Pervasive and Mobile Computing 15 (2014) 181–199

Contents lists available at ScienceDirect

Pervasive and Mobile Computing

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

Behavior analysis of elderly using topic models

Kristin Rieping [a], Gwenn Englebienne [a,b], Ben Kröse [a,b,*]

[a] Informatics Institute, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands

[b] CREATE-IT, Amsterdam University of Applied Science, Duivendrechtsekade 36-38, 1096 AH Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Available online 10 July 2014

Keywords:
Activity discovery
Sensor homes
Pervasive computing
Sequential patterns
LDA
Topic models

ABSTRACT

This paper describes two new topic models for the analysis of human behavior in homes that are equipped with sensor networks. The models are based on Latent Dirichlet Allocation (LDA) topic models and can detect patterns in sensor data in an unsupervised manner. LDA–Gaussian, the first variation of the model, is a combination of a Gaussian Mixture Model and the LDA model. Here the multinomial distribution that is normally used in the LDA model is replaced by a set of Gaussian distributions. LDA–Poisson, the second variation of the model, uses a set of Poisson distribution to model the observations. The Poisson distribution is better suited to handle counts of stochastic events but less well-suited to model time. For this we use the von Mises distribution, resulting in 'LDA–Poisson–von-Mises'. The parameters of the models are determined with an EM-algorithm. The models are evaluated on more than 450 days of real-world sensor data, gathered in the homes of five elderly people, and are compared with a baseline approach where standard k-means clustering is used to quantize the data. We show that the new models find more meaningful topics than the baseline and that a semantic description of these topics can be given. We also evaluated the models quantitatively, using perplexity as measure for the model fit. Both LDA–Gaussian and LDA–Poisson result in much better models than the baseline, and our experiments show that, of the proposed models, the LDA–Poisson–von-Mises model performs best.

1. Introduction

With the increasing number of elderly that live independently in their own homes, sensing systems that monitor the activities of the elderly are becoming popular. Most commercial systems focus on alarms when accidents happen, such as a fall. More advanced monitoring systems go beyond alarming and try to measure the functional health, for example by recognizing ADL: activities performed on a daily basis such as sleeping, toileting and cooking [1].

Wireless networks of sensors, either wearable or ambient, have been presented for measuring these activities (see [2,3] for recent surveys). For activity recognition, various machine learning techniques have shown to give good performance. Based on annotated data sets naive Bayes [4], dynamic Bayesian networks such as HMM's [5], conditional random field and hierarchial methods have been presented. However, large amounts of annotated data are required. The task of labeling a data set is time consuming and if the elderly is responsible for the annotation it can also influence the output of the data. That is why unsupervised techniques are more promising for such systems.

Unsupervised activity discovery can be done in many ways. Recently topic models have been presented to automatically find activity patterns in sensor data [6–10]. Topic models are initially designed for classifying text documents where they

* Corresponding author at: Informatics Institute, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands. Tel.: +31 610963537.
E-mail address: b.j.a.krose@uva.nl (B. Kröse).

http://dx.doi.org/10.1016/j.pmcj.2014.07.001
1574-1192/

are able to find abstract topics, such as 'politics', 'sports', and 'finances'. There is however a major difference between textual and sensor data.

In the LDA model presented by Blei [11] the documents are represented by a Bag-of-words (BOW), frequencies of words from a dictionary. To make use of the LDA model on sensor data one has to create artificial words. There are numerous ways to create words from sensor data. Some researchers create artificial words by discretization of the sensor data. The BOW representation then directly is used on these 'words' [6,10]. If a coarse granularity is used, the resulting number of words will be limited and not expressive enough to capture all variations in the original data. A fine granularity on the other hand results in a large dictionary of words and more data are needed for model building. A more advanced way of categorization is to cluster the data using machine learning techniques. The found categories can then serve as input for the LDA model [9,8]. However, in this way two (possibly conflicting) optimization problems are defined: the clustering of the sensor data and the estimation of the topics in the model.

The contribution of this paper is that we propose a principled approach for clustering the sensor data by combining the clustering and the LDA in one model and do a joint optimization of the parameters. We study two variants, the 'LDA–Gaussian' model and the 'LDA–Poisson' model. These models are able to capture similar observations/words into the same topic automatically. The parameters of both models are found with an Expectation–Maximization-procedure (EM-procedure), which uses the likelihood of the model to converge to the optimal model parameters.

In the next section an overview of related approaches is given. In Section 3 a detailed description of the data and observation vectors is given. In Section 4 the 'LDA–Gaussian' and 'LDA–Poisson' models are introduced. Section 5 contains the different experiments that are performed on real sensor data, that is obtained from the houses of solitary living elderly. In Section 6 the conclusions are presented and some suggestions for future work are given.

2. Related work

The International Classification of Functioning, Disability and Health (ICF) of the World Health Organization [12] provides the tools for extensively describing the functional health status of an individual. The activities of a human are an important indicator, and changes in the daily behavior patterns can be a sign of changes in the health of people [13]. This can be both mental or physical declines. There are different ways to monitor the health condition of people. Cameras or microphones can be very useful to monitor people's behavior [14,15], but these sensors are invading the privacy of people and often not accepted as sensors in people's homes.

Simple binary sensors such us motion sensors, contact switches or pressure mats are preferable for health monitoring in home environments. These sensors are low in cost and easy to install. Moreover, they are also experienced as non-intrusive and not disturbing by the inhabitants. Numerous researchers implemented different approaches to apply activity recognition on data generated by these kind of sensors [16]. Activity models can be built using one of two methods. The first is to learn activity models from pre-existent large-scale data sets of users behaviors using data mining and machine learning. The other method is to exploit prior knowledge in the domain of interest and to construct activity models using knowledge engineering and use formal logical reasoning for prediction and inference. In the first line probabilistic generative and discriminative models have been presented. Tapia et al. [4] use a naive Bayes classifier to find activities in annotated sensor data. They show that it is possible to find activities in ubiquitous, simple sensor data, that was obtained in real-life environments. In the work of van Kasteren et al. [17] two approaches for recognizing activities in sensor data are compared. The Hidden Markov Model and the Conditional Random Field are both applied to annotated, real-life sensor data. The authors also vary between different representations of the sensor readings and show that this can improve the results for recognizing activities. Support vector machines are used for activity recognition in [18], where not only ambient sensors are used but also wearable sensors. The second method is illustrated by the work of Hong et al. [19] who use ontologies to describe daily activities. An evidential network is used to describe activities in a hierarchical way.

All of the previous approaches used annotated data. Generating this labeled data is difficult, time consuming and the determined labels are not always accurate. For this reason unsupervised methods are preferred above supervised methods in this field. Many unsupervised methods are based on mining frequent sequences. Rashidi et al. [20] use mining techniques in combination with a clustering step to discover activities and track them.

Various authors have applied LDA with different kind of data. This topic model is able to find abstract descriptions of activities in data automatically. Chikhaoui et al. [7] use the topic model LDA in combination with sequential pattern mining to find activities in various data sets. The sequential patterns are used as words, which are the input for the LDA model. They test their method on varied annotated data sets. The topics that are found describe activities and the accuracy is measured by comparing the topics with annotation labels. In their work the focus is on detecting activities and not so much on the global idea of detecting behavior patterns as it is done in this work.

Huynh et al. [9] and Casale et al. [8] both try to discover daily routines from sensor data. Acceleration sensors that are attached to the human body generate a continuous stream of data. The data are quantized into different time intervals and in this way artificial words are created. The dictionary, which contains all unique words, is quite large, because small variations in the sensor data result in different words. Therefore the authors cluster the data beforehand with the k-means algorithm to reduce the size of the dictionary. The choice of k however is of big influence on the outcome of the LDA model, and this paper uses a different approach to handle the large amount of variation in the artificial words.