Environmental Modelling & Software xxx (2018) 1–12

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

Hybrid SOM+k-Means clustering to improve planning, operation and management in water distribution systems

Bruno Brentan [a, *], Gustavo Meirelles [b], Edevar Luvizotto Jr. [b], Joaquin Izquierdo [c]

[a] Centre de Recherche en Controle et Automatique de Nancy, 2 rue Jean L'amour, Vandoeuvre-les-Nancy, France
[b] Laboratory of Computational Hydraulics, School of Civil Engineering, Architecture and Urban Design, University of Campinas, Campinas, Brazil
[c] FluIng, Universitat Politècnica de València, Valencia, Spain

ARTICLE INFO

Article history:
Received 17 February 2017
Received in revised form
30 August 2017
Accepted 14 February 2018
Available online xxx

Keywords:
Water supply systems
Classification
Self-organizing maps
k-means clustering

ABSTRACT

With the advance of new technologies and emergence of the concept of the smart city, there has been a dramatic increase in available information. Water distribution systems (WDSs) in which databases can be updated every few minutes are no exception. Suitable techniques to evaluate available information and produce optimized responses are necessary for planning, operation, and management. This can help identify critical characteristics, such as leakage patterns, pipes to be replaced, and other features. This paper presents a clustering method based on self-organizing maps coupled with k-means algorithms to achieve groups that can be easily labeled and used for WDS decision-making. Three case-studies are presented, namely a classification of Brazilian cities in terms of their water utilities; district metered area creation to improve pressure control; and transient pressure signal analysis to identify burst pipes. In the three cases, this hybrid technique produces excellent results.

1. Introduction

To achieve sustainable development, cities should be able to plan, operate, and manage their infrastructure efficiently. With the growth of cities and scarcity of environmental resources, management must be re-thought to guarantee access to quality urban services for all citizens (Chourabi et al., 2012). Operations must be fine-tuned to this same purpose. Furthermore, planning of future government actions for urban water, such as investments to improve quality of water systems, expansion of existing systems, or reduction in energy consumption, should prioritize those cities with the worse indicators.

Many cities are unable to satisfy the needs of their populations, and smart cities respond with an intensive use of technology, information, and data to improve quality in infrastructure services. Suitably handled, this new wealth will create an ideal scenario for economic growth with a higher quality of life for citizens and less damage to the environment (Kramers et al., 2014). To achieve this goal, it is necessary to suitably handle the available information about the systems, which is based on a high level of data acquisition, and thus create large databases. In water distribution systems (WDSs), these databases can integrate such variables as flow, pressure, water demand, pipe characteristics, water quality, climatic parameters, maintenance history, etc. Due to the size of the databases, computational tools are necessary to quickly identify problems and propose solutions.

Dividing the available information into clusters is a mechanism that enables identifying patterns and the main features in databases. Among several other clustering techniques, self-organizing maps (SOMs), Kohonen (1982), based on the theory of neural networks (NNs), have gained space in environmental resource research. Kohonen (2000) highlights the use of SOMs as a clustering tool for database operation. Izquierdo et al. (2016) uses SOMs for early data labeling for the application of classification tools. Kalteh et al. (2008) highlight its extensive use in water resources problems. More specifically, in WDSs, SOMs have been widely used for water quality analysis (Blokker et al., 2016), optimal design (Norouzi and Rakhshandehroo, 2011), and pattern analysis (Laspidou et al., 2015), among other uses.

The application of data clustering in WDSs can help planning, operation, and management of these systems, since it is possible to identify anomalies, develop strategic operations for distribution, and evaluate the system through suitable indicators. In addition, in some cases, this classification enables the determination of

* Corresponding author.
E-mail address: brunocivil08@gmail.com (B. Brentan).

https://doi.org/10.1016/j.envsoft.2018.02.013
1364-8152/

Please cite this article in press as: Brentan, B., et al., Hybrid SOM+k-Means clustering to improve planning, operation and management in water distribution systems, Environmental Modelling & Software (2018), https://doi.org/10.1016/j.envsoft.2018.02.013

interesting benchmarks which can be useful to establish quality standards and goals. We consider three situations, concisely introduced in the following three paragraphs.

Cabrera et al. (2013) and Lima et al. (2015) present different proposals for WDS classification based on the energy consumption of pumping stations to create goals for systems with poor performances. Berg and Lin (2007), Thanassoulis (2000), and Scaratti et al. (2013) use data envelopment analysis (DEA) to classify cities in Peru, the United Kingdom, and Brazil respectively, in relation to the management of sanitation services, and identifying places where investment is urgently needed.

Clustering applied to WDSs can also be useful to propose strategic divisions of the entire network into manageable pieces, called district metered areas (DMAs). Considering the large extension of systems and their complex interconnections, the process to divide WDSs into DMAs can be a difficult task and result in poor scenarios when partitions are developed without appropriate tools. Campbell et al. (2016a,b) use social community algorithms to divide WDSs into DMAs to improve pressure control and reduce leakage. Herrera et al. (2010) apply graph theory to create DMAs to improve network resilience.

Also, operational data, such as pressure or flow signals, can also be grouped to create specific strategies to suitably operate the system. Pressure signals can be used to identify burst pipes and locate anomalies in WDSs (Srirangarajan et al., 2010; Covas and Ramos, 2010). Taking one of the most important challenges for WDSs, Aksela et al. (2009) present a methodology based on flow signal analysis with SOMs to detect leaks in WDSs.

The use of NNs and machine learning tools enables pattern extraction and classification of large databases. However, the number of resulting groups is usually too large. Classification methodologies for planning, operation, and management often require a small number of groups to ease and accelerate decision-making. SOMs are unable to create this reduced predefined number of groups, which is a drawback for their application in these cases.

When the number of groups can be predefined, the use of classical clustering methods, such as k-means algorithms, has been widely explored, since these methodologies can be faster than NN approaches. However, NN feature extraction combined with fast processing by k-means can be an interesting way to handle big data.

This paper proposes a classification method using SOMs as a pre-processing technique, coupled with the k-means algorithm to achieve groups that can easily be labeled and used in WDS decision-making. To exemplify the use of this methodology, three case studies are addressed. In the first, Brazilian cities are classified with respect to their water quality, operational efficiency, and economic performance, and this information can help the government plan investments to raise water quality. The second study applies the hybrid model to generate a scenario of DMAs in a fictitious water distribution network (WDN), C-town, which is frequently used as a benchmarking problem. This application enables water utilities to improve control efficiency, mainly for pressure management (which is closely related to water losses). Finally, in the third case, hydraulic transient pressure signals generated by pipe bursts and ruptures are processed with the proposed hybrid SOM + k-means algorithm to extract patterns and features in the data. This can help water companies to deal with anomalous events in WDSs, since knowledge of event groups can be associated with specific strategies to quickly solve the problems.

This paper is a substantial extension of Brentan et al., (2016), in which only a simplified version of the first case study was considered.

2. Clustering process

2.1. Self-organizing maps (SOMs)

Based on brain behavior under visual and memory stimulation, SOMs are a type of neural network with unsupervised training as proposed by Kohonen (1982). When stimulated, different regions of the brain can react according to the pattern of the stimulus. This behavior enables separating different stimuli and triggering more efficient reactions. With this inspiration, an SOM is a tool to process input data and find patterns to group similar data.

SOMs have been applied in many research fields mainly because of their ability to learn from high-dimensional input data, resulting in a low-dimensional (usually bi-dimensional) output layer (Kohonen, 2000). This property helps the visualization of topological correlations among data and leads to better understanding of a problem.

A competitive learning process is responsible for the training of SOMs. Basically, this learning process is made of three steps: competition; cooperation; and synaptic update. The competition stage is responsible for identifying the map region most activated by a certain input data. Such a region can be defined by the neighborhood of the most activated neuron, the so-called winning neuron. In the cooperation stage, the influence of the winning neuron on its neighborhood is determined. Finally, in the synaptic update stage, the winning neuron increases its weight value, leading to new and similar input data to activate the neuron again.

The iterative process of mesh adjustment to represent the feature space requires that all samples are presented to the network; in each presentation the most excited region of the mesh is adjusted. This region can be identified through the neuron exhibiting the greatest similarity between an input datum $\boldsymbol{x}$ and a neuron $\boldsymbol{w}_j$. This means the minimal distance between $\boldsymbol{x}$ and a neuron $\boldsymbol{w}_j$ selects this neuron as the most excited and initiates the process of mesh adjustment. The Euclidian distance has been widely applied to determine this similarity.

As suggested by the biological inspiration, the activation of the winner also defines a topological neighborhood that will be excited. The closer to the winning neuron, the more excited will be its neighborhood by the input $\boldsymbol{x}$. The activation power of a neuron inside the neighborhood may be written as a monotonically decaying function, $h_{j,i(\boldsymbol{x})}$, centered on the winning neuron $i(\boldsymbol{x})$ and containing the set of j neurons excited by such a winner.

The influence on the neighborhood topology reduces with time to make the model more realistic. This means that from one iteration to the next, the influence of the winning neuron decreases, making the weight adjustment more stable. After completing an iteration, each weight (neuron position) is updated with its corresponding increment $\Delta\mathbf{w}_j$ defined by (1):

$$\Delta\mathbf{w}_j = \eta h_{j,i(\boldsymbol{x})} \left(\mathbf{x} - \mathbf{w}_j\right) \tag{1}$$

where η is the forgetfulness rate, which represents the human-like learning process.

Finally, the updating for each time step n can be written as:

$$\mathbf{w}_j(n+1) = \mathbf{w}_j(n) + \eta(n) h_{j,i(\boldsymbol{x})} \left(\mathbf{x} - \mathbf{w}_j(n)\right) \tag{2}$$

The learning process finishes when either the mesh update is smaller than a pre-defined threshold, or the number of iterations is reached. At this point, the mesh can represent the feature space in a lower dimension than that of the original space and some correlations between input data dimension can be observed. Furthermore, the input data can be classified by the nearest neuron, thus forming a set of clusters.