Energy-Efficient Partitioning Clustering Algorithm for Wireless Sensor Network

Koffi V. C. Kevin de Souza[1], Catherine Almhana[1],
Philippe Fournier-Viger[2], and Jalal Almhana[1](✉)

[1] Université de Moncton, Moncton, NB, Canada
jalal.almhana@umoncton.ca
[2] Harbin Institute of Technology (Shenzhen), Shenzhen, GD, China
philfv8@yahoo.com

Abstract. Wireless Sensor Networks (WSNs) have recently achieved tremendous success at both research and industry levels. WSNs are currently implemented in many areas, such as the military, environmental monitoring, and medicine. WSN nodes are battery-operated, and energy saving is critical for their survival. Several research papers have been published on how to optimize power usage. In this paper, we focus on improving power consumption by optimizing data transfer. We propose an Energy-Efficient Partitioning Algorithm to reduce data transfer and consequently improve power consumption. Using data collected from a real WSN in the City of Moncton, we implemented and compared the performance of the proposed algorithm with another data reduction algorithm. Experimental results show that our algorithm outperforms a recent data reduction technique in terms of power saving.

Keywords: Energy saving · Clustering times series · Smart meters
Wireless Sensor Networks · Data transfer

1 Introduction

In recent years, WSNs have achieved tremendous success with several research projects have been carried out by the academic community and findings disseminated through various channels, namely conferences and journal [1, 2, 4]. Also, industry has been very active in making WSN reliable components available at a reasonable cost. Many applications are using WSNs as they are easy to deploy compared to wired systems and offer a reliable, flexible, and efficient means to gather information from variety of sensor nodes to monitor variables such as heat, light, and water levels. These sensing components are generating great interest and demand from the Internet of Things (IoT) industry, which can be viewed as an extension of the traditional WSN. Sensor nodes are generally low-power battery-operated with limited lifetime, and energy saving is critical for long-time autonomous WSN operation. Several research papers were published on energy savings. They cover a variety of WSN layers, including application, transport, network, data link, and physical layers [6]. In this paper, we are interested in power saving techniques at the application level which are more suitable for the system studied here, water meter data extraction. The main

C. Li and S. Mao (Eds.): WiCON 2017, LNICST 230, pp. 14–23, 2018.
https://doi.org/10.1007/978-3-319-90802-1_2

purpose of this system is to acquire enough data, through frequent readings, in order to detect major water leaks or Abnormal Water Consumption (AWC) and avoid water waste. Frequent Water Meter Readings (WMRs) are essential for timely AWC detection as remedy action can be taken at a very early stage of possible water leaks. However, frequent WMRs imply frequent Radio Data Transmissions (RDTs), and consequently more power draining from the sensor batteries.

Experimental results [1] showed that RDT takes a significant amount of energy when it is compared to processing, insofar as the power needed to transmit one bit is equivalent to that needed to process one thousand bits. Trading RDTs for processing seems to be profitable in terms of power saving, and this aspect is of particular interest in our study. Designing an efficient system which at the same time allows AWC detection and optimizes sensor battery usage is crucial for a reliable WSN application implementation.

A variety of techniques can be used for reducing the number of radio transmissions [7]. The most relevant ones fall into two categories: data driven approaches and data acquisition techniques. Data driven approaches include reduction schemes, data compression, and data prediction. Reduction schemes minimize the number of data transmissions by eliminating redundant or unnecessary data. Compression techniques send data, at the source node, in condensed format, which can be decompressed at the sink node. Prediction supposes that the sensed process can be modeled using time series or stochastic and algorithmic solutions. Data acquisition [7] can be achieved by several approaches: time driven, event driven, query based, and hybrid. The effectiveness of these approaches depends on the type and requirements of the application.

In this paper, we propose a new approach that combines a time driven technique and a data partitioning clustering method to reduce RDTs. To the best of our knowledge, it is the first time that data partitioning clustering is used for data transmission reduction. We implemented our approach on data collected over several months from the City of Moncton's WSN water meters which contains more than 20,000 Water Meter Nodes (WMNs). Experimental results showed that our approach offers better power consumption economy when compared to the data reduction algorithm proposed in [7].

In Sect. 2 related works are presented, in Sect. 3 we describe the case study, followed by our approach in Sect. 4. Sections 5 and 6 present experimental results and concluding remarks respectively.

2 Related Work

As previously mentioned, various power management techniques can be used in WSNs. Research and development in this area is substantial and broad, it is therefore neither possible nor appropriate to review all available power management techniques. Several approaches were proposed to be implemented at different layers, including application, transport, network, data link, mac, and physical layers [2, 3]. Taking into account the WMR application, and the architecture of the network we are studying specifically here, we are interested in those techniques implementable at the application level. In our case, we do not have access to other layers. We are mainly interested in strategies that allow to minimize the number of RDTs and preserve the main purpose of the application in terms of proper water consumption monitoring, which requires frequent WMRs. RDTs have an important impact on power saving; experimental results showed that they require much more power compared to instructions processing at the node's microcontroller. Therefore, trading RDT for data processing appears to be a very good strategy.

At the application layer, several techniques are described in the literature [9, 10]. We can classify them into three categories: data reduction, data compression, and data prediction. In data reduction [11, 12], redundant or unneeded data can be removed without any loss of information. In data compression [4], data is encoded using a variety of algorithms and transferred in condensed form and they are decompressed at the receiving node (sink node). Data prediction [7], is a more elaborate technique than the previous ones, and involves that a model is built based on the sensed data over a relatively long period of time. One copy of the model is stored at the sink node or on a server in certain cases, and another copy is stored at the sensor node. Data can be retrieved at the sink node from the model instead of from the measurement at the sensor node. In our previous work, we implemented both data reduction and prediction techniques on similar data we collected from the City of Moncton, and it proved to be effective.

Beyond the above techniques used to reduce power consumption, it is key to mention here the importance of data acquisition and how it is implemented in the whole network. The way data is extracted can define the amount of the data collected, and consequently impact the power consumption. Four procedures [6] are commonly used for data extraction: time-driven, event-driven, query-based, and hybrid.

In this paper, we will propose a new algorithm to reduce RDTs. It uses data partitioning and clustering and is a completely different approach from all of the above-mentioned techniques. To the extent of our knowledge, it is the first time that a clustering approach is used for data transmissions reduction. There are two methods for building clusters [13]; hierarchical and partitioning. The hierarchical method starts by grouping the most similar profiles at the lower level and the less similar ones at the higher level. Meanwhile, the partitioning method attempts to divide a large cluster into smaller ones. We will use the K-means algorithm [3] for data partitioning clustering.

K-means is a popular method for clusters analysis and data mining. A partition is made on a set of data (n observations) into K clusters. It assigns an observation to the cluster with the closest mean.

Given that WMRs can be seen as time series, we will use the Euclidean Distance to measure the similarity between two different WMRs. Let x and y be n-dimensional vectors [14, 15]. The Euclidean distance of similarity is given by the following formula: $dist(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$. As the number of possible clusters is unknown in our case, we need validity measures such as Sum of Squared Error (SSE) and Calinski-Harabasz (CH). All cluster validation measures are based on compactness and separation. In Euclidean spaces, compactness means the set of data is closed and bound. Separation determines how distinct or well separated clusters are. The SSE is computed as: $SSE = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$. SSE is the total sum of all sums of Euclidean distance between the center and every profile in every cluster. The optimal value is in an elbow, (see following section). The Calinski-Harabasz index evaluates the cluster validity based on the average between and within-cluster sum of squares [8].

3 A Case Study: WMRs in the City of Moncton

In this section, we will provide information about the application we are studying here, WSN WMRs in the City of Moncton.

3.1 Network Architecture [7]

Figure (1) shows the architecture of the sensor network in the City of Moncton, which includes approximately 20,000 geographically distributed nodes. The sink nodes are located in four towers, each with one square mile coverage, and data with timestamps is collected from each cluster node with some nodes reachable by more than one tower. Each meter node is equipped with a powerful narrow band UHF frequency (450–470 MHz) transmitter (FCC Part 90) with approximately more than one mile range. The sink node sends its collected data to a central server repository where data is processed using a wireless telephone line. Data extraction is based on periodic queries: every six hours for certain old meter node models and every hour for newer meter node models. In our study, we will use the data collected from the newer meter node models with hourly queries.

3.2 Data Extraction and Cleaning [7]

In collaboration with the City of Moncton, we were able to collect one year of data for more than 20,000 clients, the majority of whom have 4 readings/day, and some 24 readings/day. The results presented here are for the 1129 m, with 24 readings/day, as they offer better sampling of the water consumption process. Before implementing our approach described in the previous section, it was necessary to do some verification and cleaning of the raw data to remove any reading errors and make sure all meters have

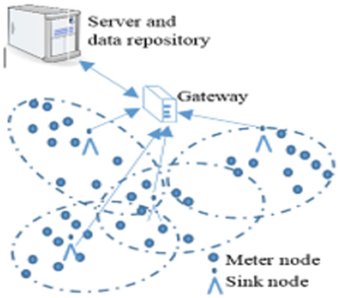


Fig. 1. The WSN architecture of the City of Moncton [7]

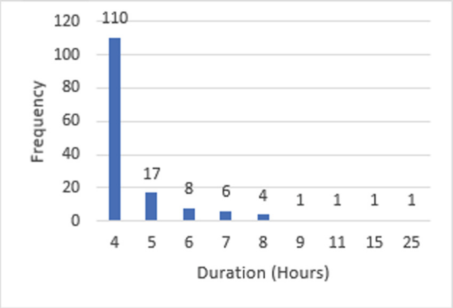


Fig. 2. Frequency of AWC readings vs. their durations for ε = 100%

exactly 24 readings/day, taken every hour. Some meters have more or less than 24 readings/day for certain days of the year. As shown in Fig. 1, certain nodes (meters) can reply to the queries coming from 2 sink nodes, as they are located within the communication range of both sink nodes, and the server will receive 2 readings for the same meter with a slight difference in time stamps. For example, we can have a reading at 4:00:00 PM and another at 4:00:10 PM, with a time difference of 10 s. In this case, we removed one of the readings. In other situations, where data is simply missing for unknown reasons, bad transmission for example, we were able to extrapolate the missing data as long as they are limited in number, i.e., less than 0.01%. Beyond this threshold, the meter data was removed from our study. It should be noted that the following results related to data transmission saving are compared to the current system used by the City of Moncton where no data transmission saving strategy is applied.

3.3 Abnormal Water Consumption

The main purpose of WMRs is to ensure proper water consumption monitoring and avoid waste of a natural resource fresh water. Consequently, it is important to detect any abnormal water consumption resulting from major leaks or from consumer abuse. There are a variety of reasons for water consumption increase/decrease, depending on the type of consumer, residential or industrial, as well as changes in users' habits. Not all sudden major water consumption increases are a result of water leaks, this why it is difficult to identify the source of increases. However, we can observe and quantify these increases through frequent WMRs. Major increases in water consumption will be described by the term 'Abnormal Water Consumption (AWC)'. An AWC event is observed when water consumption goes beyond a certain percentage ε of water consumption's moving average measured over n periods of time preceding this event. The AWC, or increase may last over extended periods during WMRs. Its importance depends on the value of ε and the duration of water consumption increase. An important AWC requires a careful monitoring, i.e. frequent WMRs.

4 Proposed Approach

Before describing our formal approach, we briefly explain the rationale behind it.

(a) As previously mentioned, the main purpose of WMR monitoring is to prevent major water waste and take the required action in the event of water leaks. Frequent WMRs, hourly in our case, should allow the detection of AWC. However, based on the data analysis, provided by the City of Moncton, the number of AWC increases is very limited and the events are of short duration. Figure 2 shows a histogram of AWC for 1129 water meters during the period from July 1st to December 31st (24 readings/day). An AWC event is detected when the water consumption exceeds a certain threshold we can define, (a formal definition will be provided later). It is obvious that the number of AWC events, 149, is very small relative to the total number of readings (4,877,280), i.e. less than 0.01%. Furthermore, the number of events where water consumption lasted more than 8 h is in fact more limited, with only 4 events. We believe that frequent WMRs are not necessary for all water meters.

(b) Knowing the history of water consumption for each subscriber (private house, rental, industry, etc.) should give us some insight to build user profiles. An AWC is difficult to define. For example, a vacant rental property may see its water consumption suddenly increase when it is rented, yet this increase cannot be considered as an AWC event. Similar events may occur when home owners decide to fill up their swimming pools during the summer. This is why it is important to categorize subscribers based on their profile to avoid false AWC events [5].

(c) It is essential to isolate certain profiles (WMNs) that show AWC and put them into a specific group to be monitored frequently. This monitoring is done hourly in our study. The remaining profiles can be classified in groups or clusters, based on their water consumption variation. In our case, we chose the standard deviation as clustering criterion. Experimental results presented in Fig. 4 show that the best sampling or observation period is 36 h (36 readings), as it provides the best saving in terms of data transmissions.

(d) As the water consumption profile may change over time, it is important that we re-analyze the WMRs and adjust our partitioning clustering. In our application, we considered two time intervals; weekly and monthly.

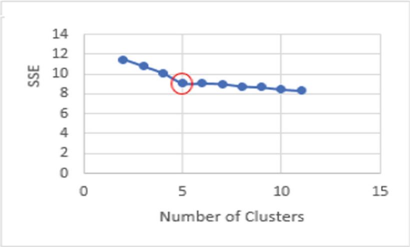


Fig. 3. Elbow method; SSE vs number of clusters

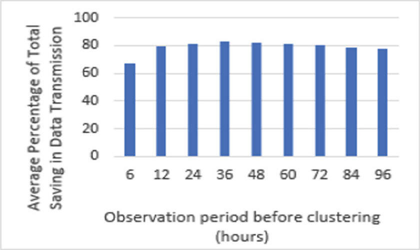


Fig. 4. Average percentage of total saving in data transmission vs observation (sampling) period

Below, we give a formal description of our approach. It includes the following steps:

Step 1. Isolate the set β of WMNs where AWC is detected. Let ε be the percentage of water consumption increase compared to the water consumption moving average $\gamma_{i,n}$ for a water meter i, computed over n previous periods of time. Let $a_i(t_j)$ be the water consumption for WMN i in the laps of time $t_j - t_{j-1}$.

i = 1, 2…. M, M is the number of water meters
j = 1, 2…. L, L is the number of readings,

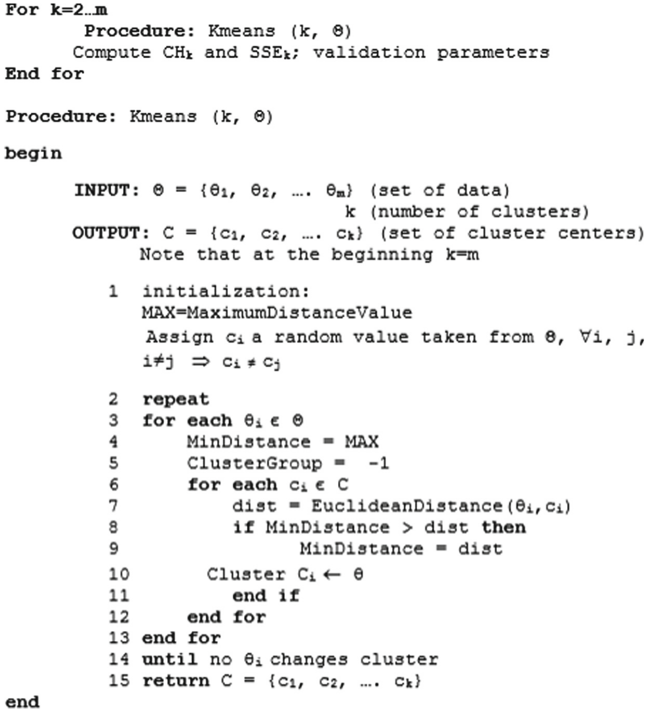
$$\forall i \in \beta,\ \left(a_i\left(t_j\right)\right) > (\gamma_{i,n} + \varepsilon^{*} \gamma_{i,n})$$

$\forall i \in \alpha,\ \left(a_i\left(t_j\right)\right) \leq (\gamma_{i,n} + \varepsilon^{*} \gamma_{i,n})$, α represents the set of WMN where water consumption doesn't exceed the threshold ($\gamma_{i,n}$ + $\varepsilon^{*}\gamma_{i,n}$).

Step 2. For each WMN $i \in \alpha$, compute the standard deviation θ_i from the set of data $\{a_i(t_j), j = 1\ldots L\}$, $i = 1\ldots m$. $\Theta = \{\theta_1, \theta_2, \ldots. \theta_m\}$, m is the number of elements in α.

Step 3. Partition Θ into g sub clusters $C_1, C_2, \ldots. C_k$, using the K-means method [3] as follows:

Let c_i be the center of sub-cluster C_i.



```
For k=2…m
      Procedure: Kmeans (k, Θ)
     Compute CHk and SSEk; validation parameters
End for

Procedure: Kmeans (k, Θ)

begin

     INPUT: Θ = {θ1, θ2, …. θm} (set of data)
                               k (number of clusters)
     OUTPUT: C = {c1, c2, …. ck} (set of cluster centers)
          Note that at the beginning k=m

          initialization:
          MAX=MaximumDistanceValue
          Assign ci a random value taken from θ, ∀i, j,
          i≠j ⇒ ci ≠ cj

          repeat
          for each θi ε Θ
              MinDistance = MAX
              ClusterGroup =  -1
              for each ci ε C
                  dist = EuclideanDistance(θi,ci)
                  if MinDistance > dist then
                        MinDistance = dist
                Cluster Ci ← θ
                  end if
              end for
          end for
          until no θi changes cluster
          return C = {c1, c2, …. ck}
end
```

Step 4. Validation: Knowing CH_k and SSE_k, see Table 1, determine the number of clusters g and validate this number using the elbow method for SSE, see Fig. 3 and the highest value given by the CH method [8].

Step 5. Compute the laps of time between two readings for each sub-cluster θ_i as follows: Upper Limit, UL = Max (c_i) Lower Limit, LL = Min (c_i)

- Assign the maximum sampling period to the sub-cluster with a center value c_i = LL and a minimum sampling period for the cluster with a center value c_i = UL. We can use a linear scale to assign a sampling period for each sub-cluster with LL < c_i < UL, a smaller sampling value for those near the UL and a bigger sampling value for those near LL. We believe that a sub cluster with small standard deviation should be monitored less frequently than the one with bigger value as they show little variation in their water consumption.

As water consumptions changes over time, it is important to retune the clustering. After a defined period T, we restart the procedure starting from step 1.

5 Experimental Results

Our experimental results are based on the data collected from 1,129 water meters from July 1st to December 31st, 24 readings/day. Figure 2 shows the frequency of AWC readings vs. their durations for ε = 100%. There are 149 AWC when we sample data hourly. When we implement our approach, which applies different data sampling periods, larger and cluster dependent, we were able to detect 122 AWC events, including all major event i.e. with more 4 h duration. This result is satisfactory in terms of the reliability of water meter monitoring. Figure 4 shows the Average Percentage of Total Saving in data transmission vs observation (sampling) period. From this figure, we can see that the best observation period is 36 h as it offers the maximum saving. Table 1. shows the values of SSE and CH as function of the number of clusters. The optimal number of cluster g is equal to 5. Figure 3 shows another representation for SSE where the elbow location indicates the optimal number of clusters.

In order to evaluate the performance of our approach we will compare our results to the data reduction techniques [7] where redundant readings are removed. Figure 5 and 6 show respectively the distribution (histogram) of data transmissions saving over 6 months when the data reduction technique and the data partitioning clustering approach are applied. From these figures, we observe that the most important saving occurs

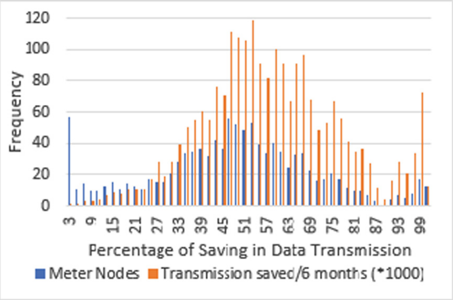


Fig. 5. Data transmission saving when data reduction technique is applied.

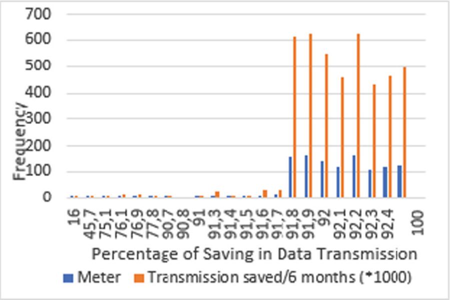


Fig. 6. Data transmission saving when the proposed partitioning clustering approach is used

between 45% and 75% for the data reduction technique and between 91% and 92% for the data partitioning clustering approach. Note that there is ‘0’ saving for the WMNs in the set β as we monitor water consumption every hour.

Figure 7 shows the average percentage saving in data transmission for our approach as compared to the reduction technique. In our approach, the clustering is repeated weekly and monthly, and in both cases our solution offers better performance than the reduction technique. The improvement of saving is approximately 10% and 40% for weekly and monthly monitoring respectively.

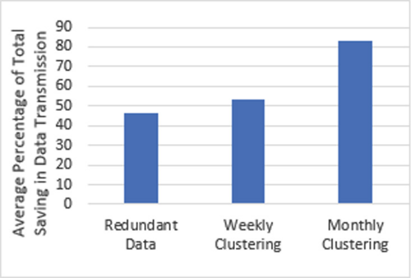


Fig. 7. Comparison between cluster and redundancy approach

Table 1. SSE and CH values according the number of clusters

Number of clusters	SSE	CH
2	11,4915	0,01083
3	10,78	0,00735
4	10,0916	0,00637
5	**9,0819**	**0,01888**
6	9,0606	0,00570
7	9,0025	0,01047
8	8,7651	0,00380
9	8,6731	0,00843
10	8,456	0,00783
11	8,3308	0,00694

6 Conclusion

In this paper, we proposed a novel approach for RDT in WSNs that is based on partitioning clustering technique. To the best of our knowledge, it is the first time this approach is applied in this area. We compared the performance of our approach to another reduction technique based on redundant readings. Experimental results showed that our algorithm performs better in terms of transmission saving. In future work, we plan to compare our approach with modeling and prediction techniques.

Acknowledgement. This work is supported in part by an NSERC Discovery Grant awarded to Prof. Jalal Almhana and the Youth 1000 Funding of Prof. Philippe Fournier-Viger from the NSFC. We are grateful to the City of Moncton who provided us with the data.

References

1. Anastasi, G., Conti, M., Di Francesco, M., Passarella, A.: Energy conservation in wireless sensor networks using data reduction approaches: a survey. Int. J. Comput. Eng. Res. **7**(3), 537–568 (2013)
2. Karim, L., Anpalagan, A., Nasser, N., Almhana, J.: Sensor-based M2M agriculture monitoring systems for developing countries: state and challenges. Netw. Protoc. Algorithm J. **5**(3), 68–86 (2013)
3. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: 5th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297. University of California Press (1967)
4. Tsai, K., Ye, M., Leu, F.: Secure power management scheme for WSN. In: 7th ACM CCS International Workshop on Managing Insider Security Threat, MIST 2015, pp. 63–66 (2015)
5. de Souza, K., Fournier-Vigier, P., Almhana, J.: Early detection of abnormal residential water consumption. Technical report (2017)
6. Said, J.E., Karim, L., Almhana, J., Anpalagan, A.: Heterogeneous mobility and connectivity-based clustering protocol for wireless sensor networks. In: ICC 2014, pp. 257–262 (2014)
7. Almhana, C., Choulakian, V., Almhana, J.: An efficient approach for data transmission in power-constrained wireless sensor network. In: ICC 2017, pp. 4058–4064 (2017)
8. Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J., Wu, S.: Understanding and enhancement of internal clustering validation measures. IEEE Trans. Cybern. **43**(3), 982–994 (2013)
9. Pantazis, N.A., Nikolidakis, S.A., Vergados, D.D.: Energy-efficient routing protocols in wireless sensor networks: a survey. Commun. Surv. Tutorials **15**(2), 551–591 (2013)
10. Ye, W., Heidemann, J.: An energy-efficient MAC protocol for wireless sensor networks. In: IEEE Computer and Communications Societies, pp. 1567–1576 (2002)
11. Younis, O., Fahmy, S.: HEED: a hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks. IEEE Trans. Mob. Comput. **3**(4), 366–379 (2004)
12. Patil, S.A., Mishra, P.: Improved mobicast routing protocol to minimize energy consumption for underwater sensor networks. Int. J. Res. Sci. Eng. **3**(2), 197–204 (2017)
13. Liao, T.W.: Clustering of time series data – a survey. Pattern Recogn. **38**(11), 1857–1874 (2005)
14. Smith, B.A., Wong, A., Rajagopal, R.: A simple way to use interval data to segment residential customers for energy efficiency and demand response program targeting. In: ACEEE Summer Study on Energy Efficiency in Buildings (2012)
15. Lavin, A., Klabjan, D.: Clustering time-series energy data from smart meters. Energy Effi. **8**, 681–689 (2015)