Information Sciences 334–335 (2016) 250–272

Contents lists available at ScienceDirect

Information Sciences

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

Evolving association streams

Andreu Sancho-Asensio[a,*], Albert Orriols-Puig[a], Jorge Casillas[b]

[a] Research Group in Electronic and Telecommunications Systems and Data Analysis, La Salle - Ramon Llull University, Quatre Camins 2, Barcelona 08022, Spain

[b] Department of Computer Science and Artificial Intelligence, University of Granada and the Research Center on Information and Communications Technology (CITIC-UGR), Granada E-18071, Spain

ARTICLE INFO

Article history:
Received 1 October 2014
Revised 10 November 2015
Accepted 19 November 2015
Available online 10 December 2015

Keywords:
Online learning
Data stream
Association rule
Concept drift
Genetic fuzzy systems

ABSTRACT

The increasing bulk of data generation in industrial and scientific applications has fostered practitioners' interest in mining large amounts of unlabeled data in the form of continuous, high speed, and time-changing streams of information. An appealing field is association stream mining, which models dynamically complex domains via rules without assuming any a priori structure. Different from the related frequent pattern mining field, its goal is to extract interesting associations among the forming features of such data, adapting these to the ever-changing dynamics of the environment in a pure online fashion-without the typical offline rule generation. These rules are adequate for extracting valuable insight which helps in decision making. This paper details Fuzzy-CSar, an online genetic fuzzy system designed to extract interesting rules from streams of samples. It evolves its internal model online, being able to quickly adapt its knowledge in the presence of drifting concepts. The different complexities of association stream mining are presented in a set of novel synthetic benchmark problems. Thus, the behavior of the online learning architecture presented is carefully analyzed under these conditions. Furthermore, the analysis is extended to real-world problems with static concepts, showing its competitiveness. Experiments support the advantages of applying Fuzzy-CSar to extract knowledge from large volumes of information.

1. Introduction

With the invention of digital computers, technology has allowed the industry to collect and store massive amounts of data concerning business processes, allowing for subsequent analysis and exploitation. Since then, the analysis and exploitation of large databases is a recurrent topic and there have been several contributions across a wide set of disciplines [5]. The need for extracting useful information from massive amounts of data, usually as a continuous, high speed and time-changing data stream has risen dramatically in industrial and scientific applications [5]. Stock markets and smart networks, among many others, are the primary targets of this kind of knowledge extraction [15,33]. However, traditional algorithms were not designed for handling data which is delivered in the form of streams and they are not able to extract accurate models in these environments [34], thus requiring online techniques. The challenges hampering the learning process under data streams are manifold: (1) obtaining a fast reaction time for changes in concept, (2) data can only be handled once, (3) storage limitations, and (4) varying noise levels.

* Corresponding author. Tel.: +34 677281045.
E-mail addresses: andreu.sancho@gmail.com, andreus@salleurl.edu (A. Sancho-Asensio), aorriols@salleurl.edu (A. Orriols-Puig), casillas@decsai.ugr.es (J. Casillas).

http://dx.doi.org/10.1016/j.ins.2015.11.043
0020-0255/

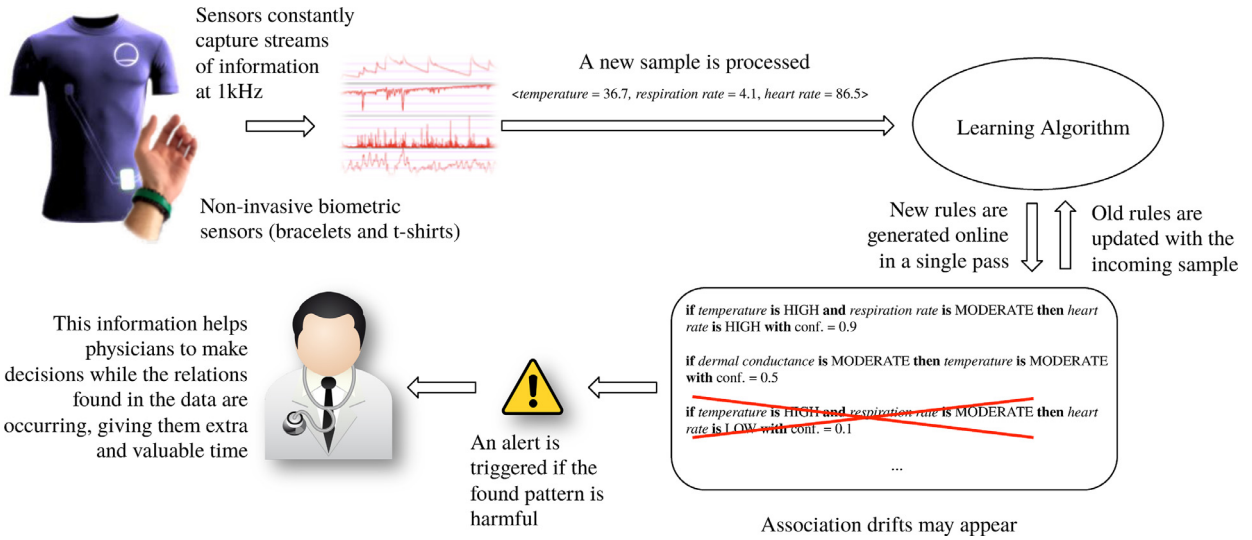

Fig. 1. An application of association stream mining. In this scenario association streams help physicians and doctors to make decisions, giving them extra and valuable time by modeling the scenario dynamically.

Despite the need for addressing the complexities of mining new, potentially useful information from data streams, current online mining field research focuses mainly on supervised methods, which assume an a priori relational structure for the set of features that define the problem. This issue is often distant from real-world situations, and it is emphasized when changes in concept occur: the undefined and ill-structured nature of the situation jointly with the dearth of labelled examples—or complete absence of them—from this new concept makes the application of a *pure* supervised algorithm ill-suited. Hence, solving the problems requires the use of hybrid methods which combine supervised with unsupervised techniques in order to deal with the problem [41].

A real case that entails a sound example of this issue is in detecting potential threats to web sites and network infrastructures. In this scenario there are a set of features that indicate suspicious acts on the infrastructure (i.e., strange characters in login interfaces or strange traffic flows, among others) by malicious users trying to identify the vulnerabilities of the system to take possession of it. It is worth mentioning that other anomaly detection strategies exist, such as statistical or based on data density. However these are typically based on labeled data and, therefore, they do not adapt themselves to concept changes. In our particular case we are not interested in directly detecting—and thus informing of the attack—but in continuously monitoring the system by adapting to the new trends that data are showing, which helps experts to decide if the system is under attack. Fig. 1 depicts a possible use-case of association stream, where multiple non-invasive biometric sensors (bracelets and t-shirts) that capture continuous physiological data in the form of streams are given to patients. In this environment, the goal is to help physicians and doctors to monitor and track the status of their patients in an online fashion.

In this regard, the unsupervised approach becomes a feasible alternative to the aforesaid issues. More specifically, *association stream mining*, focused on extracting associations among variables via production rules online and in a single pass, is specially attractive from the point of view of practitioners due to (1) the demand of interpretability of the patterns discovered in data—human-readable rules that provide valuable insight (e.g., *every time X occurs Y also happen*)—, (2) the need for discovering patterns while they are happening (e.g., the new steps of the web attack) and hence adapt to them, and (3) the high and continuous volumes of data to be processed, which require scalable learners.

Despite the importance of facing association streams, it is a new area and consequently, to the best of our knowledge, so far no fully operational approaches tackling the challenges discussed in this paper have been presented. The most similar algorithms for knowledge extraction under these conditions focus mainly on the identification of frequent variables—referred to as *online frequent pattern mining*—, relegating the generation of rules in a second place in an offline process [5], which make these mostly unpractical for handling association stream premises, where the rules discovered have to be present immediately and adapt to the changing dynamics of the continuous flow of data. Several investigations have been presented so far on frequent pattern mining, but the majority of these ignore the presence of concept drifts. Moreover, most of those algorithms are only able to deal with problems described by categorical features [16,22]. Even so, a few quantitative algorithms for frequent pattern mining have been proposed, most of them based on fuzzy logic [12,30], focusing on the identification of frequent variables and not on the final rules. Likewise, these systems are not designed for the harsh conditions of continuous flow and time-changing concepts that association stream stresses.

Association stream mining research area is rooted in traditional, offline association rule mining, one of the most well known Data Mining fields. It is used to extract interesting information from large data bases by means of describing the properties of data in the form of production rules, e.g., $X \Rightarrow Y$, where both X and Y are sets of features and $X \cap Y = \emptyset$ [1]. Association rule