Expert Systems with Applications 41 (2014) 5832–5842

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Improving data partition schemes in Smart Grids via clustering data streams

Andreu Sancho-Asensio [a,*], Joan Navarro [b], Itziar Arrieta-Salinas [c], José Enrique Armendáriz-Íñigo [c], Virginia Jiménez-Ruano [c], Agustín Zaballos [b], Elisabet Golobardes [a]

[a] Grup de Recerca en Sistemes Intel·ligents, Ramon Llull University, Quatre Camins 2, 08022 Barcelona, Spain
[b] Grup de Recerca en Internet Technologies & Storage, Ramon Llull University, Quatre Camins 2, 08022 Barcelona, Spain
[c] Departamento de Ingeniería Matemática e Informática, Universidad Pública de Navarra, Campus de Arrosadía, 31006 Pamplona, Spain

ARTICLE INFO

Keywords:
Smart Grids
Data partitions
Online learning
Clustering data streams
Learning classifier systems

ABSTRACT

Data mining techniques are traditionally divided into two distinct disciplines depending on the task to be performed by the algorithm: supervised learning and unsupervised learning. While the former aims at making accurate predictions after deeming an underlying structure in data—which requires the presence of a *teacher* during the learning phase—the latter aims at discovering regular-occurring patterns beneath the data without making any *a priori* assumptions concerning their underlying structure. The *pure* supervised model can construct a very accurate predictive model from data streams. However, in many real-world problems this paradigm may be ill-suited due to (1) the dearth of training examples and (2) the costs of labeling the required information to train the system. A sound use case of this concern is found when defining data replication and partitioning policies to store data emerged in the Smart Grids domain in order to adapt electric networks to current application demands (e.g., real time consumption, network self adapting). As opposed to classic electrical architectures, Smart Grids encompass a fully distributed scheme with several diverse data generation sources. Current data storage and replication systems fail at both coping with such overwhelming amount of heterogeneous data and at satisfying the stringent requirements posed by this technology (i.e., dynamic nature of the physical resources, continuous flow of information and autonomous behavior demands). The purpose of this paper is to apply unsupervised learning techniques to enhance the performance of data storage in Smart Grids. More specifically we have improved the eXtended Classifier System for Clustering (XCSc) algorithm to present a hybrid system that mixes data replication and partitioning policies by means of an online clustering approach. Conducted experiments show that the proposed system outperforms previous proposals and truly fits with the Smart Grid premises.

1. Introduction

Power electric distribution and transport networks belong to a deeply established but poorly evolved market (Gungor et al., 2011) that fails to provide advanced functionalities (Gungor et al., 2013; Rusitschka, Eger, & Gerdes, 2010) to both consumers and producers (also referred to as *prosumers*). According to the latest European directives, this situation must change radically in order to meet new standards concerning energy efficiency and sustainability (i.e., reducing greenhouse gas emissions, promoting energy security, fostering technological development and innovation, and limiting the amount of imported energy). In this regard, a new form of energy delivery system, coined as Smart Grid, has emerged as an alternative to handling new-generation power grid functionalities, which include real-time consumption monitoring, network self-healing, advanced metering infrastructure, or overload detection (Gungor et al., 2013).

To successfully migrate from the traditional centralized power delivery infrastructures to the distributed nature required by Smart Grids, several disciplines must be integrated (Monti & Ponci, 2010; Yan, Qian, Sharif, & Tipper, 2013): communication networks—to enable interactions between all the grid devices—cyber security—to ensure that the system is safely operated—and distributed data storage—to effectively deal with the data generated by this novel scheme (Navarro, Zaballos, Sancho-Asensio, Ravera, & Armendáriz-Iñigo, 2013). While the interaction of these

* Corresponding author. Tel.: +34 932902484.
E-mail addresses: andreus@salleurl.edu (A. Sancho-Asensio), jnavarro@salleurl.edu (J. Navarro), itziar.arrieta@unavarra.es (I. Arrieta-Salinas), enrique.armendariz@unavarra.es (J.E. Armendáriz-Íñigo), virginiaj@salleurl.edu (V. Jiménez-Ruano), zaballos@salleurl.edu (A. Zaballos), elisabet@salleurl.edu (E. Golobardes).

http://dx.doi.org/10.1016/j.eswa.2014.03.035
0957-4174/

disciplines is being explored under the context of some European-funded research projects (Gungor et al., 2013; Navarro et al., 2013), there is still not a mature framework to manage the vast amount of heterogeneous data generated by the ever-growing number of devices that populate the Smart Grid (Navarro, Armendáriz-Iñigo, & Climent, 2011; Rusitschka et al., 2010). In fact, there are several factors that prevent highly scalable (cloud) data repositories from meeting the stringent requirements of Smart Grids (Yan et al., 2013): (1) limited communication facilities in terms of delay (Selga, Zaballos, & Navarro, 2013), (2) massive amount of data streams (22 GB per day) that need to be effectively processed (Gungor et al., 2013; Rusitschka et al., 2010), (3) limited computing resources at smart devices (Navarro et al., 2013), and (4) the inability to store big sets of structured data (Stonebraker & Hong, 2012).

When addressing data storage and replication in the context of Smart Grids, practitioners are forced to select a convenient tradeoff between data consistency, data availability, and network partitioning as stated by the CAP theorem (Brewer, 2012). An effective way to deal with the CAP theorem in the specific context of Smart Grids consists in establishing a proper data partitioning layout (i.e., classifying and confining data in "logical islands" to limit the replication depth while considering the requests' locality) in order to keep data availability and consistency while leveraging system scalability. Hence, every data island (known as data partition) is able to scale up independently and meet the system requirements at every moment with a small penalty on other partitions. Therefore, initial approaches to address these concerns in distributed storage aim at replicating data in a partitioned scheme (Curino, Zhang, Jones, & Madden, 2010; Navarro et al., 2011). In fact, designing the optimal data partition configuration according to the demands posed by smart applications and the nature of these data has emerged as a hot research topic.

In this context, real-world industrial applications generate large amounts of data that are complex, ill-defined, unstructured and which contain hidden information that is potentially useful and exploitable for strategic business decision making (Antonelli et al., 2013; Orriols-Puig, Martínez-López, Casillas, & Lee, 2013). Building useful model representations from these data is a task that requires the use of unsupervised learning because this paradigm does not assume any *a priori* structure in the input information. This approach is based on algorithms that automatically explore data in order to uncover subjacent patterns. A cornerstone of unsupervised learning is found in data clustering, which consists in grouping examples into sets—the clusters—according to some proximity criterion with the aim of searching for hidden patterns (García-Piquer, 2012; Khan & Ahmad, 2013; Legara, Monterola, & David, 2013). The concept of making clusters out of input data is extrapolated to data streams in the field of *clustering data streams*.

The purpose of this paper is to introduce an online unsupervised Michigan-style LCS algorithm in order to build a data partitioning scheme that adapts itself to the specific needs of the Smart Grid. More specifically, this system is targeted at analyzing the data streams generated by smart devices in order to build a set of clusters, each containing the most frequently retrieved datum patterns in order to minimize the amount of accesses to multiple sites. This valuable information is used by the data replication protocol of the Smart Grid when selecting the proper device to place every datum. With this novel approach, the data management system of the Smart Grid is able to build a dynamic set of partitions and, thus, minimize the overhead associated to data movement over a distributed system (Brewer, 2012).

1.1. Paper contributions

The contributions of this work are the following:

- It explores the problem of data replication and partitioning policies under the Smart Grid domain.
- It goes beyond the state-of-the-art definition of XCSc, an online unsupervised Michigan-style LCS algorithm, and details how it has been modified to make it suitable for online domains.
- It details the online and evolving behavior capacities of this enhanced version of XCSc.
- It demonstrates the competitive behavior of this new approach by conducting a series of experiments on (1) a classic data stream synthetic environment, (2) an extended data stream synthetic environment with evolving components, and (3) a realistic scenario using the standard benchmarks proposed by the Yahoo! Cloud Serving Benchmark (YCSB).
- It encourages practitioners to use this system to address the problem of implementing data partitioning policies in the Smart Grid's storage layer.

1.2. Paper organization

The remainder of this work is organized as follows. Section 2 describes the related work in Smart Grids and in clustering data streams, stressing the critical concerns of data storage, replication and partitioning in the specific context of Smart Grids. Section 3 presents the deployed system architecture to handle the aforementioned requirements. Section 4 depicts the obtained results with our approach and compares them with previous work. Finally, Section 5 concludes the paper and outlines some future research lines.

2. Related work

Data play a key role in the Information and Communication Technology infrastructure that supports the Smart Grid (Navarro et al., 2013). In fact, smart functions that provide advanced functionalities on the electric domain rely on the quality and richness of the gathered information. In addition, these collected data are aggregated and computed at very geographically distinct points with different requirements—according to every smart function constraint (Gungor et al., 2013)—in terms of consistency, availability and network partitioning, which drives distributed systems practitioners into a critical compromise referred to as CAP theorem (Brewer, 2012).

Classic relational databases fail at finding an optimal trade-off between these features while providing scalable solutions (Brewer, 2012). Therefore, the latest cloud-based repositories, devoted to addressing the storage challenges of what has been recently coined as Big Data (Stonebraker & Hong, 2012), have relaxed the relational properties (by relying on key-value stores) and consistency constraints (using weak consistency models such as eventual consistency) of data in order to build highly scalable storage systems (Gulisano, Jiménez-Peris, Patiño-Martínez, Soriente, & Valduriez, 2012). Although this strategy has been successfully used to cope with the ever-growing amounts of data generated by several real-world applications (e.g., Twitter, Facebook, Google), preliminary results obtained in the Smart Grid domain (Rusitschka et al., 2010) are far from acceptable (Yan et al., 2013). Indeed, it has been shown that these general purpose storage repositories are unable to handle the specificities of Smart Grids in terms of variable consistency (Gungor et al., 2013), response time (Selga et al., 2013), and resource adaptability (Navarro et al., 2011).

2.1. Data partitioning concerns and Smart Grids' storage layer

In order to (1) meet the aforesaid storage requirements, (2) overcome scalability limitations of fully replicated solutions, and (3) exploit data locality, the construction of a proper data partitioning scheme has become necessary (Curino et al., 2010), especially in