity took place, the transport mode (chain) and the travel time per mode, if relevant, accompanying individuals, and whether the activity was planned. Open time intervals were used to report the start and end times of activities. A pre-coded scheme was used for activity reporting. More details can be found in Arentze and Timmermans (2000).

3.2 Study Design

The original data set is split into two subsets. A training set, containing the first 75% of the cases, on which the different models will be built and optimised. The remaining 25% of the cases make up the validation or test set that can be used to compute the accuracies (percentage of correctly classified instances), etc. These percentages are arbitrary but are common practice in validation studies (see e.g. Wets *et al.*, 2000).

We will first build decision trees for each of the nine choice facets, using the C4.5 algorithm (Quinlan, 1993). This approach will be called the full approach. The C4.5 trees were induced based on one simple restriction: the final number of cases in a leaf node must meet a minimum. For eight out of the nine choice facets, this minimum was set to 15 (except for the very large data set of the `select'-dimension, where this number was set to 30). In a second approach, the feature selection approach, we will first identify the relevant attributes for each of the nine choice facets separately, based on the Relief-F feature selection method with the *k* parameter set equal to 10. Next, the C4.5 trees were built based on the same restriction as in the full approach, though only the remaining relevant attributes were used. To determine the selection of variables, the following procedure was adopted. Several decision trees were built, each time removing one more irrelevant attribute, as they appeared lowest in the ranking that has been provided by the FS method. For each of these decision trees, the accuracy was calculated and compared to the accuracy of the decision tree of the full approach. The smallest decision tree, which resulted in a maximum decrease of 2% in accuracy compared to the decision tree including all features, was chosen as the final model for a single choice facet in the feature selection approach. This strategy was applied to all nine dimensions of the Albatross model.

3.3 Results

At first, we will take a closer look at the average length of the observed and predicted sequences of activities. In the observed patterns, the average number of activities equals 5.160 for the training set and 5.155 for the test set. This average length offers room for 1-3 flexible activities complemented with 2-4 in-home activities. Considerable variation occurs, however, as indicated by the standard deviation of approximately 3 activities.

Method	Training set	Test set
Full approach	5.286	5.286
	(2.953)	(2.937)
FS approach	5.014	4.907
	(3.033)	(2.921)

Table 1: Average number of predicted activities in sequences (standard deviation between brackets)

We observe in Table 1 that in general the full approach predicts activity sequences that are somewhat too long, while those of the feature selection approach are rather a little bit too short.

The results of these different methods will now be compared at two levels of aggregation: the choice facet level and the trip matrix level. At the choice facet level, we will discuss the number of attributes that remained in the final decision tree model for each of the two approaches and the probability of a correct prediction for each decision tree. At the trip matrix level, correlation coefficients are calculated to measure the degree of correspondence between the observed and the predicted Origin-Destination matrices.

3.3.1 Choice Facet Level

Tables 2 and 3 provide the results of the analyses conducted to assess model performance at the choice facet level. The first column of these tables presents the nine choice facets of Albatross. The second column lists the levels of the Y-variable, while the third column gives the total number of attributes that were considered to build the final decision tree. The fourth column depicts the size of the decision tree. Column five reports the probability of a correct prediction and in the last column a measure of relative performance, where the probability of a correct prediction is compared to the probability of a correct prediction under a null model. This null model assigns a new case to a category of the Y-variable with a probability, equal to the number of observed cases in the category divided by the total number of cases in the data set.

Decision tree	# alts	# attrs	# leafs	E	e_{ratio}
Mode for work	4	32	8	0.598	0.155
Selection	2	40	35	0.686	0.052
With-whom	3	39	72	0.499	0.223
Duration	3	41	148	0.431	0.145
Start time	6	63	121	0.408	0.285
Trip chain	4	53	8	0.802	0.576
Mode other	4	35	63	0.524	0.222
Location 1	7	28	30	0.540	0.264
Location 2	6	28	47	0.372	0.214

Table 2: Model performance: choice facet level (full approach)

Decision tree	# alts	# attrs	# leafs	E	e_{ratio}
Mode for work	4	2	6	0.595	0.147
Selection	2	1	1	0.669	0.000
With-whom	3	4	51	0.467	0.173
Duration	3	4	38	0.368	0.051
Start time	6	8	1	0.172	0.000
Trip chain	4	10	13	0.811	0.596
Mode other	4	11	60	0.508	0.196
Location 1	7	6	15	0.513	0.222
Location 2	6	8	14	0.312	0.141

Table 3: Model performance: choice facet level (FS approach)

The results of the previous analyses show that, in general, the full approach outperforms the FS approach on the dimensions separately. On the other hand, feature selection generally generates considerably less complex decision trees than the full approach. One exception is the 'trip chaining' choice facet, which more leafs in the final tree with FS than in the tree without feature selection. A logical consequence of this result is that the measure of relative performance of the models with FS is somewhat smaller.
The most important variables for both approaches do not differ that much, but if differences can be discerned, they can then often be explained by high correlations between variables.

3.3.2 Trip Matrix Level

At trip matrix level, we compare the number of trips made from a certain origin to a certain destination. Correlations were calculated between observed and predicted matrix entries in general and for trip matrices that are disaggregated on transport mode. The variation of the correlation coefficient can be largely explained by the variation in the number of cells between matrices. The general OD matrix has 400 cells (20 origins and 20 destinations) and the OD matrix by mode 2000. As could be expected, the fit decreases with an increasing number of cells.

Matrix	ρ (o, p) (Full approach)	ρ (o, p) (FS approach)
None (train)	0.962	0.957
Mode (train)	0.885	0.887
None (test)	0.942	0.947
Mode (test)	0.856	0.849

Table 4: Model performance: trip matrix level

In Table 4 the performance of the different models on the training and the test data set is given. The results indicate that all correlation coefficients are similar. Both approaches perform equally well and if there is a difference it does not exceed the 1% level.

4. Conclusion

In the last decade, rule-based models that predict travel behaviour based on activity diary data have been suggested in the literature. These models usually perform very well, though, very often, they are based on a very complex set of rules. Moreover, research in the field of psychology (Gigerenzer *et al.* 1999) has learned us that simple models often predict human behaviour very well. In fact, the call for simplicity is a question of all ages. Occam's razor, that has to be situated already in the Middle Ages, being an important example. It is in this light that this paper should be regarded. We tried to simplify the complex set of rules used to determine the Albatross system by performing two similar analyses: one with and one without irrelevant variables, while in the second analyses, at same time we cut back in the number of variables. The results of the tree-induction algorithms can namely be heavily influenced by the inclusion of irrelevant attributes. On the one hand, this may lead to over-fitting, while one the other hand, it is not evident whether the inclusion of irrelevant attributes would lead to a substantial loss in accuracy and/or predictive performance. The aim of this study reported therefore was to further explore this issue in the context of the Albatross model system.

The results show that the models that make up their decisions based on one or a few variables are not in any case second to the complex analysis. This comes as a welcome bonus. In fact, more or less the same results were obtained at the trip matrix level. At the choice facet level, one can observe that a strong reduction in the size of the trees as well as in the number of predictors is possible without adversely affecting predictive performance too much. Thus, at least in this study, there is no evidence of substantial loss in predictive power in the sequential use of decision trees to predict activity-travel patterns.

The results indicate that using feature selection in a step prior to tree induction can improve the performance of the resulting model. It should be noted, however, that predictive performance and simplicity are not the only criteria. The most important criterion is that the model needs to be responsive to policy sensitive attributes and for that reason policy sensitive attributes, such as for example service level of the transport system, should have a high priority in the selection of attributes if the model is to be used for predicting the impact of policies. The feature selection method allows one to identify and next eliminate correlated factors that prevent the selection of the attributes of interest during the construction of the tree, so that the resulting model will be more robust to policy measures.

These findings endorse the primary belief that people, because of their limitations in knowledge and time, rely for their choices on some simple heuristics. Since, in the Albatross system, we are trying to predict nine different choices on travel behaviour made by human beings, this might give an idea on why these simple models do not necessarily perform worse than the complex models. However, if simple models are able to predict the choices of a human being, this can mean two things: either the environment itself is perceived as simple, or the complex choice process can be described by simple models. Since activity-based transport modellers keep developing systems with an increasing complexity in order to try to understand the travel behaviour undertaken by humans, we acknowledge that the environment is not simple. However, whether it is perceived as simple by human beings, remains an open question.

5. References

Arentze, T. A. and Timmermans, H. J. P. (2000) *Albatross: A Learning-Based Transportation Oriented Simulation System* Eindhoven University of Technology, EIRASS.

Arentze, T., Hofman, F., van Mourik, H., Timmermans, H. and Wets, G. (2000), Using decision tree induction systems for modeling space-time behavior, *Geographical Analysis,* **32,** pp. 330-350.

Arentze, T. A., Borgers, A. W. J., Hofman, F., Fujii, S., Joh, C.-H., Kikuchi, A., Kitamura, R., Timmermans, H. J. P. and van der Waerden, J. (2001) Rule-based versus utility-maximizing models of activity-travel patterns: a comparison of empirical performance. In *Travel Behaviour Research: The Leading Edge* Ed. D A Hensher, Pergamon Press, Oxford, pp. 569-584.

Domingos, P. (1998) Occam's two razors: The sharp and the blunt. Proceedings of the fourth international conference on knowledge discovery and data mining, pp. 37-43.

Gigerenzer, G., Todd, P.M. and the ABC Research Group. (1999) *Simple Heuristics That Make Us Smart.* Oxford University Press, New York.

Hall, M. A. (1999a) *Correlation-based Feature Selection for Machine Learning,* Ph.D. dissertation, University of Waikato.

Hall, M. A. (1999b) Feature selection for Machine Learning: Comparing a correlation-based filter approach to the wrapper. Proceedings of the Florida Artifical Intelligence Symposium (FLAIRS), Orlando, Florida.

Kira, K. and Rendall, L. A. (1992) A practical approach to feature selection. Proceedings of the Ninth International Conference on Machine Learning, Aberdeen, Scotland, UK, Morgan Kaufmann Publishers, San Mateo, pp 249-256.

Koller, D. and Sahami, M. (1996) Toward optimal feature selection. Proceedings of the 13th International Conference on Machine Learning, Bari, Italy, pp 284-292.

Kononenko, I. (1994) Estimating attributes: analysis and extensions of relief. Proceedings of the 7th European Conference on Machine Learning, Catania, Italy, Springer Verlag, pp 171-182.

Moons, E., Wets, G., Aerts, M., Vanhoof, K., Arentze, T. A. and Timmermans, H. J. P. (2002) The impact of irrelevant attributes on the performance of classifier systems in generating activity schedules. Proceedings of the 81st Annual Meeting of the Transportation Research Board, Washington, D.C.

Moons, E. (2005) *Modelling Activity-Diary Data: Complexity or Parsimony?* Ph. D. dissertation, Limburgs Universitair Centrum, Diepenbeek, Belgium.

Quinlan, J. R. (1993) *C4.5 Programs for Machine Learning.* Morgan Kaufmann Publishers, San Mateo.

Tornay, S. (1938) *Ockham: Studies and selections*. La Salle, IL: Open Court..

Wets, G., Vanhoof, K., Arentze, T. A., Timmermans, H. J. P. (2000) Identifying Decision Structures Underlying Activity Patterns: An Exploration of Data Mining Algorithms *Transportation Research Record,* **1718**, pp. 1-9.