Environmental Modelling & Software xxx (2018) 1–13

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

# Environmental data stream mining through a case-based stochastic learning approach

Fernando Orduña Cabrera, Miquel Sànchez-Marrè [1, *]

ARTICLE INFO

*Article history:*
Received 4 March 2017
Received in revised form
29 November 2017
Accepted 17 January 2018
Available online xxx

*Keywords:*
Data science
Data stream mining
Dynamic case learning
Stochastic learning
Case-based reasoning
Air quality detection
Environmental modelling

ABSTRACT

Environmental data stream mining is an open challenge for Data Science. Common methods used are static because they analyze a static set of data, and provide static data-driven models. Environmental systems are dynamic and generate a continuous data stream. Dynamic methods coping with the temporal nature of data must be provided in Data Science. Our proposal is to model each environmental information unit, timely generated, as a new case/experience in a Case-Based Reasoning (CBR) system. This contribution aims to incrementally build and manage a Dynamic Adaptive Case Library (DACL). In this paper, a stochastic method for the learning of new cases and management of prototypes to create and manage the DACL in an incremental way is introduced. This stochastic method works with two main moments. An evaluation of the method has been carried using a data stream of air quality of the city of Obregon, Sonora. México, with good results. In addition, other datasets have been mined to ensure the generality of the approach.



## 1. Introduction

Data science discipline has a main goal to obtain effective knowledge from data. This effective data analysis must be translated into useful knowledge and information for decision support. In the usual application of data mining techniques, there is a gap between the mined models and the effective knowledge needed for a reliable decision-making process. Many works in the literature have applied some data mining methods to build assessment models and/or predictive models in environmental domains. The main problem on most of the approaches is that assessment and predictive models are based on static supervised environmental databases gathered from a concrete period of time. Therefore, on one hand the models are static and cannot capture the dynamic nature of such environmental data streams, and on the other hand, the number of different environmental conditions or states are determined *a priori*. These models are not very reliable for the stakeholders' decision-making process. The temporal component of data must be taken into account jointly with quantitative and qualitative data, and the number of different environmental conditions or states must be discovered through the mining of the environmental data stream. The approach presented in this paper integrates these different information sources and build an incremental data-driven method based on a case-based stochastic learning approach for assessment of environmental conditions. To illustrate the approach, it has been applied to a case study on air quality assessment conditions.

In recent years, the problem of mining data streams has grown the attention of many researchers. Many real-world applications generate data continuously. For example, in network monitoring, environmental data sensors, telephone record calls, multimedia data, customer transactions, customer click streams, and so on. Advances in technology have facilitated new ways of continuously collecting data. In many applications, the volume of such data is so large that it may be impossible to store the data on disk. Furthermore, even when the data can be stored, the volume of the incoming data may be so large that it may be impossible to process any particular record more than once. Therefore, many data mining and database operations such as classification, regression, clustering, frequent pattern mining and indexing become significantly more challenging in this context (Aggarwal, 2007). The monitoring of many events in real time produces much information. In recent years, data stream mining field has grown rapidly. In (Hulten and Domingos, 2001) outlined some desirable properties for learning tasks in data streams: incrementality, constant time to process each

* Corresponding author.
*E-mail address:* miquel@cs.upc.edu (M. Sànchez-Marrè).
[1] Intelligent Data Science and Artificial Intelligent Research Centre (IDEAI-UPC), Knowledge Engineering & Machine Learning Group (KEMLG), Dept. of Computer Science, Universitat Politècnica de Catalunya (UPC), Barcelona, Catalonia.

https://doi.org/10.1016/j.envsoft.2018.01.017
1364-8152/

example, single scan over the training set, and taking *concept drift* into account. Learning from data streams require *incremental learning* algorithms that take into account the problem of *concept drift*: the underlying concept or distribution of the data can change over time, and the mined models should be aware of the changes, and adapt themselves to the changes.

Our proposal is to model the new environmental information timely generated as a new case/experience in a Case-Based Reasoning (CBR) system. CBR (Richter and Weber, 2013; López de Mántaras et al., 2005) solves new problems/cases using old similar solved problems/cases in the same domain. The designed method aims to incrementally learn the new cases, build new prototypes of cases, and store the cases in the most similar prototype. A prototype of cases is a generalization of similar cases, which represents a similar environmental situation (i.e., an air quality environmental condition). This contribution aims to incrementally create and manage a Dynamic Adaptive Case Library (DACL) as a way to strengthen its structure. The proposed method implements a DACL, which can be used as a predictive system for new environmental data situations, warning the experts about dangerous situations for the citizen. In this paper, a stochastic method for the learning of new cases and management of prototypes to create and manage the DACL in an incremental way is introduced. This stochastic method works with two main moments. The first moment guides the learning of new cases and decides where to store the cases. The second moment evaluates the candidate prototype, and select it or builds a new prototype. This way, through this data-driven process, an automatic, adaptive and dynamic system for supporting decisions can be deployed.

### 1.1. Background knowledge

In machine learning literature, several works have addressed the problem of learning from data streams (Gama, 2010; Gama and Gaber, 2007), and other works studied the time changing concept problem (Hulten et al., 2001; Klinkenberg and Joachims, 2000; Maloof and Michalski, 2000; Kubat and Widmer, 1995). Most common techniques used are temporal windows, which determines the training set for the learning algorithm, and the weighting of examples, which attempts to decrease the relevance of the older examples, and increase the relevance of the new ones. Also there are some mixture techniques. Some authors (Klinkenberg, 2004; Klinkenberg and Renz, 1998; Widmer and Kubat, 1996) propose the use of adaptive time windows in order to minimize the generalization error of the classification models.

Continuous problem domains (i.e., domains where cases are generated from a continuous data stream) require different underlying representations and place additional constraints on the problem solving process (Ram and Santamaria, 1997). define three characteristics where the problem domain is continuous, and those are: First, they require *continuous representations*. Second, they require *continuous performance*. Third, these problem domains require *continuous adaptation* and learning. Reasoning about continuous domains is not an easy task. Moreover, this is a domain where CBR can rapidly extend its benefits because data is systematically collected for its analysis. A CBR system that continuously interacts with an environment must be able to create autonomously new situation cases (new concepts or clusters) based on its perception of the local environment in order to select the appropriate steps to achieve the current mission goal (Haris and Slobodan, 2005), but a general framework is still missing. Some systems that use case-based methods in continuous environment are described in (Urdiales et al., 2006; Kruusmaa, 2003; Ram et al., 1997).

There are two other central problems derived from the continuous nature of some domains. First of all, the *size of the case library* could grow very fast as the CBR system is learning new cases without an extensive improvement in the competence of the system, as pointed out in (Miyashita and Sycara, 1995). Two natural human cognitive tasks appear as the solution to these problems: forgetting (Keane and Smith, 1995) and sustained relevant learning (Sànchez-Marrè et al., 1999). On the other hand, learning many cases could provoke an *overhead in the case library organization*. As new cases are stored in the case library, it will be necessary to update the case library organization (Meléndez et al., 2001).

Sànchez-Marrè in (Sànchez-Marrè et al., 2000) introduced the idea of using prototypes for improving the retrieval of cases in a static multiple case library. This multiple case library was composed of expert-defined prototypes and its corresponding hierarchical case libraries.

Finestrali and Muñoz in (Finestrali and Muñoz-Avila, 2013) implements a stochastic explanation to determine the learning goal while plays Wargus. This game is an example of a stochastic domain. Their studied the problem of explaining events in stochastic environments using Case-Based Reasoning (CBR). The center of their approach has three ideas: (1) Using the notion of Stochastic Explanations (2) Retaining as cases (event, stochastic explanation) pairs when such unexpected events occur. (3) Learning the probability distribution in the stochastic explanation as the cases were retrieved. Their proposal is novel and somewhat similar to our work. In our work, the stochastic method is the core of the learning of new cases and building new prototypes, while they use Q-Learning to give the stochastic explanation of the learning and labelling of the case.

Stochastic clustering methods have been introduced in clustering field like in the case of Chuan in (Tan et al., 2010), where their proposal examines a practical stochastic clustering method that has the ability to find clusters in datasets without requiring users to specify the centroids or the number of clusters. Their experimental setup confirms that the proposed method performs competitively against the traditional clustering methods in terms of clustering accuracy and efficiency. In their work, an estimation of the similarity thresholds for $n$ items is computed. The estimation of similarity thresholds is widely used. For instance see (Orduña Cabrera, 2016).

In (Tan et al., 2010), authors implement a stochastic method where the classification process and building of the clusters are the aim. This is done without human interaction. They use the time variable to evaluate the probability of belonging to some cluster. In our proposed method, we use a time variable to indicate the time of the acquired data. Both methods learn cases and store it, but our approach differs from theirs in the method of learning the cases. In our case, the proposal is focused on the representative prototypes. In addition, the learning of cases in our proposal is conditioned to accomplish with the maximal acceptable dispersion.

Air quality assessment is important for air pollution control and environmental management. Air quality is an important concern over the world, and especially in urban areas. Local city authorities and governments are responsible for the continuous monitoring and improvement of the urban air quality. There are several contributions in the study of the air pollution. These contributions study different aspects related to the air contamination. Some of this research works are (Halonen et al., 2014; Costabile and Allegrini, 2007). In (Halonen et al., 2014) the effect of long-term exposure to traffic pollution is studied. They study the associations between traffic pollution and emergency hospital admissions for cardio-respiratory diseases. In (Costabile and Allegrini, 2007) the study aims to analyze the relationship between air quality and air pollution from transport, and they develop a framework to understand its relationship. The effect of the air pollution and the

Please cite this article in press as: Orduña Cabrera, F., Sànchez-Marrè, M., Environmental data stream mining through a case-based stochastic learning approach, Environmental Modelling & Software (2018), https://doi.org/10.1016/j.envsoft.2018.01.017