A Data Imputation Method with Support Vector Machines for Activity-Based Transportation Models

Supplementary material

YANG, Banghua; JANSSENS, Davy; RUAN, Da; BELLEMANS, Tom & WETS, Geert
(2013) A Data Imputation Method with Support Vector Machines for Activity-Based Transportation Models. In: Wang, W. and Wets, G. (Ed.). Computational Intelligence for Traffic and Mobility, p. 159-170.

DOI: 10.2991/978-94-91216-80-0_9
Handle: http://hdl.handle.net/1942/16214

Chapter 9

A Data Imputation Method with Support Vector Machines for Activity-Based Transportation Models

Banghua Yang, Davy Janssens, Da Ruan, Tom Bellemans, and Geert Wets

Transportation Research Institute (IMOB) – Hasselt University, Wetenschapspark 5 bus 6, 3590 Diepenbeek, Belgium

Activity-based approaches in transportation models aim at predicting which activities are conducted where, when, for how long, with whom, the transport mode involved and so on. An activity-based framework named FEATHERS (Forecasting Evolutionary Activity Travel of Households and their Environmental RepercussionS) has been developed for Flanders in Belgium. During the establishment of the framework, lots of data are needed. One of the main data sources are activity-based diaries. However, activity diaries tend to contain incomplete information due to various reasons. More recently, with the development of computer science and technology, some artificial intelligence and machine learning techniques have arisen to process the missing data. In this chapter, a data imputation method with a Support Vector Machine (SVM) is proposed to solve the issue of missing data in activity-based diaries. In order to verify the efficiency of SVMs, other methods such as LDA (Linear Discriminant Analysis) and PNN (Probabilistic Neural Network) are also used to process the same data imputation problem. Compared with accuracies obtained by SVMs, the accuracies obtained by LDA and PNN are lower. The initial experimental results show that missing elements of observed activity diaries can be accurately inferred by relating different pieces of information. Therefore the proposed SVM data imputation method in this chapter serves as an effective data imputation method that can induce complete activity diaries in the case of missing information.

9.1 Introduction

Activity-based approaches in transportation models aim at predicting which activities are conducted where, when, for how long, with whom, the transport mode involved and so on. The activity-based approach is a sound option to model people's travel behavior, which has set the standard for travel demand modeling during the last decade (Moons and Wets, 2007). The basic premise of this approach is that travel demand is derived from

W. Wang and G. Wets, *Computational Intelligence for Traffic and Mobility*,
Atlantis Computational Intelligence Systems 8,
DOI: 10.2991/978-94-91216-80-0_9,

the activities that individuals and households need or wish to perform (Jones *et al*., 1983). A dynamic activity-based travel demand framework, FEATHERS has been developed for Flanders (the Dutch speaking region of Belgium) based on the above aim (Janssens *et al*., 2007). The FEATHERS to be applied for the whole Belgium is currently being developed at IMOB. To build the FEATHERS model that can predict all of those above facets, one requires data on all these facets. Clearly, the data collection is a huge challenge. One of the main data sources is activity-based diaries. These diaries are considered to be the most important source of information that benefits the establishment of transportation models. However, activity-based diaries have also been proven to have many disadvantage. One is that the diaries demand high effort to plan and implement and also require high costs in terms of time, finances and other resources. The other is that the collection of diary data frequently brings along a huge burden on respondents to maintain and recall exact details. Consequently, activity diaries tend to contain incomplete information due to various reasons, which is a serious problem because activity-based models require complete diary information.

Activity diaries used in the existing FEATHERS mainly contain individual surveys and household surveys. The surveys are composed of 8551 samples of households and 15888 samples of personals in Belgium. Each sample includes many variables. Among all samples, about 10% samples of households and 5% samples of personals exist missing information. The missing information in a sample may contain one missing variable value, two missing variable values or more than two. If all samples that contain any missing values are deleted and the analysis is then carried out on the samples that remain, some serious drawbacks will be brought. One of drawbacks is the reduction of samples, which will affect the predicting reliability and quality of the FEATHERS model. The other is that the elimination of useful information in the sample will result in serious biases if the samples are not missing completely. The interest of this chapter has centered on performing data imputation, the process by which missing values in a data set are estimated by appropriately computed values, thus constructing a complete dataset.

More recently, with the development of computer science and technology, some artificial intelligence and machine learning techniques have arisen in the area of missing data treatment, such as artificial neural networks (ANN), fuzzy logic systems, and rough sets, which stimulate the missing data research to a new stage. An ANN is a mathematical model or computational model that is inspired by the structure and/or functional aspects of biological neural networks. Modern neural networks are usually used to model complex