Information Systems 57 (2016) 207–222

Contents lists available at ScienceDirect

Information Systems

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

Labeling sensing data for mobility modeling

Jesse Read *, Indrė Žliobaitė, Jaakko Hollmén

Helsinki Institute for Information Technology HIIT, Aalto University School of Science, Department of Information and Computer Science, PO Box 15400, FI-00076 Aalto, Espoo, Finland

ARTICLE INFO

Available online 14 September 2015

Keywords:
Sensory data
Sensor fusion
Hidden Markov models
Multi-label

ABSTRACT

In urban environments, sensory data can be used to create personalized models for predicting efficient routes and schedules on a daily basis; and also at the city level to manage and plan more efficient transport, and schedule maintenance and events. Raw sensory data is typically collected as time-stamped sequences of records, with additional activity annotations by a human, but in machine learning, predictive models view data as labeled instances, and depend upon reliable labels for learning. In real-world sensor applications, human annotations are inherently sparse and noisy. This paper presents a methodology for preprocessing sensory data for predictive modeling in particular with respect to creating reliable labeled instances. We analyze real-world scenarios and the specific problems they entail, and experiment with different approaches, showing that a relatively simple framework can ensure quality labeled data for supervised learning. We conclude the study with recommendations to practitioners and a discussion of future challenges.

1. Introduction

The availability and penetration of smart mobile devices is increasing; smartphone penetration in Europe is already more than 50% [1], and is forecast to continue at a double-digit annual rate through to the end of 2017. Mobile sensing systems are finding their way in many application areas, such as monitoring human behavior, social interactions, commerce, health monitoring, traffic monitoring, and environmental monitoring [2].

Pervasiveness of mobile phones and the fact that they are equipped with many sensor modalities makes them ideal sensing devices. Since mobile phones are personal devices, we can use the idea of mobile sensing to probe the owner of the phone and the environment, in which the user is moving. Our general interest is to use mobile phones to learn about the mobility patterns of people and to reason and predict about their mobility patterns in urban traffic environment.

The idea of using mobile phones as sensors is not new: mobile phones have been used for context recognition (e.g., [3]) and for measuring social interactions (e.g., [4]) in complex social systems already about a decade ago.

Nowadays, smart phones are equipped with a wide range of sensors, including motion, location and environment sensors, that allow collecting rich observational data about human mobility in urban areas. Various predictive modeling tasks can be formulated based on such data. For example, one can be interested in recognizing the current activity of a person [5], their levels of stress or depression [6] or other metrics of health, predicting the next location [7], or predicting a trajectory of movement [8,9].

In this study, we present a methodology for preprocessing such sensory data for machine learning purposes and its use for analyzing, modeling and predicting human mobility in urban areas. Note that although our experiments involve activity recognition, solving this particular task is not our focus. There is already considerable literature on this topic

* Corresponding author.
E-mail addresses: jesse.read@aalto.fi (J. Read), indre.zliobaite@aalto.fi (I. Žliobaitė), jaakko.hollmen@aalto.fi (J. Hollmén).

http://dx.doi.org/10.1016/j.is.2015.09.001
0306-4379/

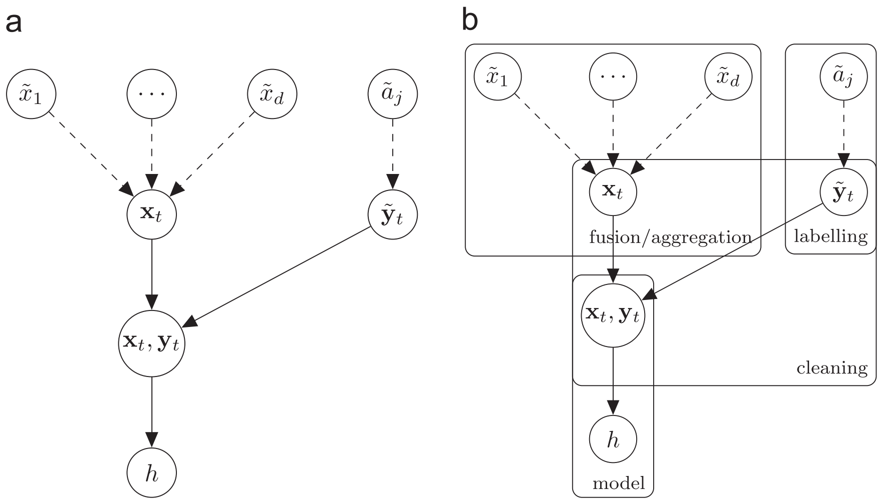

Fig. 1. The overall methodology from raw data to predictive models and analysis. See Table 1 for details of notation. Dashed lines represent irregular arrival of data, solid lines represent time-segmented data into $t = 1, 2, \ldots$. The plates in Fig. 1(b) partition the sequence into processes which we deal with section by section in the paper.

(see, e.g., [5,10,11]). Rather, we focus on cleaning partially labeled data, and general analytics and classification of this data, in particular with respect to the manual annotations. The goal is to ensure a degree of reliability such that the data can be used by supervised learning algorithms.

The main contributions of this study are a survey of tasks involved with mobile sensing in urban environments, via case study identification of issues that arise in this domain, and formulation of a methodology for preprocessing and cleaning sensory data for predictive modeling, in particular to creating reliable labeled instances, as well as highlighting important questions for future research. We focus on the need to automate a process of cleaning and pre-processing, rather than relying on human analysis. This paper extends the preliminary report of [12].

We continue the paper with Section 2, giving an overview of our methodological approach. The sections following are organized with respect to the plates of Fig. 1(b): Section 3 deals with preprocessing for aggregation and fusion of both the input data and the output data (the latter case we term simply 'labeling'), to form a set of time-indexed instances. Section 4 outlines a general methodology for the intermixed process of cleaning and classification of data. Section 5 deals with some of the analytical and modeling issues that can be approached once with reliable labeled data. Section 6 discusses overall results obtained from the experiments throughout the paper, offers recommendations to practitioners, and comments on future work. Finally, Section 7 provides conclusions.

2. Preprocessing methodology

We begin by presenting a methodological approach at a conceptual level. Then the following sections we discuss corresponding algorithmic techniques to be used at different steps of the preprocessing process.

2.1. Challenges in preprocessing

The task of data preprocessing in mobile sensing is not trivial. Data from sensors is collected as a sequence of time stamped observation records, but these records are not equally spaced in time. Moreover, the timestamps of records from different sensors are not matching, i.e., not aligned. In addition, observation records can be of different types: recording discrete events (e.g., battery charger plugged in), processes realized over a period of time (e.g., acceleration), or static measurements (e.g., current temperature) of continuous fluctuations.

For example, consider the battery sensor data (left) with accelerometer data (right):

```
timestamp,temperature,voltage
1371211281,330,4191
1371211281,330,4191
1371211292,330,4190
1371211293,330,4190
1371211293,330,4190
1371211300,330,4119
1371211300,330,4119
1371211301,330,4152
1371211341,330,4190
```

```
timestamp,X,Y,Z
1371211283,-3.027305,7.893985,6.5144534
1371211283,-3.027305,6.1312504,7.817344
1371211283,-3.027305,6.1312504,7.8556643
1371211283,-3.1039455,6.1695704,7.7790236
...
1371211283,-3.027305,6.207891,7.664063
1371211283,-3.1422658,6.207891,7.664063
1371211284,-3.1039455,6.207891,7.817344
1371211284,-3.180586,6.09293,7.7790236
1371211284,-3.027305,6.246211,7.664063
1371211284,-3.027305,6.09293,7.6257424
1371211284,-3.1039455,6.1312504,7.664063
1371211284,-3.027305,6.246211,7.7790236
1371211284,-2.9889846,6.09293,7.7407036
```