Travel Behaviour and Society 6 (2017) 100–109

Contents lists available at ScienceDirect

Travel Behaviour and Society

journal homepage: www.elsevier.com/locate/tbs

Estimating motorized travel mode choice using classifiers: An application for high-dimensional multicollinear data

Anabele Lindner, Cira Souza Pitombo *, André Luiz Cunha

Department of Transportation Engineering, São Carlos School of Engineering, University of São Paulo, Av. Trabalhador São-carlense, 400, São Carlos, SP 13566-590, Brazil

ARTICLE INFO

Article history:
Received 24 March 2016
Received in revised form 4 August 2016
Accepted 22 August 2016

Keywords:
Artificial Neural Networks
Decision tree algorithms
Travel mode choice
Multicollinear data

ABSTRACT

Studies in the field of discrete choice analysis are crucial for transportation planning. Generally, travel demand models are based on the maximization of the random utility and straightforward mathematical functions, such as logit models. These assumptions lead to a continuous model that presents constraints concerning fitting the data. Artificial Neural Networks (ANN) and Classification Trees (CT) are classification techniques that can be applied to discrete choice models. These techniques can overcome some disadvantages of traditional modeling, especially the drawback of not being able to model high-dimensional multicollinear data. This research paper compares the performance of estimating *motorized travel mode choice* through ANN and CT with a binary logit in a multicollinear study case (aggregated and disaggregated covariates). The dataset refers to an Origin-Destination Survey carried out in São Paulo Metropolitan Area, Brazil in 2007. Classification techniques have shown a good ability to forecast (approximately 80% match rate), as well as to recognize travel behavior patterns. Furthermore, by using the classifier application, the most important covariates within all the datasets can be selected. These covariates can be related to households, as well as to Traffic Analysis Zones.

1. Introduction

Given the need to improve public transportation as a key to dealing with existing urban transportation issues, transportation planners are expected to investigate human behavior regarding travel mode choices.

Mode choice modeling, a part of the classic sequential four-step model, is based on the utility maximization hypothesis and is subdivided into deterministic and probabilistic approaches built on econometrics concepts. Mode choice probabilities and discrete choices were first explored by Ben-Akiva (1973), McFadden (1974) and Domencich and McFadden (1975).

The traditional methods for discrete choice issues, e.g. Multinomial (MNL) and Nested Logit (NL) regression, are mathematical structures which require low computational complexity. On one hand, such models can be easy to handle, on the other hand, they imply limitations related to an attribute recognized as IIA (Independence of Irrelevant Alternatives) and to multicollinear aspects.

The IIA attribute involves the constraint that random error terms are independent (no correlation) and equally (the same variance) distributed (Koppelman and Wen, 2000). Considering this, the estimation of travel mode choice using traditional approaches may result in biased, and hence, uncertain results.

As well as this main disadvantage regarding the IIA attribute, traditional modeling in travel mode choice estimation is based on straightforward mathematical assumptions. This means that the entire data is modeled based on a single continuous function. The mentioned disadvantages may be overcome by different techniques, e.g. Data Mining approaches. Classification Trees (CT), Artificial Neural Networks (ANN), Fuzzy theory, Ensembles of Decision Trees, Support Vector Machine, Knowledge-Based Neural Network and Class Association Rule are examples of data mining approaches mentioned in the literature review. These techniques identify travel patterns and significant explanatory variables, given a dataset. They can be useful to analyze travel demand as they result in convenient estimations through a non-linear function model, or a model with various functions.

Using the mentioned data mining approaches has been widely adopted in activity based approaches (Arentze and Timmermans, 2007; Pitombo et al., 2011). Various authors have also used data mining techniques to estimate the travel mode choice (Vythoulkas and Koutsopoulos, 2003; Senbil et al., 2005; Arentze and Timmermans, 2007; Zhang and Xie, 2008; Xianyu et al., 2008; Lu and Kawamura, 2010; Diana, 2012; Seyedabrishami and Shafahi, 2013; Rasouli and Timmermans, 2014; Pitombo et al., 2015).

* Corresponding author.
E-mail address: cira@sc.usp.br (C.S. Pitombo).

http://dx.doi.org/10.1016/j.tbs.2016.08.003
2214-367X/

Some studies have used CTs to estimate travel mode choices and have explored the comparison between the logit regression and the former technique results (Wets et al., 2000; Moons et al., 2007).

Other references used ANN to analyze discrete travel choices and compared its results with a logit model. Hensher and Ton (2000) and Cantarella and de Luca (2005) applied a multilayer feedforward architecture. Both papers are based on disaggregated data related to households.

On the other hand, Xie et al. (2003) applied both techniques (CT and ANN) to estimate business travel mode. They concluded that the non-parametric approaches outperformed the traditional one.

The mentioned authors who have explored the use of data mining and MNL modeling to predict travel choice behavior have also considered covariates at different levels of aggregation, e.g. household and individual related, as well as trip length and travel time of different choice alternatives. This paper introduces patterns on a different level, i.e. considering Traffic Analysis Zones (TAZs), as well as covariates at a household level. In addition, the reviewed studies on the subject do not take into account high-dimensional multicollinear data (multiple highly correlated variables), as seen in the proposed method of this study.

Additionally, the high performance of data mining applications in the travel mode choice literature shows the interest in applying and developing data mining techniques to different datasets in travel demand. In other words, this study presented the data mining application to modeling mode choice and aims to compare CT, ANN and the logit model to multicollinear data with multiple dimensions. Methodologies with similar purposes, to what is seen in the current study, are still rare in the literature (Omrani et al., 2013).

This article presents five sections besides this introductory one. The second section provides definitions of the covered techniques. The third section describes the proposed method and materials. The fourth section presents the results and main discussions. Finally, conclusions are drawn in the fifth section.

2. Classification techniques in data mining

A classification technique (classifier) is a data mining approach for categorical dependent variables that not only finds patterns, but can also provide capabilities to predict the outcome of a future observation from an input dataset (Tan et al., 2006). Classification methods can be found in the literature as the k-nearest neighbor classifier, Naïve Bayes classifier, Decision Trees (DT), Artificial Neural Networks (ANN) and Support Vector Machines (SVM), for instance.

DT can be used both for continuous and categorical data as a means of a regression or a classification tree (CT), respectively. Taking this into account, in this study the DT is represented as a CT. In order to analyze and estimate the motorized travel mode choice, the CT was used as a data mining process.

Furthermore, this paper also provides basic information regarding the use of ANN for the same purpose of finding patterns and estimating the dependent categorical variable.

2.1. Classification tree

A DT represents a function that takes as input a vector of attribute and returns a "decision" – a single output value, by performing a sequence of tests (Russell and Norvig, 2002).

DTs are a hierarchical representation of the entire data, where each segment is designated as a node. The root node consists of the complete database and is divided sequentially, generating child nodes. Terminal nodes or leaves derive from subgroups in which the data subdivision is no longer conceivable.

As previously mentioned, a Classification Tree (CT) is a DT specific to categorical data. Breiman et al. (1984) developed the primary algorithm for Classification and Regression Trees (CART). CART trees work by dividing a population into binary splits (Fig. 1), which sequentially breaks the data up into smaller parts with maximum homogeneity.

The CART algorithm assesses all possible splits of all explanatory variables and selects the "best" split, starting from the root node. Therefore, the process is repeated to the subsequent nodes. The "best" split is defined by the least impurity overall (Berk, 2008).

Supposing the terminology given by Fig. 1, where t is a node, x_j is an explanatory variable j; x_j^R is the best splitting value of variable x_j and P is the probability of each node, CART solves the maximization problem by Eq. (1). The algorithm searches all possible values of all variables in a matrix X (with M number of variables x_j and N observations) for the best split question $x < x_j^R$, which will maximize the change (decreasement) of impurity measure $\Delta i(t)$.

$$argmax_{x_j \leqslant x_j^R, j=1,\ldots,M}[i(t_{Parent}) - P_{Left} i(t_{Left}) - P_{Right} i(t_{Right})] \quad (1)$$

Eq. (1) implies the definition of the impurity function. Three definitions have been used in the past: Bayes error, cross-entropy function and the Gini index (Breiman et al., 1984) and yet only Gini and the Twoing splitting rules are widely used in practice.

The Gini index uses the impurity function $i(t)$, presented in Eq. (2).

$$i(t) = \sum_{k \neq 1} p(k|t)p(l|t) \quad (2)$$

In this paper, the dependent variable is a binary with classes of "0" for private motor vehicle choice and "1" for transit choice per household, represented by k and l, respectively. Hence, $p(k|t)$ and $p(l|t)$ are conditional probabilities of class k and l given t.

The splitting rule adopted in this study is established by applying the Gini impurity function in Eq. (2) to the maximization problem in Eq. (1).

The CART algorithm chooses the splitting variable that maximizes the explained variance of the predictions of the classes.

The measure of importance of an explanatory variable X in relation to the final tree is defined as the (weighted) sum across all splits in the tree of the improvements that X has when it is used as a primary or surrogate splitter. The variable importance of X is expressed in terms of a normalized quantity relative to the variable having the largest measure of importance. It ranges from 0 to 100, with the variable having the largest measure of importance scored as 100. In this article, the explanatory variable importance is a good indicator to measure the importance of disaggregated variables, as well as the aggregate variables.

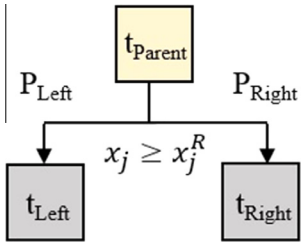

Fig. 1. Splitting algorithm. (Adapted from Timofeev, 2004).