Information Fusion 43 (2018) 33–46

Contents lists available at ScienceDirect

Information Fusion

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

Mode tracking using multiple data streams

Mohamed-Rafik Bouguelia[*,a], Alexander Karlsson[b], Sepideh Pashami[a], Sławomir Nowaczyk[a], Anders Holst[c]

[a] *Center for Applied Intelligent Systems Research, Halmstad University, Sweden*
[b] *University of Skövde, Sweden*
[c] *Swedish Institute of Computer Science, Sweden*

ARTICLE INFO

Keywords:
Mode tracking
Clustering
Data streams
Time series
Knowledge discovery

ABSTRACT

Most existing work in information fusion focuses on combining information with well-defined meaning towards a concrete, pre-specified goal. In contradistinction, we instead aim for autonomous discovery of high-level knowledge from ubiquitous data streams. This paper introduces a method for recognition and tracking of hidden conceptual modes, which are essential to fully understand the operation of complex environments, and an important step towards building truly intelligent aware systems. We consider a scenario of analyzing usage of a fleet of city buses, where the objective is to automatically discover and track modes such as highway route, heavy traffic, or aggressive driver, based on available on-board signals. The method we propose is based on aggregating the data over time, since the high-level modes are only apparent in the longer perspective. We search through different features and subsets of the data, and identify those that lead to good clusterings, interpreting those clusters as initial, rough models of the prospective modes. We utilize Bayesian tracking in order to continuously improve the parameters of those models, based on the new data, while at the same time following how the modes evolve over time. Experiments with artificial data of varying degrees of complexity, as well as on real-world datasets, prove the effectiveness of the proposed method in accurately discovering the modes and in identifying which one best explains the current observations from multiple data streams.

1. Introduction

In the current era of massive amount of information that is generated in real-time, new methods for extracting meaningful knowledge from the data are desperately needed. The majority of information fusion research nowadays focuses on information that has been analyzed and pre-processed by human experts, both in terms of what the data is expected to contain, as well as what is the purpose of the analysis. The ubiquitous streaming context, however, requires systems that can be used in a fully autonomous way, ones that adapt to the task at hand and can themselves discover high-level knowledge from data.

The rapid growth of *streaming data* generated throughout all of society brings the need to perform different types of analytics for different tasks at hand. It is important that automatic knowledge discovery methods are able to deal with various characteristics of such unprocessed data. *Fusing* various information sources to reveal *higher order structure* in the data is challenging, in particular because of the very open-ended nature of the task. One important capability is tracking of hidden conceptual states, referred to as a *mode estimation*. The operation of most, if not all, complex environments can only be understood by recognizing high-level order, often on several different levels of abstraction. One common type of such a structure in data that is typically useful to detect is *clusters* [1]. In case of data streams describing a dynamic environment, these clusters can be interpreted as underlying “hidden states”, or *modes*, of the environment. By *tracking* such modes, i.e., analyzing which best explains the currently seen data, one can estimate and predict the behavior of the environment and take actions accordingly.

Tracking in general is one of the core problems within the research field of *information fusion* (IF) [2]. However, most often these tracking methods are designed to handle states directly coupled to some physical property of the environment of interest [3], e.g., *position* and *velocity* of some object, and not more *abstract* states such as a road vehicle having being driven in an aggressive or calm way. To be more specific, the IF research field is often divided [4] into two main parts referred to as *Low-Level IF* (LLIF), where tracking and filtering of physical states within the environment of interest is the main focus and *High-Level IF* (HLIF) where the goal is to obtain a high-level, composite,

[*] Corresponding author.
E-mail addresses: mohbou@hh.se (M.-R. Bouguelia), alexander.karlsson@his.se (A. Karlsson), sepideh.pashami@hh.se (S. Pashami), slawomir.nowaczyk@hh.se (S. Nowaczyk), aho@sics.se (A. Holst).

https://doi.org/10.1016/j.inffus.2017.11.011
Received 15 July 2016; Received in revised form 10 October 2017; Accepted 17 November 2017
Available online 23 November 2017
1566-2535/ © 2017 Elsevier B.V. All rights reserved.

understanding of the current situation of interest, i.e., mode estimation, often framed as achieving *situation awareness* [5]. Historically, methods within HLIF have not to the same extent been studied in depth, however, they are very important for many practical purposes.

One of the main underlying motivations for our proposed method within this paper is a real-world scenario of analyzing the usage of fleet of city buses. The modes of interest within such a scenario can include concepts such as highway route, heavy traffic, or aggressive driver. Our goal is to develop a method that can discover theses modes, as well as detect and track them, based on available on-board signals. Hence, in a sense we aim to perform LLIF-tasks in an HLIF setting due to the compound abstract, situational oriented, nature of the states that we track, i.e., the modes. However, methods and models from LLIF rely on assumptions that are hard or impossible to satisfy when abstract states are of interest, in particular when used in systems that handle a large number of data streams. There are several reasons for this, and in this work we will focus on the following aspects: (I) one cannot, in advance, adequately predict, describe or interpret these modes, i.e., identifying the modes that exist in the environment, which are initially unknown, is one of the core parts of the automatic knowledge discovery process; s (II) the streaming information sources may be unpredictable, i.e., they may emerge and disappear at any time, have irregular sample intervals, or many missing values; and (III) many of the high-level, conceptual modes within dynamic environments only emerge based on multiple features analyzed over extended time frame, and not from single measurements, as is the case with traditional tracking.

There are several rather recent literature surveys on learning from streaming data [6–10] depicting the state-of-the-art with respect to both unsupervised as supervised learning. Many approaches are based on ensemble learning due to the fact that such models have been proven to be useful for many issues related to stream mining [10]. On this theme, a comprehensive taxonomy and discussion of ensemble methods for various data stream mining tasks is provided in [9]. Besides, methods for data preprocessing that cope with streaming data have been categorized and analyzed in [11] and issues related to concept drift have been surveyed in [12]. Many of the stream mining methods rely on geometric properties of the points to be clustered, e.g., distance-based ones [13–15], or on the distribution of these points within the cluster space, e.g., density-based ones [16–18]. In addition, several model-based approaches to the problem of learning on streaming data have been proposed, of which we highlight here one: SWEM, an EM-based algorithm, which uses a sliding window and a Gaussian mixture model (GMM) with incremental learning [19].

In this paper we propose an approach for tracking high-level modes of the environment of interest that specifically aims to handle the challenges listed above. Our contribution is to use the combination of three important ideas: first, we *aggregate* each low level signals over time into a number of features that represent longer periods of interest, sufficient for the more abstract conceptual states to emerge; second, we use *unsupervised clustering* in order to fuse these features from multiple sources, instead of using the original signal space, to find candidates for higher level modes; and finally we apply *Bayesian tracking* [20], back in the original signal space, with the clusters as seeds, to more accurately follow mode transitions and to fine-tune the definitions of the modes, as indicated by the low level signals.

In addition to proposing and implementing this novel method, and as part of showcasing its effectiveness, we present a *city bus fleet scenario* [21]. During a four-year longitudinal study we have been collecting data concerning the operation and usage of a fleet of buses in regular operation in around a city in western Sweden. The primary purpose is improving uptime, by comparing one vehicle against the rest of the group to detect deviations and predict component failures. In this context it is important to determine how the vehicle is being used at different points in time, to enable "within-mode" analytics and to more fairly compare the behavior of the vehicles. High-level mode tracking will also enable other applications that require hidden context information, such as *advanced driver assistance- and active safety functions* (ADAS) [22]. Some example modes of interest include whether or not the vehicle is currently being driven in an "aggressive way," indicating a stressed or inexperienced driver; whether the vehicle is being driven in light or heavy traffic; or the type of the road and environment where the driving takes place, e.g., city, countryside or highway.

The selected scenario clearly showcases the four challenges mentioned previously: (I) the modes of interest are not known *a priori* – we have certain expectations based on domain knowledge, however, it is clear that we are not aware of all the important hidden states; (II) the data is originating at different subsystems of the vehicle, and is transferred over unreliable network; and (III) concepts such as aggressive or inexperienced driver do not manifest in individual sensor readings, but are only discernible over longer periods of time, from a combinations of several signals. This, together with another evaluation based on a publicly available Electromyography dataset, completes the contribution of our paper: we present a novel method designed to handle specific deficiencies in existing solutions, and showcase its effectiveness and generality through experimental evaluation.

2. Proposed approach

In this section we provide the formal problem definition, as well as detailed explanation of the proposed method. The aim of the proposed approach is to find the underlying modes using aggregated features and to update these modes based on the continuous streams of data.

2.1. Problem definition

Let us assume that we have n data streams $\{y_t^i \in \mathbb{R}\}_{i=1}^n$ sampled at time points $t \in \{0, 1, ...\}$ that can be associated with some entity within a dynamic environment of interest. The goal is then to track the underlying mode $M_t \in \mathscr{M}$ of the entity, i.e., a certain underlying hidden state related to the entity that best explains some aspects of the currently seen data, based on multiple sources of data streams. We assume that we cannot pre-determine the mode space $\mathscr{M}$ but rather need to infer it from some historic data. We also assume that the information sources $i \in \{1, ..., n\}$ can appear or disappear at any time, are not necessarily synchronized, and likely to contain missing values.

2.2. Approach

As mentioned in the Introduction, for estimation of abstract states, or modes, it is not sufficient to consider individual values at separate points in time from the raw data signals. In our example scenario, the values of vehicle or engine speed at any given time point are very weak indicators of whether the mode is "aggressive" or "calm" driving of that vehicle. Instead, it is necessary to provide aggregations of this data over longer periods. Therefore, the first step is to consider all values over some time frame $\mathscr{T} \triangleq \{t_i, ..., t_{i+w}\}$, where $w = |\mathscr{T}|$ is the time window size. Our proposed approach consists of three steps: *aggregate* the data into higher order features; perform *clustering* on the fused (joint) feature space of the aggregated data; and finally apply *tracking* of the time series. We refer to the proposed approach as the *ACT* (Aggregation, Clustering, and Tracking) method.

2.2.1. Aggregation of data

As previously mentioned, performing clustering directly on the raw signal values coming from the streams, not only suffers from the missing values problem, but is also not sufficient to discover interesting abstract conceptual modes. In order for such modes to emerge, we need a number of different representations of the data, computed over sufficient periods of time.

For this purpose, we investigate two complementary classes of approaches for taking the history of values into account. One such class of methods is based on calculating higher order features from the time

Download English Version:

https://daneshyari.com/en/article/6937961

Download Persian Version:

https://daneshyari.com/article/6937961

Daneshyari.com