Expert Systems with Applications 42 (2015) 5705–5722

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Characterizing activity sequences using profile Hidden Markov Models

Feng Liu [a,*], Davy Janssens [a], JianXun Cui [b], Geert Wets [a], Mario Cools [c]

[a] Transportation Research Institute (IMOB), Hasselt University, Wetenschapspark 5, bus 6, B-3590 Diepenbeek, Belgium
[b] Department of Transport Engineering, Harbin Institute of Technology (HIT), 1500 Harbin, China
[c] Local Environment Management & Analysis (LEMA), University of Liège, Chemin des Chevreuils 1, Bât B.52/3, 4000 Liège, Belgium

ARTICLE INFO

Article history:
Available online 12 March 2015

Keywords:
Profile Hidden Markov Models (pHMMs)
Sequence Alignment Methods (SAM)
Multiple sequence alignments
Activity sequences
Activity-travel diaries
Mobile phone data

ABSTRACT

In literature, activity sequences, generated from activity-travel diaries, have been analyzed and classified into clusters based on the composition and ordering of the activities using Sequence Alignment Methods (SAM). However, using these methods, only the frequent activities in each cluster are extracted and qualitatively described; the infrequent activities and their related travel episodes are disregarded. Thus, to quantify the occurrence probabilities of all the daily activities as well as their sequential orders, we develop a novel process to build multiple alignments of the sequences and subsequently derive profile Hidden Markov Models (pHMMs). This process consists of 4 major steps. First, activity sequences are clustered based on a pre-defined scheme. The frequent activities along with their sequential orders are then identified in each cluster, and they are subsequently used as a template to guide the construction of a multiple alignment of the cluster of sequences. Finally, a pHMM is employed to convert the multiple alignment into a position-specific scoring system, representing the probability of each frequent activity at each important position of the alignment as well as the probabilities of both insertion and deletion of infrequent activities.

By applying the derived pHMMs to a set of activity-travel diaries collected in Belgium as well as a group of mobile phone call location data recorded in Switzerland, the potential and effectiveness of the models in capturing the sequential features of each cluster and distinguishing them from those of other clusters, are demonstrated. The proposed method can also be utilized to improve activity-based transportation model validation and travel survey designs. Furthermore, it offers a wide application in characterizing a group of any related sequences, particularly sequences varying in length and with a high frequency of short sequences that are typically present in human behavior.

1. Introduction

1.1. Problem statement

Efficient characterization of people's activity behavior is of great importance in transportation management and planning as the daily activities of individuals and households have long been hypothesized to be the key determinants of travel demand (e.g. Bhat & Koppelman, 1999; Pendyala & Goulias, 2002). Towards this perspective, a number of studies have been conducted on activity behavior, such as the quantification of the number and the characterization of the types of activities conducted each day (Buliung, 2001). In order to take into account sequential dependencies of the daily activities, Sequence Alignment Methods (SAM), i.e. Optimal String Matching (OSM), originally developed for the analysis of protein sequences (e.g. Durbin, Eddy, Krogh, & Mitchison, 1998), have been brought into transportation science (Wilson, 1998) and social science in general (Abbott & Forrest, 1986). In this adoption, the methods are modified to suit the specific context of sequences related to human activity behavior (e.g. Joh, Arentze, Hofman, & Timmermans, 2002; Joh, Arentze, & Timmermans, 2001; Wilson, 2006; Wilson, 2008). From *activity-travel diaries* that document the full activity-travel behavior performed by respondents during a time frame of one or a few days (e.g. Bhat & Singh, 2000; Spissu, Pinjari, Bhat, Pendyala, & Axhausen, 2009), sequences of activities can be extracted, generating time-resolved activity trajectories of the corresponding individuals. SAM can then be applied to the analysis of the sequences, taking into account not only the separate activities but also the orders of the activities.

The typical application of SAM to the analysis of activity sequences has been conducted in the following process. (i) Given a set of activity sequences, *pairwise sequence alignment algorithms*

* Corresponding author. Tel.: +32 0 11269125; fax: +32 0 11269199.
E-mail addresses: feng.liu@uhasselt.be (F. Liu), davy.janssens@uhasselt.be (D. Janssens), cuijianxun@hit.edu.cn (J. Cui), geert.wets@uhasselt.be (G. Wets), mario.cools@ulg.ac.be (M. Cools).

http://dx.doi.org/10.1016/j.eswa.2015.02.057
0957-4174/

are first employed to compare and score each pair of the sequences on the basis of the composition and sequencing of the activities. (ii) Based on the sequence alignments, *Patterns,* which consist of a substring frequently occurring in the set of sequences, are extracted (e.g. Joh, Arentze, & Timmermans, 2007; Joh, Ettema, & Timmermans, 2008). Alternatively, the obtained alignment scores can be further utilized as distance measures to classify the sequences, through a certain clustering method, commonly the neighbor-joining algorithms. The end nodes (leaves) of a tree consist of all the sequences, while the internal nodes represent the distance relationship between the sequences. (iii) Clusters can subsequently be formed by cutting the tree at certain particular internal nodes, with all the sequences descendent from each of these nodes making up a separate cluster. (iv) Finally, in each obtained cluster, the typical behavior of the sequences is analyzed. *Patterns,* which characterize the particular cluster of sequences, are also derived (e.g. Martin, Schoon, & Ross, 2008; Saneinejad & Roorda, 2009; Shoval & Isaacson, 2007a). If the socio-economic data of the respondents is available, the derived patterns are further associated with the personal information, in order to extrapolate the typical activity and travel behavior represented by the patterns to a wider subpopulation who shares similar socio-economic background (e.g. Wilson, 2008).

Until now, SAM has successfully demonstrated its value and feasibility on the study of activity sequences, by uncovering patterns which capture distinct sequential features buried in the sequences, offering additional insights into activity-travel behavior (e.g. Delafontaine, Versichele, Neutens, & Van de Weghe, 2012; García-Díez, Fouss, Shimbo, & Saerens, 2011; Riedel, Venkatesh, & Liu, 2008; Shoval & Isaacson, 2007a). The sequential information has been dismissed by other traditional sequence analysis methods e.g. Euclidean or Hamming distances (e.g. Wilson, 2006). However, although patterns reveal the common characteristics of a cluster of activity sequences, they have intrinsic limitations in profiling the sequential data. For instance, they focus solely on frequent activities; infrequent activities are thus ignored. Consequently, instead of characterizing all the cluster of sequences as a whole, they only target a part of consensus sequences which share common activities in a particular order. In addition, they simply qualitatively describe the cluster, in the sense that they just disclose what the consensus sequences are without providing a further measure (e.g. a probability) on how likely a consensus sequence or any other particular sequence occurs in the cluster. Furthermore, when the obtained patterns are used to evaluate a new individual's activity sequence, they either match the sequence or do not, providing only a binary answer. There is no information on assessing how similar (or dissimilar) the individual's activity performance is to the cluster of behavior. Thus, from both aspects of characterizing a cluster and evaluating an individual sequence, a pattern is rigid; it exclusively accommodates frequent activities at certain positions, and it only identifies a given sequence that contains all these frequent activities in the corresponding order. Consequently, the infrequent activities and their related travel episodes are disregarded. A sequence, which shares a part of these frequent activities in the corresponding order, is also excluded from the cluster. A model, which quantitatively characterizes an entire cluster of sequences by disclosing the probabilities of both frequent and infrequent activities as well as their sequential orders, and which is able to provide a specific probability measuring the similarity between the cluster and any given sequence, has so far been lacking.

1.2. Profile Hidden Markov Models

To complement patterns, sequence profile methods, and *profile Hidden Markov Models (pHMMs)* in particular, have been developed in bio-informatics and widely applied to *protein sequences,* i.e. strings of letters representing chemical compounds called *residues* (e.g. Finn et al., 2010; Finn et al., 2014). PHMMs are a position-specific scoring system, built upon *a multiple alignment* of a cluster of sequences, which is an extension of pairwise alignment to incorporate more than two sequences at a time (e.g. Durbin et al., 1998). The models classify the positions of the multiple sequence alignment into *match, insertion,* and *deletion* states. The match states model the important positions where a few particular common letters are present for the majority of the sequences; they underlie the basic structure of the cluster. In contrast, the insertion and deletion states correspond to the introduction of additional letters and the omission of certain common letters in a sequence, respectively; they account for the results of random occurrences of the letters. Once built, the pHMMs can be used to score a new sequence and evaluate the relationship between this sequence and the corresponding cluster, in order to classify this new sequence.

Compared with patterns, pHMMs are designed as quantitative descriptors generating numerical weights for each letter at each specific match state, making them more complete than patterns in characterizing both frequent and infrequent activities that occur at the match positions. They are also more effective in modeling both insertion and deletion of activities, thus allowing similarities to be detected from an activity sequence which shares a part of the activities highly habituated in the cluster but may nevertheless skip the other part of the frequent activities or execute extra activities spontaneously on a particular day. When a pHMM is used to evaluate a new activity sequence, the quantitative weights serve to define a score (probability) for the sequence. The score can be used to classify the sequence; in addition, it gives an estimation of how similar or dissimilar the new sequence is to the cluster of behavior.

Apart from SAM, activity sequences have also been analyzed by other methods, including Hidden Markov Models (HMMs) as well as a number of other machine learning and data mining techniques. The sequential data is collected either from sensors (e.g. Chernbumroong, Cang, Atkins, & Yu, 2013) or from various other big data sources, e.g. mobile call locations (Liu et al., 2014), WIFI traces (Danalet, Bilal, & Bierlaire, 2014) and location-based services (Hasan & Ukkusuri, 2014). However, important differences exist between the established methods and pHMMs in terms of both model building process and the types of sequential information that is focused on. HMMs analyze a cluster of sequences through a two stage stochastic process (e.g. Fang, Chen, & Srinivasan, 2014; Zhang, Liu, Jian, & Guan, 2013). First, a set of finite states are defined for the cluster. In the second stage, transition probabilities between the states as well as emission probabilities to generate the observed activities at each of the states are derived. In this modeling process, all the transition and emission probabilities are position independent, in the sense that each position of the sequences could be assigned to any of the defined state set, depending on the states of its previous (one or a few) positions, but not on the absolute location of the position itself. Different positions could be assigned to a same state. As a result, HMMs concentrate only on each part of the sequences, capturing the dependence of the adjacent activities. PHMMs, instead, are position-specific probabilistic matrices, in which each position is modeled through a unique state with its specific transition, emission, insertion and omission probabilities. They belong to a type of HMMs particularly designed for multiple sequence alignments. PHMMs analyze the entire length of the sequences at the same time, through a multiple alignment which reveals the activities at all important positions along the sequences. They are more effective in characterizing the probability distribution of the frequent activities at each of the important positions as well as of