A multiple-layer clustering method for real-time decision support in a water distribution system

Alexandru Predescu[1], Cătălin Negru[1], Mariana Mocanu[1], Ciprian Lupu[1], and Antonio Candelieri[2]

[1] Department of Computer Science, University POLITEHNICA of Bucharest
[2] Department of Computer Science, University of Milano Bicocca

Abstract. Machine learning provides a foundation for a new paradigm where the facilities of computing extend to the level of cognitive abilities in the form of decision support systems. In the area of water distribution systems, there is an increased demand in data processing capabilities as smart meters are being installed providing large amounts of data. In this paper, a method for multiple-layer data processing is defined for prioritizing pipe replacements in a water distribution system. The identified patterns provide relevant information for calculating the associated priorities as part of a real-time decision support system. A modular architecture provides insights at different levels and can be extended to form a network of networks. The proposed clustering method is compared to a single clustering of aggregated data in terms of the overall accuracy.

Keywords: Decision Support System · Machine Learning · Multiple-Layer Clustering

1 Introduction

Nowadays, smart cities use Information and Communications Technology (ICT) systems for management of water resources. There are some challenges that need to be overcome, such as efficient water exploitation, elimination of water loss because of broken pipes and ensuring good water quality. Also, these ICT systems gather data about water production, distribution and consumption, aiming to optimize different stages of water cycle and to inform all the actors implies (e.g. operators, city services providers and citizens) [1].

In the ever expanding large scale water distribution systems, the task of the operator has become increasingly difficult and most commonly there is little insight on the system as a whole. Therefore, a decision support system is useful for increasing the efficiency and handling events based on priorities and cost effectiveness. A typical situation where Machine Learning algorithms are used is the classification of consumers based on their behavior.

Water resource management for complex networks is a subject of ongoing research. This includes the estimation of water demand from smart meter data, which allows for efficient planning and detection of anomalies, as described in [2].

There are also described solutions which allow for handling the vast amount of data in such systems. The k-means algorithm is used for data clustering in the context of unsupervised learning. The demand patterns of different consumer types can be extracted from AMR (Automatic Meter Reading) meters which provide 24-hour time series, as described in [3].

The paper is organized as follows: In section 2 we present the context of research in the area of Machine Learning and defined strategies for increasing the efficiency of resource spending in water distribution systems. In section 3 we propose a method for integrating modern technologies and results from the field of Machine Learning into a decision support system for prioritizing maintenance works and repairs, defining a method for data clustering, priority evaluation and implementation. In section 4 we use the proposed method on a data set provided by measurements and we present the results in terms of accuracy and the level of detail that is obtained by a multiple-layer data clustering architecture, in comparison with a single layer architecture. The effect of an additional consumer on the existing clusters is also evaluated.

2 Related work

Machine Learning (ML) is the state-of-the-art for Artificial Intelligence (AI), based around the idea that the machines should have access to and learn from available data. It is currently being used in a broad range of applications to extract relevant information in fields such as Internet of things (IoT), where wireless sensor networks (WSN) provide increasingly large amounts of data. The algorithms and strategies for addressing functional and non-functional requirements for this domain are described in [4]. Data clustering and aggregation is one of the main functional requirements for data processing in a WSN. Outlier detection is defined as a performance enhancement, though it can be the main focus in some applications.

Clustering methods have been applied in domains such as energetics for finding the energy consumption patterns [5], finance for evaluating the financial trends [6], medicine for detecting brain activity [7], robotics for improving autonomous capabilities [8] and psychology for analyzing user behavior in the context of social networks [9].

For water distribution systems, a proactive strategy allows for increasing the efficiency of resource spending for pipe renewal by addressing the most important assets first, as stated in [10]. A proactive strategy is recommended for low probability and high consequence situations, while a reactive strategy implies analyzing the breakdown history and scheduling the repair accordingly and is often used for high probability and low consequence situations.

In [11], by using the k-means algorithm and the results of multiple clusterings, it is possible to accumulate evidence in the context of unsupervised learning. In this paper we study a multiple clustering strategy by defining multiple layers, each providing additional information about the entire data set, in a proactive approach to water network management.

3 Proposed solution

We propose the integration of a multiple-layer clustering algorithm into a decision support system for prioritizing pipe replacements in a water distribution system. The priorities associated to the network segments are defined as a function of the calculated daily demand patterns and the consumer types. There are different types of consumers such as residential, commercial and industrial, each having different requirements for water and service quality [3]. In our proposed algorithm, each consumer node is assigned a priority. With pipes being the scope of most repair works, the associated priorities have to be calculated. Considering a tree structure and using a bottom-up approach, each parent node is assigned the average priority of its child nodes. The pipe priority is defined as the average priority of its end nodes. We propose two questions related to the integration of this method into a real-time decision support system that we further address in section 4:

(i) **Network reconfiguration**
The first question is related to the effect of adding a new consumer to the network on the associated priorities (e.g. by installing an additional smart meter providing data for the particular node). When comparing this scenario to the ideal case where all the data is known in advance, the discrepancy between the two scenarios is a measure of the overall accuracy of the method in real-life operating conditions.

(ii) **Comparison and parameter optimization**
The second question is related to the advantages and disadvantages of using a multiple-layer clustering method over a single-layer clustering method and the effect on the overall accuracy correlated with the number of clusters.

3.1 Data clustering method

For identifying the consumer patterns from measured data, unsupervised learning algorithms are used. We considered a two-layer clustering method which finds patterns for each consumer and then for the entire network, as shown in Fig. 1. This method allows for distributed processing of consumer patterns for the first layer, while reducing the bandwidth required for the second. An extension to n-layer clustering, where the system can be divided into subsystems, is a subject of a future paper. The following steps outline the proposed clustering method:

(i) **Individual consumer clustering**
The clustering algorithm is run for each individual consumer node, to obtain the daily consumption pattern over the entire time frame. This allows for an accurate visual representation of the consumer behavior for the particular node. The optimal number of clusters for each consumer node is determined using a silhouette analysis on k-means clustering, as having a silhouette coefficient that indicate a lower degree of overlapping with the other clusters [12].

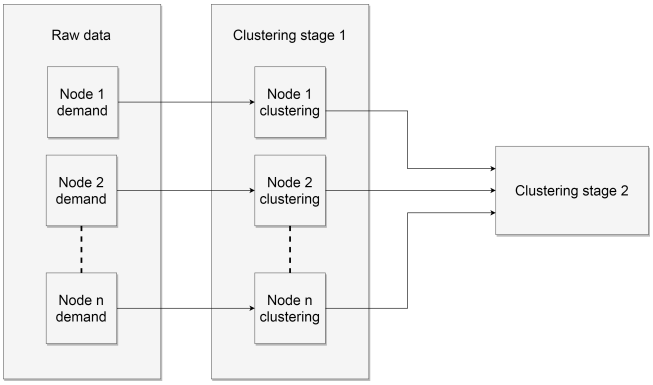

Fig. 1. Dual-layer clustering method

(ii) **Consumer group clustering**
The cluster centroids from the individual consumer clustering are used as the input to the second clustering algorithm. The result of this second-layer represents the consumer demand patterns for the network/network section. The method is the same as for individual consumer clustering, with the additional specification of the number of clusters that should emphasize the patterns for the consumer group. The requirement is that the number of clusters for the higher layer is not higher than the number of aggregated clusters from a particular layer. The effect is that the main characteristic of the current layer is used for data processing for the higher layers. Therefore, the data can be processed in a more organized way, similar to the Hierarchical Clustering method.

(iii) **Priority evaluation**
Each time series corresponding to the particular node is assigned to one of the identified patterns for the entire network. The assignment of individual consumers to a cluster is evaluated by finding the most frequent assignment of the associated time series. Therefore, the consumer priority is selected based on the most relevant assignment that results from the measured data.

As a result, the main consumer categories can be defined as residential consumers with peaks in the morning and in the evening, commercial consumers with higher demand during the business hours and industrial consumers which can have either a uniform demand during the day or an irregular pattern in the case of irrigation systems. In a supervised learning strategy, these characteristics can be defined and compared to the identified clusters, using a similarity measure such as the euclidean distance [13].

For multiple layers, the clustering method is similar. Each layer would provide patterns for the lower layers. The consumer layer is considered as the first layer, showing the individual consumer patterns.

3.2 Priority assignment method

The network is represented as a directed graph. For the sake of simplicity, we considered a single source (i.e. pumping station) as the root node, consumers as leaf nodes and intermediary nodes. After the individual consumers are clustered using the method described in section 3.1, the associated priorities can be calculated as a function of the consumer type and demand profile.

The final objective is to assign priorities to network sections in order to provide decision support for the maintenance operations. For this, we define an indirect method, using the previously calculated priorities for consumer nodes, where the priority of the segment is calculated as the average priority of the connected nodes. This assumes that the priority of the intermediary nodes has to be calculated, in this case as the average priority of the adjacent consumer nodes. For this, a breadth-first graph traversal is used and the priorities are calculated in reverse order, from the leaf nodes up to the root node. For the n-layer clustering, the node priorities are calculated for each layer, using the data from the lower layers. Each layer should be able to operate either as an autonomous entity or as part of a higher layer decision system. The priority associated to a layer can be defined according to the specific requirements (e.g. location, average priorities from lower layers). The edge priority is then calculated in a similar way, as described in the two-layer scenario.

3.3 Python implementation

The application is built with a client-server architecture. The Python API handles the data and algorithms, and a SPA (Single Page Application) shows the user interface for the proposed decision support system. The data exchange is implemented using HTTP requests and the MVC architecture allows for separation of concerns. The Python language is used, as there is a vast ecosystem of open-source libraries which can be used for scientific purpose. The first step is loading the data from CSV files into a *numpy* array that is used in many libraries requiring data processing capabilities, such as *numpy*, *scipy* and *matplotlib*. The *scikit-learn* package provides tools for Machine Learning applications and implements some of the most common algorithms in this field [12]. For unsupervised learning, the *KMeans* class from the *sklearn.cluster* package implements the general clustering algorithm that can be used for a broad range of applications. The algorithm clusters data with the specified parameters e.g. number of clusters, maximum number of iterations, and allows the initialization of centroids with previous values. The results (e.g. centroids, assignments) can be extracted from the returned object. The method *predict* returns the assignment of the data samples to the closest corresponding cluster, without modifying the centroids. This is useful for assignment of additional data to a cluster as it allows using the training data set, as well as additional data. In the case of new samples that have to be added to a cluster, the algorithm can be initialized with the previous centroids and ensures the same order when updating the clusters. For graph representation and related algorithms, the

networkx package is used. The network is defined in JSON format, suitable for data exchange with a GIS (Geographic Information System) server. The measurements are assigned to the corresponding network node and the priorities for the other nodes and segments are calculated, from the consumer layer to the upper layers.

4 Results

The data that we used in this paper is obtained using the smart meters from the urban water distribution system of Milan. The data is stored in CSV files corresponding to each node. Each file contains daily 24-hour time series for individual consumers having a sampling time of 1 hour over a time-frame of several months (September-December 2014). The structure of the network is not required for the experiments presented in this paper and an experimental model is used instead.

First, we obtain the consumer patterns from the data set for individual consumers. This initial clustering reveals some typical consumer patterns. For example, in Fig. 2, the characteristic shows two peaks, corresponding to the average residential consumer pattern and in Fig. 3, the pattern is that of an industrial consumer with a single peak during the working hours.

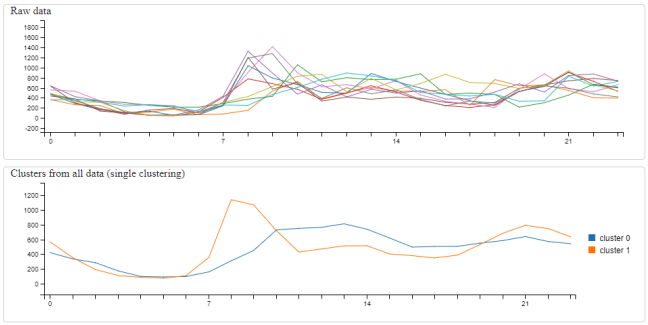

Fig. 2. Residential consumer pattern

Then, using the second-layer clustering, the general consumer patterns are obtained and the individual consumers are assigned to one of the identified patterns to calculate the associated priority. For an effective visual overview, the network nodes and segments and the two-layer clustering results are shown as a graph, having a colormap representation of the associated priorities. The higher priorities are shown in red and the lower priorities are shown in blue. The size of the element is also proportional to the priority. In Fig. 4, a simple network is shown with the results obtained from the available data set. When a node is selected, the time series and cluster centroids are shown to the operator.

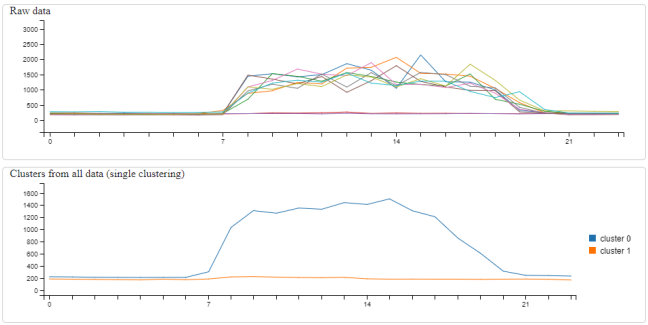

Fig. 3. Industrial consumer pattern

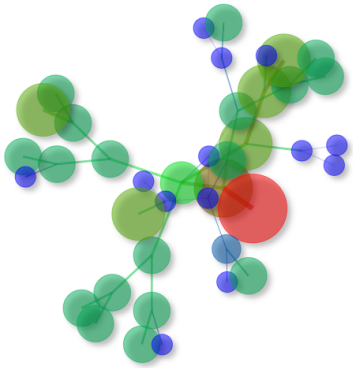

Fig. 4. Overview

Therefore, we defined a method for priority evaluation using clustering algorithms as part of a visual decision support system. There are some aspects that have to be clarified regarding the accuracy and reliability of the proposed method in real-life scenarios.

4.1 Network reconfiguration

For answering the first question that we proposed in section 3, we analyze the partial recalculation (update) of the clusters when a consumer node is added to the network configuration. In Fig. 5, the clusters are calculated from two residential consumers and in Fig. 6, the clusters are updated with the data from two additional residential consumers.

The cluster assignments and main characteristics are maintained between the two cases and the third cluster shows a significant change in terms of consumer demand. The other clusters present only a slight change from the initial clusters. This shows the possibility of using the clustering algorithm under dynamic conditions. In the case of incoming new time series on the same

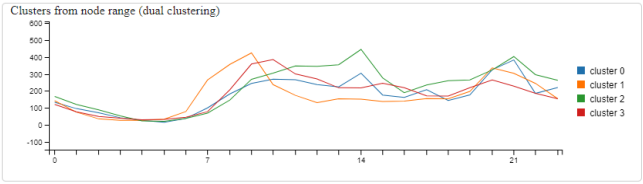

Fig. 5. Network reconfiguration (before)

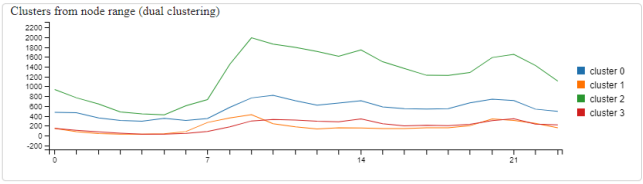

Fig. 6. Network reconfiguration (after)

network configuration, the assignment to one of the clusters does not require a full recalculation of centroids, as described in [14].

4.2 Comparison and parameter optimization

For evaluating the accuracy of the two-layer clustering method, the resulting centroids are compared to a control method. The control method implements a single-layer clustering using the data from each node as input. We consider the silhouette analysis to find the optimal number of clusters for the first-layer, as we are not interested in a specific number of individual consumer patterns. The second-layer clustering has a predefined number of clusters that represent the general patterns for the entire data set (e.g. residential consumers, industrial consumers).

The resulting centroids obtained using the optimal number of clusters for each node and 3 final clusters are shown in Fig. 7. The control method (green) provides a reference for testing the accuracy of the proposed method (blue).

For answering the second question that we proposed in section 3, we define an extensive evaluation of the overall deviation from the control method for the entire range of parameters. Instead of using the optimal number of clusters for the first-layer clustering, we evaluate the deviation over the entire range (2-82). The deviation is calculated as the euclidean norm of the individual cluster deviation.

$$x_i \in \mathbb{R}^N,\ i = 1 \dots NC \tag{1}$$

$$y_i \in \mathbb{R}^N,\ i = 1 \dots NC \tag{2}$$

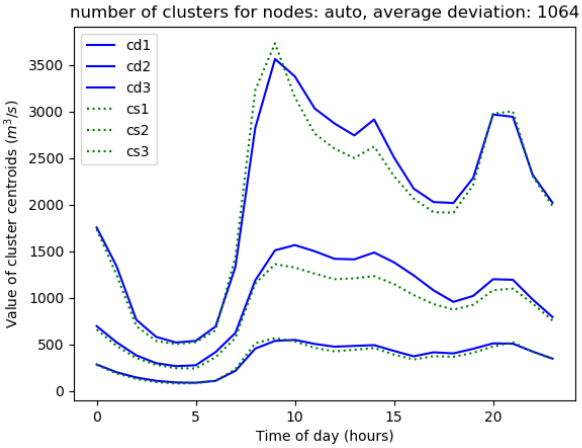

Fig. 7. Dual-layer with optimal number of individual clusters

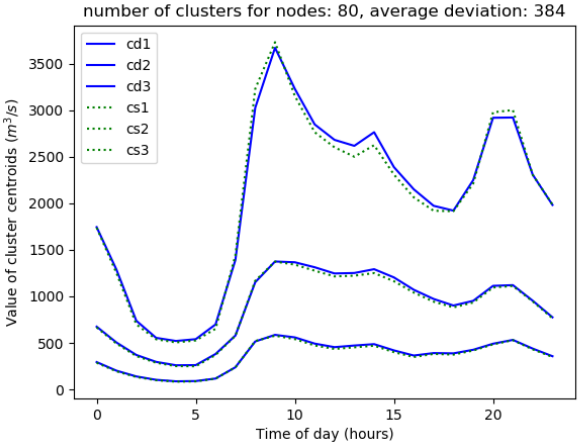

Fig. 8. Dual-layer with highest number of individual clusters

$$\delta_i = \sqrt{\sum_{j=1}^{N}(x_{i,j} - y_{i,j})^2} \in \mathbb{R},\ i = 1 \dots NC \tag{3}$$

$$\|\delta\| = \sqrt{\sum_{i=1}^{NC}\delta_i^2} \in \mathbb{R} \tag{4}$$

The formulas describe the method for calculating the overall deviation between the two methods, where $N = 24$ is the dimension of each cluster centroid and $NC = 3$ is the number of clusters used for the second layer, x_i represents a cluster centroid obtained using the control method and y_i represents a cluster centroid obtained using the two-layer clustering, δ_i represents the deviation between two individual cluster centroids.

The results are shown in Fig. 9. As expected, the deviation is inversely proportional to the number of clusters. The deviation of the results obtained with the optimal number of clusters for each node is shown in orange, and is generally higher than the method with fixed number of clusters. Therefore, we propose that for the second-layer clustering, a higher number of clusters should be used. Nonetheless, for analyzing the data at the current layer, the optimal number of clusters should be used, so that there is no redundant data for the decision system. By using the highest possible number of clusters (i.e. 82), we obtain the most similar resulting centroids which are shown in Fig. 8, using the same convention as in Fig. 7.

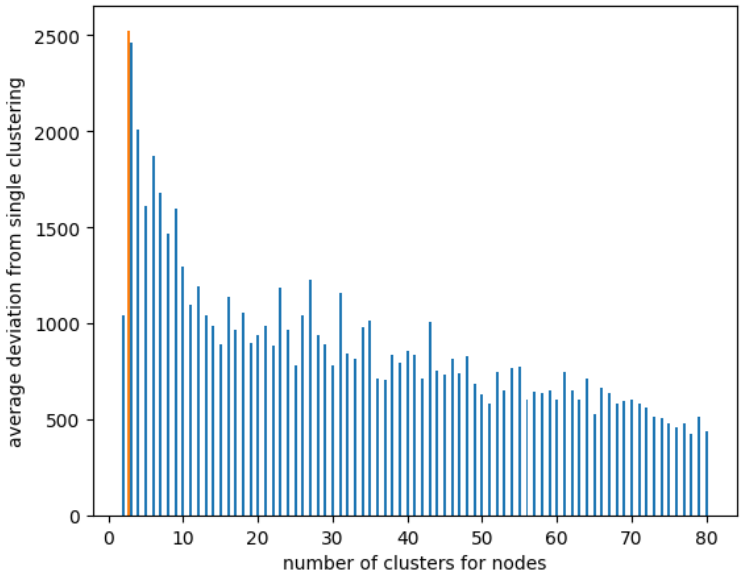

Fig. 9. Overall deviation

4.3 Overview

The final result of the clustering algorithm consists in the identified consumer patterns, that allow for an accurate classification of the consumers and calculating the associated priorities. In the case of a multiple-layer clustering, the identified patterns characterize a particular sub-network that can have an associated priority as well. The resulting decision support system is characterized by a similar organized architecture that allows for coordinating a large-scale water distribution system. With additional results presented in [14], the real-time requirements can range from a daily evaluation of priorities to a much more frequent evaluation, in sync with the actual sampling time of the data.

5 Conclusion

In this paper, Machine Learning algorithms are used as the basis of a decision support system for water distribution systems. Data clustering is used for consumer type identification and further data from consumers is then assigned to the identified clusters.

The two-layer clustering reveals patterns at the consumer level and for the network as a whole. The proposed decision support system uses the result of the clustering method to evaluate the priorities of network sections, for efficient maintenance operations in water distribution systems. This idea can be extended in the case of multiple sub-divisions of the network, using parallel computing and a decentralized architecture. The multiple sub-divisions and hierarchical model allow for an increased scalability of the proposed solution.

Assuming that there is enough data to create an accurate model, the method can be extended for real-time scenarios. The upcoming time series can be assigned to the previously identified clusters without recalculating the centroids, as long as the accuracy of the model is within the requirements.

Therefore, the possibilities given by newly emerging paradigms and open-source software provide multiple directions for improvement in the quality of service in areas such as utility networks and the particular case of water distribution systems.

Acknowledgement

We are thankful to the PN III Program P3 – European and International Cooperation, UEFISCDI, that supported the research activity and part of the presentation in conference, as well as to the H2020 Twinning Program, that partially supported the publication under the 690900 project - Data4Water

References

1. M. Umar and W. Uhl, *"Integrative Review of Decentralized and Local Water Management Concepts as Part of Smart Cities (LoWaSmart)"*, 2016, Norsk institutt for vannforskning
2. D. García, D. Gonzalez, J. Quevedo, V. Puig and J. Saludes *"Water demand estimation and outlier detection from smart meter data using classification and Big Data methods"*, Conference report 2015.
3. D. García, D. Gonzalez, J. Quevedo, V. Puig and J. Saludes, *"Clustering and classification of aggregated smart meter data to better understand how demand patterns relate to customer type"*, Universitat Politecnica de Catalunya (UPC), Conference report 2015
4. M. A. Alsheikh, S. Lin, D. Niyato and H. P. Tan, *"Machine Learning in Wireless Sensor Networks: Algorithms, Strategies, and Applications"*, IEEE Communications Surveys & Tutorials, vol. 16, no. 4, pp. 1996-2018, Fourthquarter 2014.
5. F. Iglesias and W. Kastner, *"Analysis of similarity measures in times series clustering for the discovery of building energy patterns"*, Energies, vol. 6, no. 2, 2013, pp. 579–597.
6. M. Kumar, N.R. Patel and J. Woo, *"Clustering Seasonality Patterns in the Presence of Errors"*, Proceedings of Eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 2002, pp. 557–563.
7. A. Wismüller, O. Lange, D.R. Dersch, G.L. Leinsinger, K. Hahn, B. Pütz and D. Auer, *"Cluster Analysis of Biomedical Image Time-Series"*, International Journal of Computer Vision, vol. 46, no. 2, 2002, pp. 103–128.
8. M. Ramoni, P. Sebastiani and P. Cohen, *"Multivariate Clustering by Dynamics"*, Proceedings of the national Conference on Artificial Intelligence, 2000, pp. 633–638.
9. J. Zhu, B. Wang and B. Wu, *"Social network users clustering based on multivariate time series of emotional behavior"*, The Journal of China Universities of Posts and Telecommunications, vol. 21, no. 2, 2014 pp. 21–31.
10. M. Moglia, S. Burn and S. Meddings, *"Decision support system for water pipeline renewal prioritisation"*, 2006, ITcon vol. 11, pp. 237-256
11. A. L. N. Fred and A. K. Jain, *"Data clustering using evidence accumulation"*, Object recognition supported by user interaction for service robots, 2002, pp. 276-280 vol.4.
12. scikit-learn developers (BSD License). *"Selecting the number of clusters with silhouette analysis on KMeans clustering"* [Online] Available: `http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html`
13. F. Iglesias * and W. Kastner, *"Analysis of Similarity Measures in Times Series Clustering for the Discovery of Building Energy Patterns"*, 2013, Energies, vol. 6, pp. 579-297
14. A. Predescu, C. Negru, M. Mocanu, C. Lupu *"Real-time clustering for priority evaluation in a water distribution system"*, AQTR2018, Cluj, Romania